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

Supplementary material

Annex A. Cohorts, diarrhoeal disease definitions and genotyping

ALSPAC (discovery and replication)

Study design

The Avon Longitudinal Study of Parents And Children (ALSPAC) is a large, prospective cohort study based in the South West of England. 14,541 pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992 were recruited and detailed information has been collected on these women and their offspring at regular intervals (1). The study website contains details of all the data that is available through a fully searchable data dictionary (http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/). Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Only Caucasian children were included in the study. Half of the population was included in the discovery step [ALSPAC (Disc)] and the other half in the replication [ALSPAC (Repl)].

Outcome definition: diarrhoea and doctor diagnosis of diarrhoea (questionnaire)

Diarrhoeal disease was assessed by questionnaires administrated to the parents at 18 months and at 30 months of age and they refer to the period from last visit. Questions at 18 and 30 months were respectively: “Has your child had diarrhoea since she was 6 months old? Yes and saw the doctor, Yes but did not visit the doctor, No” and “Has your child had diarrhoea since she was 18 months old? Yes and saw the doctor, Yes but did not visit the doctor, No”. Diarrhoea was defined by combining doctor and non doctor affirmative answers.

Genetic data and statistical analysis

GWAS data was generated by Sample Logistics and Genotyping Facilities at the Wellcome Trust Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe. 975 individuals were genotyped at the WTSI and 9,382 were genotyped at LabCorp, both on the Illumina 550K Custom chip. All individuals of non-European ancestry, ambiguous sex, extreme heterozygosity (<0.32 or >0.345 in the WTSI set and <0.31 or >0.33 in the

LabCorp set), cryptic relatedness (>10% IBD) and high missingness (>3%) were removed. SNPs with low genotyping rate (<95%), with low minor allele frequency (<1%), out of Hardy Weinberg equilibrium (p<5x10-7) or from the pseudo-autosomal region of the X chromosome were excluded. 8,365 individuals typed on 464,311 probes remained. Phasing of the SNPs was carried out using MaCHv1.0 Markov Chain Haplotyping software (2) and imputation was carried out using Minimac (3) using phase 1 of 1000 Genomes reference panel (v3.20101123, ALL populations, no monomorphic/singletons). The final imputed dataset consisted of 8,365 individuals and 31,337,615 variants. Genome-wide association analysis was carried out using mach2datv1.0.23 (4). Summary statistics were available for 30M variants that were successfully analyzed. This study only includes children with a Caucasian origin.

CHOP (replication)

Study design

The Center for Applied Genomics (CAG) has recruited ~80K pediatric patients from The Children's Hospital of Philadelphia (CHOP) (5). Enrolment into the study is random and therefore encompasses all of the major common pediatric disorders. Upon enrollment, CAG is authorized to extract the patients' medical history from their electronic medical record (EMR) and store the information in a de-identified database. Approximately 90% of enrollees consent to a yearly EMR update. Biological samples including DNA have been collected from all patients. The study was approved by the CHOP Institutional Review Board (IRB). Written informed consent for participation in the study was obtained from all participants and their parents or guardians. CHOP participated in the replication step.

Outcome definition: doctor diagnosis of diarrhoea (medical record)

Information on diarrhoeal diseasewas obtained by a clinical researcher going through the general practitioner records of each child and recording when each general practitioner diagnosis was made. Cases were defined as "diarrhoea diagnosis or medication for diarrhoea or mention of diarrhoea in the reports". Controls were children not in the previous category among records reviewed and with at least one visit to the hospital. Two time windows were explored: from 6 to 18 months, and from 18 to 30 months.

Vaccinated children identified through the CHOP medical records were excluded from the analysis. Given that vaccination is extensive against Rotavirus is the USA, we carried out a sensitivity analysis excluding all children from CHOP and results did not change substantially (not shown).

Genetic data and statistical analysis

Samples were genotyped on either the Illumina HumanHap550 or the HumanHap610 at CAG, following the manufacturers' instructions. Standard quality control parameters were applied to the dataset, samples with chip-wide genotyping failure rate < 5% were excluded; SNPs with minor allele frequencies of < 1%; genotyping failure rates of greater that 5% or Hardy-Weinberg P-Values less that $1x10^{-6}$ were excluded from further analysis.

Imputation of untyped markers (~39M) was carried out using IMPUTE2 (6) after prephasing with SHAPEIT (7). Each chromosome was prephased separately. Reference phased cosmopolitan haplotypes and recombination rates were obtained from the 1000 Genomes project (1000 Genomes phase I integrated release 3). Imputation was carried out in 5Mb intervals using an effective population size of 20,000 as recommended. Statistical tests for association were carried out using the SNPTEST v2 package (8). The analysis was restricted to Caucasian children.

COPSAC2000 (discovery)

Study design

The Copenhagen prospective study on asthma in childhood (COPSAC2000) is a prospective clinical study of a birth cohort of 411 infants born to mothers with a history of asthma. The newborns were enrolled at the age of 1 month, the recruitment of which was previously described in detail (9). The study was approved by the Ethics Committee for Copenhagen (KF 01- 289/96) and The Danish Data Protection Agency (2008-41-1754) and informed consent was obtained from both parents.

Outcome definition: doctor diagnosis of diarrhoea (medical interview)

Diarrhoea was determined in medical interviews of the parents. Parents were interviewed by the research doctor and asked if the child had suffered from any illnesses, including diarrhoea and/or vomiting, since their last visit. The interviews took place at age 6, 12, 18 and at 24 months and collected information from the past 6 months.

Genetic data and analysis

High throughput genome-wide SNP genotyping were performed using the Illumina Infinium™ II HumanHap550 v1 and v3 platform (Illumina, San Diego), at the Children's Hospital of Philadelphia's Center for Applied Genomics. Phasing of the SNPs was carried out using MaCHv1.0 Markov Chain Haplotyping software (2) and imputation was carried out using

Minimac (3) using phase 1 of 1000 Genomes admixed reference panel. Genome-wide analysis was carried out using mach2dat v1.0.21 (4).

COPSAC2010 (replication)

Study design

The Copenhagen prospective study on asthma in childhood (COPSAC2010) is a population based longitudinal clinical study of 700 pregnant women and their offspring (10). The families were monitored closely from week 24 of mothers' pregnancy till age 3 year of the offspring. From the age of 1 week the child was examined at 10 scheduled visits to the research center and additional visits at onset of any skin or respiratory symptoms.

Outcome definition: doctor diagnosis of diarrhoea (medical interview)

Diarrhoea was determined in medical interviews of the parents. Parents were interviewed by the research doctor and asked if their child had any illness, including diarrhoea and/or vomiting, since their last visit. The interviews took place every 6 months and collected information from the last 6 months. Between visits, the parents kept a diary on infectious symptoms, including gastroenteritis.

Genetic data and statistical analysis

Genotyping of 951,117 genetic markers were carried on the Illumina Infinium Human Omni Express Exome Bead chip at the AROS Applied Biotechnology AS center, in Aarhus, Denmark. Genotypes were called with Illumina's GenomeStudio software. All samples underwent quality control (QC) filters, where genetic variants with Hardy-Weinberg equilibrium p values >10-6, minor allele frequency (MAF>0.01), genotyping call rate > 0.95 were retained. We excluded individuals with genotyping call rate < 0.95, gender mismatches, genetic duplicates, outlying heterozygosity >0.27 and <0.037, and those individuals not clustering with the CEU individuals (Utah residents with ancestry from northern and Western Europe) through a multi-dimensional clustering analyses (MDS) seeded with individuals from the International Hap Map Phase 3. Imputation to 1000 Genomes reference panel using SHAPEIT v2 (7) and IMPUTE2 (6) was carried out at Barcelona Supercomputing Center (www.bsc.es) after doing the appropriate post-imputation QC (MAF>0.01, HWE in controls>5e-6, info>0.882) and successfully imputed ~7.2 M polymorphic variants. Genome-wide analysis was carried out using SNPTESTv.2.5 (8) using the score method, frequentist statistics, assuming and additive model, and using the first five population principal components as covariates.

GenerationR (discovery)

Study design

The Generation R Study is a population-based prospective cohort study of pregnant women and their children from fetal life onwards in Rotterdam, The Netherlands (11). All children were born between April 2002 and January 2006, and currently followed until young adulthood. Of all eligible children in the study area, 61% were participating in the study at birth. Cord blood samples including DNA have been collected at birth. The study protocol was approved by the Medical Ethical Committee of the Erasmus Medical Centre, Rotterdam (MEC 217.595/2002/20). Written informed consent was obtained from parents of all participants.

Outcome definition: diarrhoea and doctor diagnosis of diarrhoea (questionnaire)

Diarrhoeal disease was assessed by questionnaire administrated to the parents at the infant's age at 12 months and at 24 months. Questions at 12 months were: "Has your child had infections of the stomach/intestine or gastric flu in the last 6 months? No; Yes; Yes went to general practitioner; Yes went to hospital." Diarrhoea was defined as the combination "yes, did not go to doctor", "yes went to general practitioner" and "yes went to hospital". Questions at 24 months were: "Has your child had infections of the stomach/intestine or gastric flu in the last 12 months? No; Yes, did not go to doctor; Yes went to general practitioner; Yes went to hospital." Diarrhoea was defined as the combination "yes", "yes went to general practitioner" and "yes went to hospital". Doctor diagnosis for gastroenteritis was the combination of the last two answers.

Genetic data and statistical analysis

Samples were genotyped using IlluminaHumanHap610/660 Quad Arrays following standard manufacturer's protocols. Intensity files were analyzed using the Beadstudio Genotyping Module software v.3.2.32 and genotype calling based on default cluster files. Any sample displaying call rates below 97.5%, excess of autosomal heterozygosity (F<mean-4 SD), and mismatch between called and phenotypic gender were excluded. In addition, individuals identified as genetic outliers by the IBS clustering analysis (>3 standard deviations away from the HapMap CEU population mean) were excluded from the analysis. Genotypes were imputed for all polymorphic SNPs from phased haplotypes in autosomal chromosomes using the 1000 Genomes GIANTv3 panel in Minimac (3). Twins were excluded from the analyses. Ethnicity was grouped into Caucasians and non-Caucasians, based on genetic ancestry. Ancestry determination analysis included the genomic data of all Generation R individuals merged with the three reference panels of the HapMap Project Phase II (YRI, CEU and CHB/JPT). Only

Caucasians were analyzed in this study. Detailed information of the Generation R genetic dataset can be found elsewhere (12).

Association between diarrhoeal disease and GWAS SNPs was performed using a regression framework adjusting for population stratification in the Generation R cohort using mach2dat (4) as implemented in GRIMP (13). Four principal components were used to control for population stratification in the Caucasian subpopulation.

GINIplus (replication)

Study design

A total of 5,991 mothers and their newborns were recruited into the German Infant study on the influence of Nutrition Intervention PLUS environmental and genetic influences on allergy development (GINIplus) between September 1995 and June 1998 in Munich and Wesel. Infants with at least one allergic parent and/or sibling were allocated to the interventional study arm investigating the effect of different hydrolyzed formulas for allergy prevention in the first year of life. All children without a family history of allergic diseases and children whose parents did not give consent for the intervention were allocated to the non-interventional arm. Detailed descriptions of the GINIplus study have been published elsewhere (14). DNA was collected at the age 6 and 10 years. Approval by the local Ethics Committees and written consent from participant's families were obtained.

Outcome definition: doctor diagnosis of diarrhoea (questionnaire)

Diarrhoea definition was based on questionnaire data administrated to the parents at 24 months of age. The question was: "Did a doctor diagnose your child with diarrhoea (with or without vomiting) during the last 12 months?".

Genetic data and statistical analysis

794 children from the GINIplus study were analyzed using the Affymetrix Human SNP Array 5.0. Genotypes were called using BRLMM-P. Criteria for exclusion of individuals were: a call rate below 95%, a heterozygosity outside mean +/- 4 SD, a failure of the sex check or a failure of the similarity quality control using MDS analysis based on IBS. Criteria for exclusion of variants were: a call rate below 95%, a MAF < 0.01 and a HWE p-value < 0.00001. The filtered data sets were prephased using SHAPEIT v2 (7) and imputation was done using IMPUTE2 (6) considering the haplotypes from the 1000 Genomes project phase I v3 as a reference (March 2012 release, updated version from 26 Aug 2012, all ancestries, limited to variants with more than one minor allele copy). Genome-wide association analysis of gastroenteritis was carried

out in SNPTEST v2.4 (8) regressing expected allelic dosage on case-control status, including sex as a covariate.

LISAplus (discovery)

Study design

The influence of Life-style 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 is a population-based birth cohort study. A total of 3,094 healthy, full-term neonates were recruited between 1997 and 1999 in Munich, Leipzig, Wesel and Bad Honnef. Detailed descriptions of the LISIplus study have been published elsewhere (15). DNA was collected at the age 6 and 10 years. Approval by the local Ethics Committees and written consent from participant's families were obtained.

Outcome definition: doctor diagnosis of diarrhoea (questionnaire)

In LISA, diarrhoea was based on questionnaire data administrated to the parents at 12 months and at 24 months of age. The same question was asked at the both surveys: "Did a doctor diagnose your child with diarrhoea (with or without vomiting) during the last 6 months?".

Genetic data and statistical analysis

585 from the LISA study were analyzed using the Affymetrix Human SNP Array 5.0 and 88 individuals with the Affymetrix Human SNP Array 6.0. Genotypes were called using BRLMM-P algorithm (5.0), respectively BIRDSEED V2 algorithm (6.0). Criteria for exclusion of individuals were: a call rate below 95%, a heterozygosity outside mean +/- 4 SD, a failure of the sex check or a failure of the similarity quality control using MDS analysis based on IBS. Criteria for exclusion of variants were: a call rate below 95%, a MAF < 0.01 and a HWE p-value < 0.00001. The filtered data sets were prephased using SHAPEIT v2 (7) and imputation was done using IMPUTE2 (6) considering the haplotypes from the 1000 Genomes project phase I v3 as a reference (March 2012 release, updated version from 26 Aug 2012, all ancestries, limited to variants with more than one minor allele copy). Genome-wide association analysis of gastroenteritis was carried out in SNPTEST v2.4 (8) regressing expected allelic dosage on case-control status, including sex as a covariate.

INMA_SAB (discovery)

Study design

The INMA—INfancia y Medio Ambiente—(Environment and Childhood) Project is a network of birth cohorts in Spain that aim to study the role of environmental pollutants in air, water and diet during pregnancy and early childhood in relation to child growth and development (16). The study has been approved by Ethical Committee of each participating centre and written consent was obtained from participating parents. Data for this study comes from INMA Sabadell (children born between 2004 and 2007).

Outcome definition: diarrhoea (questionnaire)

In INMA Sabadell, information on diarrhea was obtained from questionnaires administrated to the parents by a nurse. The questions at 14 months and at 24 months of age were: "From last interview (6 months), has your child had diarrhoea and vomiting? Yes/No" and "During last 12 months, has your child had diarrhoea or vomits? Yes/No".

Genetic data and statistical analysis

DNA was obtained from cord blood or whole blood collected at 4 years of age using the Chemagen protocol at the Spanish National Genotyping Centre (CEGEN). Children whose parents reported to be white and to be born in Spain or in European countries and that were not lost during the follow-up were selected for genotyping. Genome-wide genotyping was performed using the HumanOmni1-Quad Beadchip (Illumina) at CEGEN. Genotype calling was done using the GeneTrain2.0 algorithm based on HapMap clusters implemented in the GenomeStudio software. Quality control was done using PLINK (17) and following standard criteria. First of all, SNPs were flipped to the human genome + strand. We applied the following initial quality control thresholds: sample call rate>98% and/or LRR SD<0.3. Then, we checked sex, relatedness, heterozygosity and population stratification. Genetic variants were filtered for SNP call rate>95%, MAF>1% and HWE p value>1.10E-6. Imputation was done using IMPUTE2 (6) and a cosmopolitan panel from the 1000 Genomes project (release March 2012, downloaded from http://mathgen.stats.ox.ac.uk/impute/ALL_1000G_phase1integrated_v3_impute.tgz) as a reference. Genome-wide association analysis was carried out using SNPTEST (8). Summary statistics were available for approximately 30M variants that were successfully analyzed.

INMA_VAL (replication)

Study design

The INMA—INfancia y Medio Ambiente—(Environment and Childhood) Project is a network of birth cohorts in Spain that aim to study the role of environmental pollutants in air, water and

diet during pregnancy and early childhood in relation to child growth and development (16). The study has been approved by Ethical Committee of each participating centre and written consent was obtained from participating parents. Data for the replication came from INMA Valencia (children born between 2004 and 2006).

Outcome definition: diarrhoea and doctor diagnosis of diarrhoea (questionnaire)

In INMA Valencia data was obtained similarly. The question at age 12 months was: "Has your child visited the doctor for gastroenteritis (diarrhoea and vomiting)? Yes/No"; and at age 24 months: "From last interview (12 months), has your child had diarrhoea and vomiting? Yes/No".

Genetic data and analysis

DNA was obtained from cord blood using the Chemagen protocol at the Spanish National Genotyping Centre (CEGEN). Children whose parents reported to be white and to be born in Spain or in European countries and that were not lost during the follow-up were selected for genotyping. Genome-wide genotyping was performed using the HumanOmni1-Quad Beadchip (Illumina) at CEGEN. Genotype calling was done using the GeneTrain2.0 algorithm based on HapMap clusters implemented in the GenomeStudio software. Quality control was done using PLINK (17) and following standard criteria. First of all, SNPs were flipped to the human genome + strand. We applied the following initial quality control thresholds: sample call rate>98% and/or LRR SD<0.3. Then, we checked sex, relatedness, heterozygosity and population stratification. Genetic variants were filtered for SNP call rate>95%, MAF>1% and HWE p value>1.10E-6. Imputation was done using IMPUTE2 (6) and a cosmopolitan panel from the 1000 Genomes project (release March 2012, downloaded from http://mathgen.stats.ox.ac.uk/impute/ALL_1000G_phase1integrated_v3_impute.tgz) as a reference.Genome-wide association analysis was carried out using SNPTEST (8) program. Summary statistics were available for approximately 30M variants that were successfully analyzed.

MoBa (replication)

Study design

The Norwegian Mother and Child Cohort Study (MoBa) is a prospective population-based pregnancy cohort conducted by the Norwegian Institute of Public Health (18). Participants were recruited from all over Norway from 1999-2008, and 41% of invited women consented to participate. The cohort includes 114,000 children, 95,000 mothers and 75,000 fathers. Blood

samples were obtained from both parents during pregnancy and from mothers and children (umbilical cord) at birth. Follow-up is conducted by questionnaires at regular intervals and by linkage to national health registries. MoBa has obtained a license from the Norwegian Data Inspectorate. MoBa participated in the replication stage.

Outcome definition: diarrhoea (questionnaire)

Only the children born at term were included in the present study. Diarrhoea was defined from questionnaire Q5 administrated to the mothers when a child was approximately 18 months of age. Four questions were selected: EE240 "Has your child had gastric flu/diarrhoea between 6-11 months of age?", EE241 "How many times?", EE242 "Has your child had gastric flu/diarrhoea between 12-18 months of age?", EE243 "How many times?". Every child could have values at both age-ranges. A positive answer to EE240 or (when EE240 answer was missing) EE241 values in the range of 1-20 were considered as indicators for diarrhoea at 6-11 months of age, while control status was assigned to children with negative answer to EE240 with concordant values ("0" or missing) at EE241, or when EE240 was missing, but EE241 had a value "0". Similarly, cases and controls were defined for the age-range of 12-18 months. For the association analysis we combined diarrhoea indicators from two age-ranges: reported diarrhoea episode(s) at any age-range indicated cases, while controls were children without reported diarrhoea episodes at any age-range.

Genetic data and analysis

Preterm born and term born children were selected for genotyping. Genotyping was carried out using Illumina Human660W-Quad genotyping BeadChip. Genotyping quality control excluded SNPs with the call rate lower than 97%, SNPs with HWE p-value <10E-5, ambiguous SNPs (A/T, C/G), monomorphic SNPs, SNPs that "changed chromosome" (when comparing genotyping manifest file with 1000G reference), X chromosome SNPs located in pseudo-autosomal regions, SNPs with 1000G-incompatable alleles (e.g., genotyped A/C vs 1000G A/G). Genotyping quality control excluded samples with more than 3% missing genotypes, samples with X chromosome heterozygosity problems (male F<0.8, female F>0.2), individuals with the fraction of heterozygous genotypes being 2 SD from the group's mean, sample duplicates, one sample (with a higher missingness rate) from each of related sample-pairs (PI_HAT>0.1875), non-Europeans (PCA, visual exclusion). In imputation procedures, haplotypes from 1000 Genomes project (phase1, v3, release date March 2012) were used as a reference panel, software SHAPEIT v2.644 (7) was used to phase the genotypes, software IMPUTE2v2.3.0 (6) was use to impute genotypes. For statistical analyses statistical computing program "R"

(version 3.2.3, 2015-12-10) and a custom-build analysis script were used. Generalized linear model was fitted on the case/control status using function "glm" (family - "binomial"), the effect allele dosage and the child's gender.

Annex B. Vaccination against *Rotavirus*

In 2006, two new live, oral, attenuated *Rotavirus* vaccines were licensed for infants less than six months old: the two dose monovalent human *Rotavirus* vaccine (Rotarix, GSK) and the three-dose pentavalent bovine-human, reassortant vaccine (RotaTeq, Sanofi Pasteur MSD). Vaccines confer up to 85% protection against *Rotavirus* disease of any severity (19, 20).

The inclusion of vaccination with either *Rotavirus* vaccine in national immunization programs has been recommended worldwide by the World Health Organization (WHO) since 2009 (21). Since 2006, the Centers for Disease Control and Prevention Advisory Committee on Immunization Practices have recommended routine *Rotavirus* vaccination of infants in the USA (22). The same year, vaccination was implemented in this group.

In Europe, vaccination against *Rotavirus* has not been introduced systematically in the public health system except for a few countries: Austria, Belgium, Luxemburg, Finland, UK, Germany and Norway (23). Moreover, in UK, Germany and Norway it was not introduced until recently (2013-2014). However before 2013-2014, in countries such as Germany and Spain there were recommendations about *Rotavirus* vaccination from different health care agencies and as a result of this around 20% of the population was vaccinated (Table Annex-B1).

Most of the cohorts participating in this study enrolled their children before systematic introduction of vaccination against *Rotavirus*, except for CHOP in the USA (Supplementary Table S11). In CHOP, children with genome-wide genetic data were enrolled from 2006, the same year when the vaccine was licensed. In 2008 approximately 60% of the children in Philadelphia were vaccinated (24). In the present study, the proportion of *Rotavirus* vaccination according to The Children's Hospital of Philadelphia medical records was around 7-10% (Table Annex-B2). Because of missing data on vaccination (children vaccinated in other health centers different from The Children's Hospital of Philadelphia) and passive immunization, a sensitivity analysis excluding CHOP cohort was conducted and results did not change substantially.

Table Annex-B1. *Rotavirus* vaccination programs, coverage and years of cohort enrollment.

Cohort	Years of enrollment	Years when children in the study were 1y old	Year of introduction in the national immunization programme	Vaccine coverage from literature within each country	Overlap between cohort recruitment and vaccination
ALSPAC	1991-1992	1992-1993	2013 (recommended from 2012)	0% in 2010	No
CHOP	1987-present	from 2006	2006	60% in 2008	Yes[a]
COPSAC2000	1998-2000	1999-2001	not introduced	1-4% in 2010	No
COPSAC2010	2009-2010	2010-2011	not introduced	1-4% in 2010	Very limited
Generation R	2002-2006	2003-2007	not introduced	0% in 2010	No
GINI	1995-1998	1996-1999	2013 (recommended from 2010)	24.8% in 2010	No
LISA	1997-1999	1998-2000	2013 (recommended from 2010)	24.8% in 2010	No
INMA_SAB	2004-2006	2005-2007	not introduced (recommended from 2008)	21.7% in 2010	Very limited
INMA_VAL	2003-2005	2004-2006	not introduced (recommended from 2008)	21.7% in 2010	No
MoBa	1999-2008	2000-2009	2014 (recommended from 2011)	0-2% in 2010	Very limited

Vaccines against *Rotavirus* were commercialized in 2006

[a]Vaccinated children based on medical records have been excluded from the analysis

Table Annex-B2. Proportion of *Rotavirus* vaccinated children in CHOP according to medical records.

Age	N total	% vaccinated	N controls	% vaccinated	N cases	% vaccinated
6 to 18 months	3595	7.80%	3399	6.80%	196	25%
18 to 30 months	3595	10.30%	3387	10.45%	208	8.65%

References

1. Fraser A., Macdonald-Wallis C., Tilling K., Boyd A., Golding J., Davey Smith G., Henderson J., Macleod J., Molloy L., Ness A. *et al.* (2013) Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int. J. Epidemiol.,* **42,** 97-110.
2. 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.
3. Howie B., Fuchsberger C., Stephens M., Marchini J. and Abecasis G. R. (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.,* **44,** 955-959.
4. Li Y., Willer C., Sanna S. and Abecasis G. (2009) Genotype imputation. *Annu Rev Genomics Hum. Genet.,* **10,** 387-406.
5. Hakonarson H., Grant S. F., Bradfield J. P., Marchand L., Kim C. E., Glessner J. T., Grabs R., Casalunovo T., Taback S. P., Frackelton E. C. *et al.* (2007) A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature,* **448,** 591-594.
6. 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.
7. Delaneau O., Marchini J. and Zagury J. F. (2011) A linear complexity phasing method for thousands of genomes. *Nat. Methods,* **9,** 179-181.
8. Marchini J., Howie B., Myers S., McVean G. and Donnelly P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.,* **39,** 906-913.
9. Bisgaard H. (2004) The Copenhagen Prospective Study on Asthma in Childhood (COPSAC): design, rationale, and baseline data from a longitudinal birth cohort study. *Ann. Allergy Asthma Immunol.,* **93,** 381-389.
10. Bisgaard H., Vissing N. H., Carson C. G., Bischoff A. L., Folsgaard N. V., Kreiner-Moller E., Chawes B. L., Stokholm J., Pedersen L., Bjarnadottir E. *et al.* (2013) Deep phenotyping of the unselected COPSAC2010 birth cohort study. *Clin. Exp. Allergy,* **43,** 1384-1394.

11. Jaddoe V. W., van Duijn C. M., Franco O. H., van der Heijden A. J., van Iizendoorn M. H., de Jongste J. C., van der Lugt A., Mackenbach J. P., Moll H. A., Raat H. *et al.* (2012) The Generation R Study: design and cohort update 2012. *Eur. J. Epidemiol.,* **27,** 739-756.
12. Medina-Gomez C., Felix J. F., Estrada K., Peters M. J., Herrera L., Kruithof C. J., Duijts L., Hofman A., van Duijn C. M., Uitterlinden A. G. *et al.* (2015) Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study. *Eur. J. Epidemiol.,* **30,** 317-330.
13. Estrada K., Abuseiris A., Grosveld F. G., Uitterlinden A. G., Knoch T. A. and Rivadeneira F. (2009) GRIMP: a web- and grid-based tool for high-speed analysis of large-scale genome-wide association using imputed data. *Bioinformatics,* **25,** 2750-2752.
14. Berg A., Kramer U., Link E., Bollrath C., Heinrich J., Brockow I., Koletzko S., Grubl A., Filipiak-Pittroff B., Wichmann H. E. *et al.* (2010) Impact of early feeding on childhood eczema: development after nutritional intervention compared with the natural course - the GINIplus study up to the age of 6 years. *Clin. Exp. Allergy.,* **40,** 627-636.
15. Heinrich J., Bolte G., Holscher B., Douwes J., Lehmann I., Fahlbusch B., Bischof W., Weiss M., Borte M. and Wichmann H. E. (2002) Allergens and endotoxin on mothers' mattresses and total immunoglobulin E in cord blood of neonates. *Eur. Respir. J.,* **20,** 617-623.
16. Guxens M., Ballester F., Espada M., Fernandez M. F., Grimalt J. O., Ibarluzea J., Olea N., Rebagliato M., Tardon A., Torrent M. *et al.* (2012) Cohort Profile: the INMA--INfancia y Medio Ambiente--(Environment and Childhood) Project. *Int. J. Epidemiol.,* **41,** 930-940.
17. Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M. A., Bender D., Maller J., Sklar P., de Bakker P. I., Daly M. J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.,* **81,** 559-575.
18. Magnus P., Birke C., Vejrup K., Haugan A., Alsaker E., Daltveit A. K., Handal M., Haugen M., Hoiseth G., Knudsen G. P. *et al.* (2016) Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). *Int. J. Epidemiol*.
19. Ruiz-Palacios G. M., Perez-Schael I., Velazquez F. R., Abate H., Breuer T., Clemens S. C., Cheuvart B., Espinoza F., Gillard P., Innis B. L. *et al.* (2006) Safety and efficacy of an attenuated vaccine against severe rotavirus gastroenteritis. *N. Engl. J. Med,* **354,** 11-22.

20. Cortese M. M., Immergluck L. C., Held M., Jain S., Chan T., Grizas A. P., Khizer S., Barrett C., Quaye O., Mijatovic-Rustempasic S. *et al.* (2013) Effectiveness of monovalent and pentavalent rotavirus vaccine. *Pediatrics,* **132,** e25-33.
21. WHO (2009) Rotavirus vaccines: an update. *Wkly. Epidemiol. Rec.,* **84,** 533-540.
22. Cortese M. M. and Parashar U. D. (2009) Prevention of rotavirus gastroenteritis among infants and children: recommendations of the Advisory Committee on Immunization Practices (ACIP). *MMWR Recomm. Rep.,* **58,** 1-25.
23. Parez N., Giaquinto C., Du Roure C., Martinon-Torres F., Spoulou V., Van Damme P. and Vesikari T. (2014) Rotavirus vaccination in Europe: drivers and barriers. *Lancet Infect. Dis.,* **14,** 416-425.
24. Clark H. F., Lawley D., Mallette L. A., DiNubile M. J. and Hodinka R. L. (2009) Decline in cases of rotavirus gastroenteritis presenting to The Children's Hospital of Philadelphia after introduction of a pentavalent rotavirus vaccine. *Clin. Vaccine. Immunol.,* **16,** 382-386.

Legends to Supplementary Figures

Supplementary Figure S1. A) Quantile-quantile (Q-Q) plots showing the probability values from GWAS meta-analysis of doctor diagnosis of diarrhoea at age one year (DD1Y). The red line indicates the distribution under the null hypothesis. Lambda values are shown. **B)** Manhattan plot of the GWAS meta-analysis of doctor diagnosis of diarrhoea at age one year (DD1Y). The x-axis represents the autosomal chromosomes and the y-axis represents–$\log_{10}$ (*p*). The red dashed line indicates genome-wide significance (*p*=5.00E-08) and the blue dashed line indicates suggestive genome-wide significance (*p*=1.00E-05). Variants with genome-wide significance are represented with red dots.

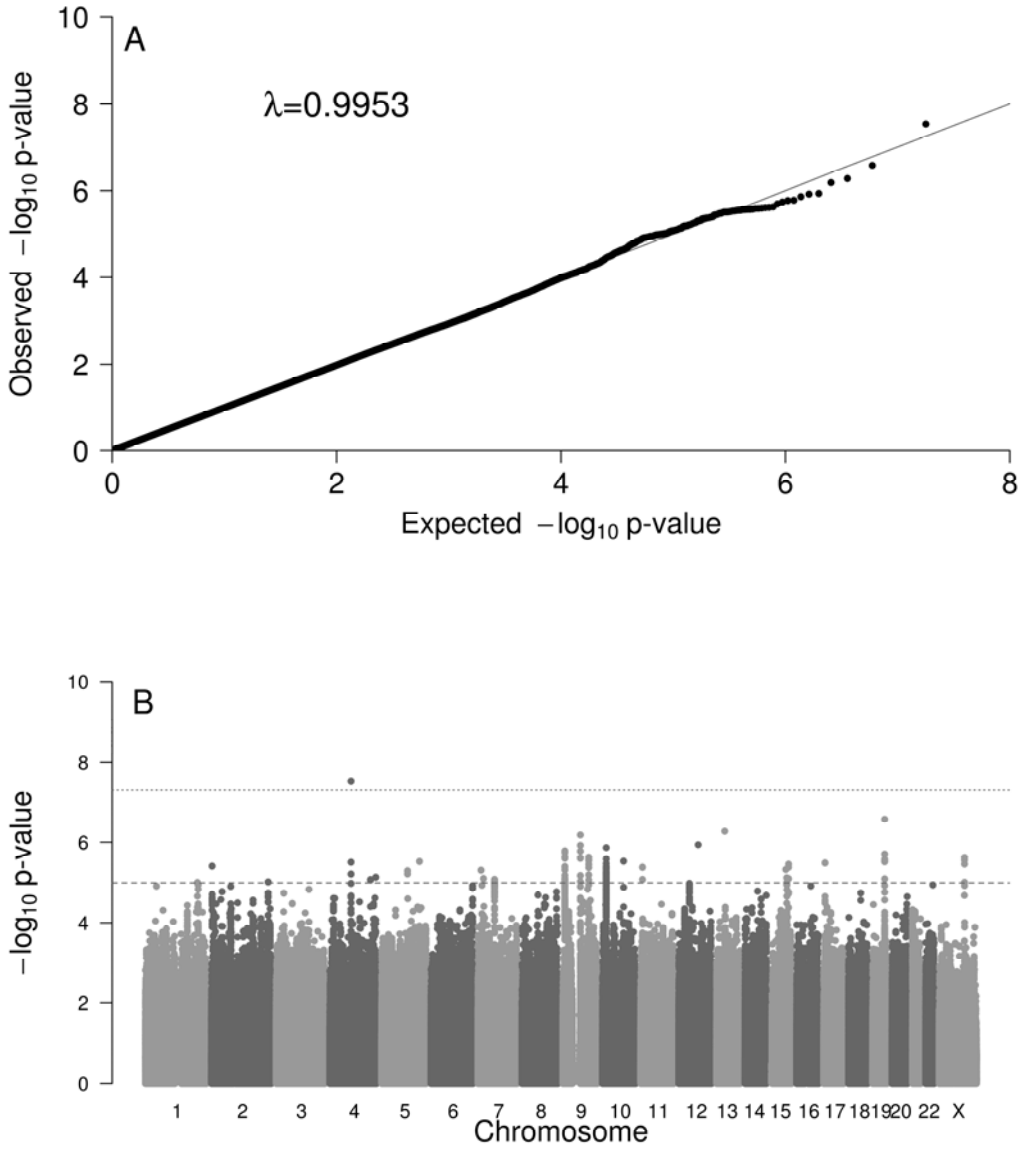
Supplementary Figure S2. A) Quantile-quantile (Q-Q) plots showing the probability values from GWAS meta-analysis of diarrhoea at age two years (D2Y). The red line indicates the distribution under the null hypothesis. Lambda values are shown. **B)** Manhattan plot of the GWAS meta-analysis of diarrhoea at age two years (D2Y). The x-axis represents the autosomal chromosomes and the y-axis represents–$\log_{10}$ (*p*). The red dashed line indicates genome-wide significance (*p*=5.00E-08) and the blue dashed line indicates suggestive genome-wide significance (*p*=1.00E-05). Variants with genome-wide significance are represented with red dots.

Supplementary Figure S3. A) Quantile-quantile (Q-Q) plots showing the probability values from GWAS meta-analysis of doctor diagnosis of diarrhoea at age two years (DD2Y). The red line indicates the distribution under the null hypothesis. Lambda values are shown. **B)** Manhattan plot of the GWAS meta-analysis of doctor diagnosis of diarrhoea at age two years (DD2Y). The x-axis represents the autosomal chromosomes and the y-axis represents–$\log_{10}$ (*p*). The red dashed line indicates genome-wide significance (*p*=5.00E-08) and the blue dashed line indicates suggestive genome-wide significance (*p*=1.00E-05). Variants with genome-wide significance are represented with red dots.

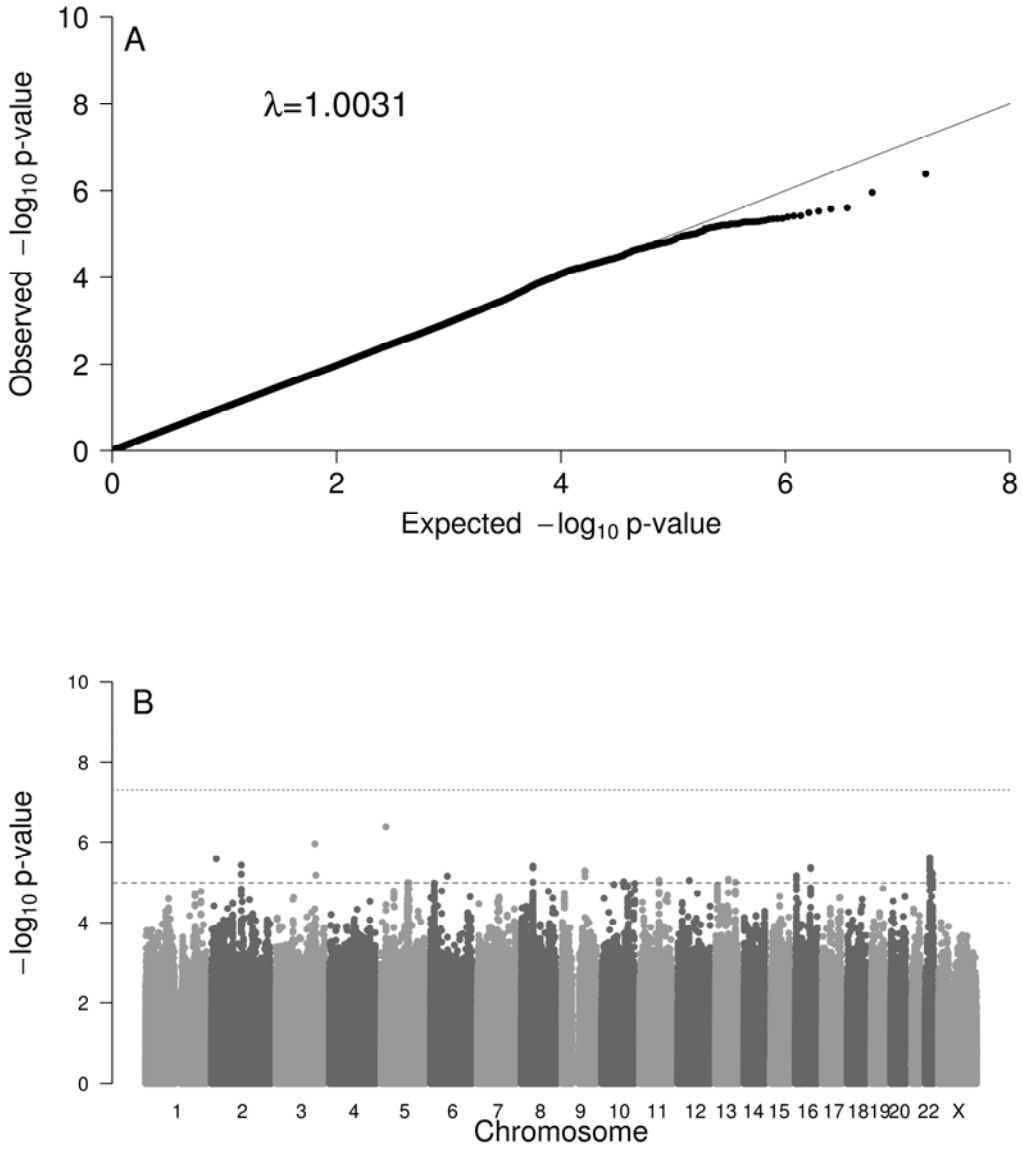
Supplementary Figure S4. Forest plots for rs601338 at 19q13.33 for the four diarrhoeal outcomes: **A)** Diarrhoea at age one year (D1Y); **B)** Doctor diagnosis of diarrhoea at age one year (DD1Y); **C)** Diarrhoea at age two years (D2Y); **D)** Doctor diagnosis of diarrhoea at age two

years (DD2Y). In the vertical panel, the studies participating in the discovery or replication phase are presented. In the horizontal lines, the 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.

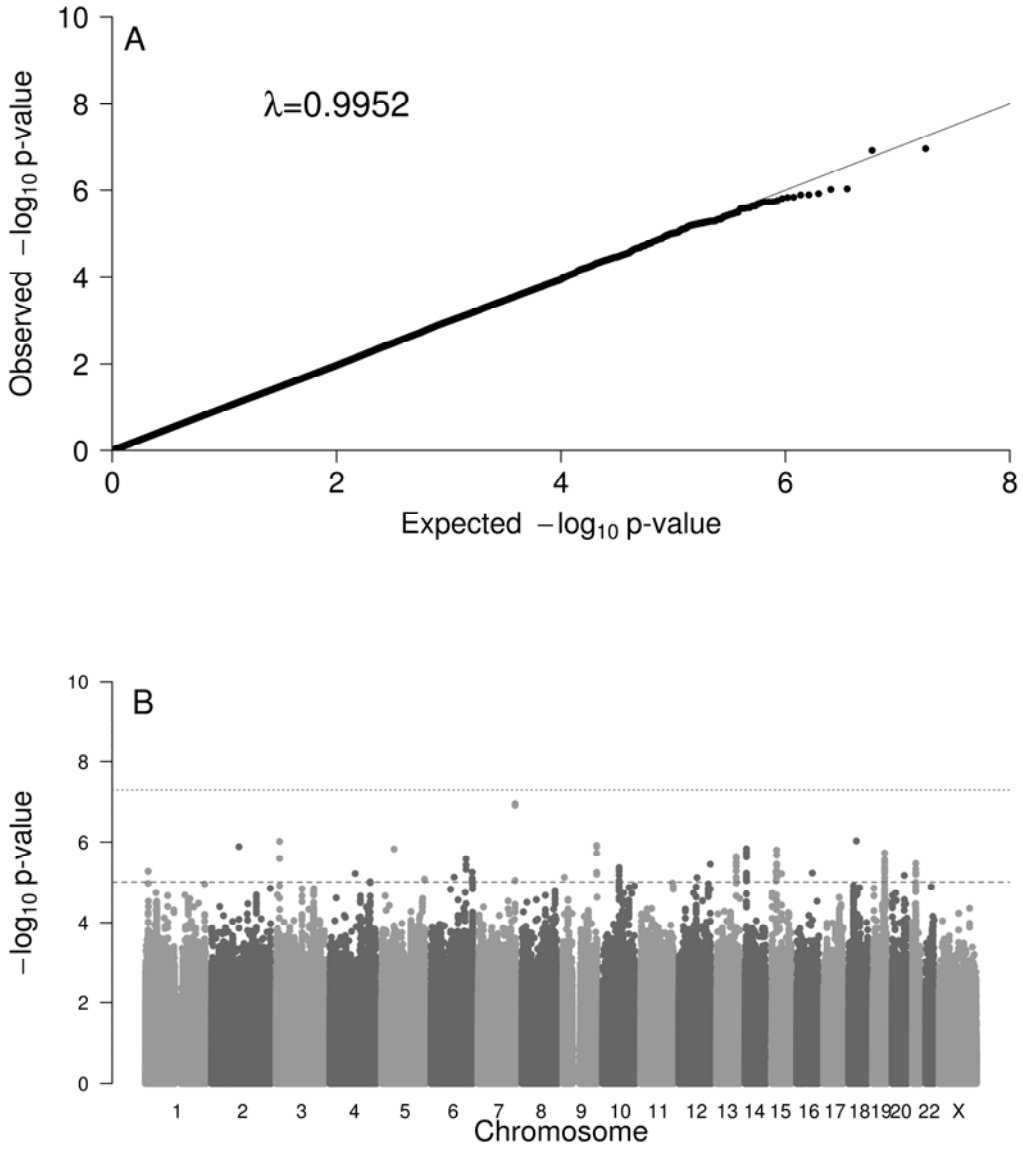
Supplementary Figure S1



Supplementary Figure S2

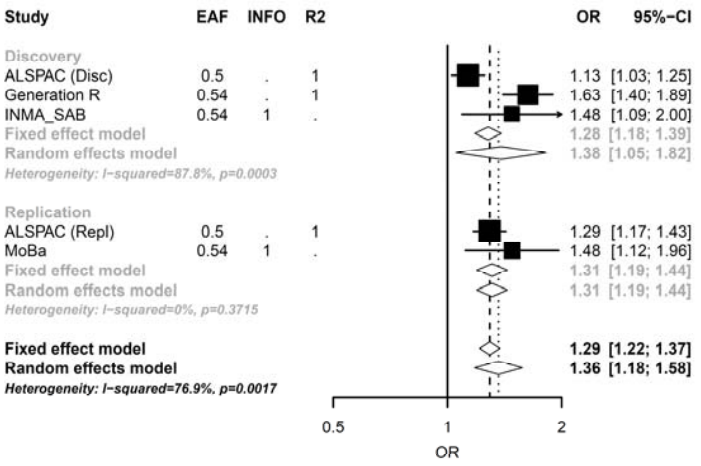


Supplementary Figure S3

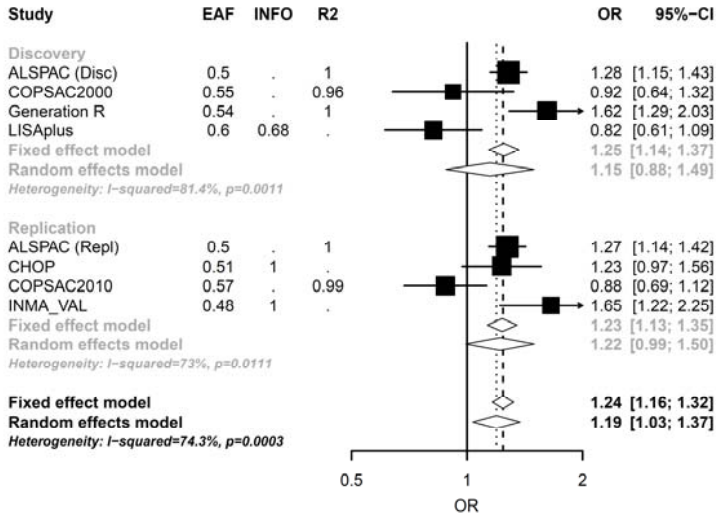


Supplementary Figure S4

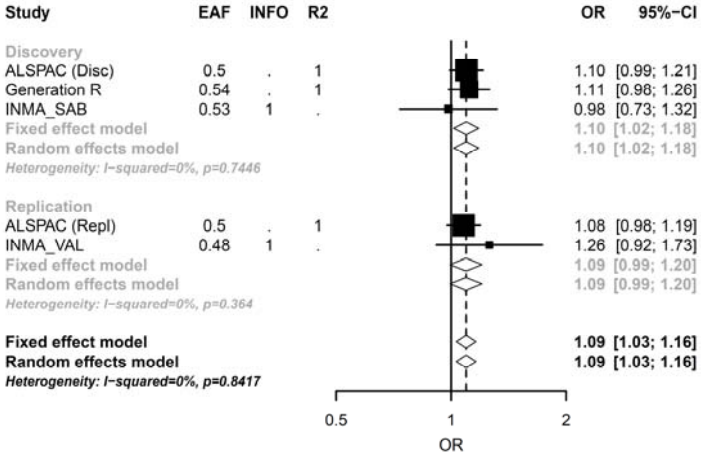
A. Diarrhoea at age one year (D1Y)



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)