ORIGINAL ARTICLE

Multiplex genomewide association analysis of breast milk fatty acid composition extends the phenotypic association and potential selection of FADS1 variants to arachidonic acid, a critical infant micronutrient

Josyf C Mychaleckyj, Uma Nayak, E Ross Colgate, Dadong Zhang, Tommy Carstensen, Shahnawaz Ahmed, Tahmeed Ahmed, Alexander J Mentzer, Masud Alam, Beth D Kirkpatrick, Rashidul Haque, Abu Syed Golam Faruque, William A Petri Jr

ABSTRACT

Background Breast milk is the sole nutrition source during exclusive breastfeeding, and polyunsaturated fatty acids (FAs) are critical micronutrients in infant physical and cognitive development. There has been no prior genomewide association study of breast milk, hence our objective was to test for genetic association with breast milk FA composition.

Methods We measured the fractional composition of 26 individual FAs in breast milk samples from three cohorts totalling 1142 Bangladeshi mothers whose infants were genotyped on the Illumina Mega chip and replicated on a custom Affymetrix 30K SNP array (n=616). Maternal genotypes were imputed using IMPUTE.

Results After running 33 separate FA fraction phenotypes, we found that SNPs known to be associated with serum FAs in the FADS1/2/3 region were also associated with breast milk FA composition (experiment-wise significance threshold 4.2x10^{-9}). Hypothesis-neutral comparison of the 33 fractions showed that the most significant genetic association at the FADS1/2/3 locus was with fraction of arachidonic acid (AA) at SNP rs174556, with a very large per major allele effect size of 17% higher breast milk AA level. There was no evidence of independent association at FADS1/2/3 with any other FA or SNP after conditioning on AA and rs174556. We also found novel significant experiment-wise SNP associations with: polyunsaturated fatty acid (PUFA) 6/PUFA3 ratio (sorting nexin 29), eicosanoid (intergenic) and capric (component of oligomeric Golgi complex 3) acids; and six additional loci at genomewide significance (<5x10^{-8}).

Conclusions AA is the primary FA in breast milk influenced by genetic variation at the FADS1/2/3 locus, extending the potential phenotypes under genetic selection to include breast milk composition, thereby possibly affecting infant growth or cognition. Breast milk FA composition is influenced by maternal genetics in addition to diet and body composition.

INTRODUCTION

Long-chain polyunsaturated fatty acids (LCPUFAs) are important for growth and cognitive development during early life since they are structural components of membrane phospholipids, precursors of inflammation-mediating eicosanoids, and also modulate gene expression by acting as agonists or ligands for transcription factors. In postnatal exclusively breastfeeding infants, breast milk is the sole source of these compounds and other essential FAs. Since the first enzyme in the omega-6 and omega-3 pathway conversion of precursor essential linoleic (LA, C18:2n6) and alpha-linolenic (ALA, C18:3n3) FAs, delta-6 desaturase encoded by the FA desaturase 2 gene (FADS2), is likely rate limiting, and LCPUFAs are beta-oxidised or used in developmental processes, endogenous neonatal LCPUFAs synthesis is unable to meet total demand and preformed LCPUFA metabolites in breast milk are necessary to prevent depletion. Hence, the composition of breast milk is important in ensuring that the infant obtains the correct balance of macronutrients and micronutrients, particularly the critical arachidonic acid (AA, 20:4n6) and docosahexaenoic acid (DHA, 22:6n-3) acids.

Previous candidate gene and genomewide association studies (GWAS) of FA levels have focused on circulating FA levels in blood plasma or erythrocyte membranes in adults with fewer in infants and children and were motivated by hypotheses of lipids as mediators of cardiovascular disease, inflammation and cancer. By comparison, there are limited extant data exploring the association of genetic variation with breast milk composition, and previous studies have generally tested a few candidate SNPs in the FADS1/2/3 region against selected breast milk LCPUFA components.

Given the dearth of information about genetic influences on breast milk composition and hence the nutritional supply to the early developing ex utero infant, we undertook a GWAS of genetic variants associated with the percentage composition of 26 FAs in breast milk samples from more than 1100 Bangladeshi mothers, enrolled with their infants in three cohorts in two locations in Bangladesh, from the Performance of Rotavirus and Oral Polio Vaccines in Developing Countries (PROVIDE) study, and a more recent study of Cryptosporidium infection...
Complex traits

in infants (Field Studies of Cryptosporidiosis in Bangladesh, ‘CRYPTO’ study, manuscript in preparation). We hypothesised that GWA analysis could identify genetic variants that are associated with breast milk FA composition, and by comparing the GWA results of the 26 assayed FAs, we could identify the FA components most likely to be under the influence of genetic variation. The study design included infant genotyping only, but the breast milk compositional data prompted us to ask whether we could impute maternal genotypes using the uniparental obligate allelic transmissions and hence perform a GNAS in the mothers accounting for uncertainty in the imputation. Our results suggest that the approach worked well, particularly benefiting from the extraordinary proportion of variance in the FA traits explained by individual SNPs.

MATERIALS AND METHODS

Study populations

Study participants were drawn from two separate studies of three birth cohorts conducted in two different locations in Bangladesh: urban Mirpur Thana (in Dhaka) and rural Mirzapur Thana (25 miles North-West of Mirpur). Both populations generally lacked access to treated water, but Mirzapur had reduced household crowding compared with the denser slums of Mirpur.

PROVIDE study

The study design and population has been described elsewhere.27 28 Seven hundred predominantly slum-dwelling mother–infant dyads, with no known maternal or fetal complications, were enrolled in Mirpur, Dhaka within 7 days post-delivery, into a birth cohort between May 2011 and November 2012.

Field studies of Cryptosporidiosis in Bangladesh (CRYPTO) study

Seven hundred and fifty-eight mother–infant pairs were recruited into two cohorts, one in Mirzapur (n=258) and one in Mirpur (n=500), Bangladesh, enrolled within 7 days post-delivery between June 2014 and March 2016.

The protocol and procedures were approved by the Ethical Review Committee for human subjects protection and Research Review Committee for scientific merit at the International Centre for Diarrhoeal Diseases Research, Bangladesh and the Institutional Review Boards at the University of Virginia and University of Vermont. All participants provided written informed consent at study entry.

Breast milk FA determination

A description of breast milk collection and quantification of a comprehensive panel of FAs to include saturated, monounsaturated and polyunsaturated FAs has been published for the PROVIDE study.29 Similar sample collection and processing was adopted in CRYPTO study. A single breast milk sample from 1419 mothers (PROVIDE n=683 and CRYPTO n=736), collected within 6-weeks postpartum using a dried milk spot protocol, was assayed for 26 FAs by gas chromatography at OmegaQuant Analytics Laboratory (Sioux Falls, South Dakota, USA). Individual breast milk FAs were expressed as %wt/wt of total identified FA.30 31

DNA extraction and handling

Blood samples from 700 PROVIDE and 713 CRYPTO infants were drawn in the field clinics and processed using standard laboratory protocols. More details are in online supplementary methods.

Custom Affymetrix Axiom 30K SNP ‘MalChip’ Array—PROVIDE study

A custom Affymetrix SNP array was developed during 2013 to investigate the genetic associations of a broad spectrum of infant and maternal phenotypic traits linked to impaired infant growth, development, metabolism, infectious disease susceptibility, enteric and systemic inflammation and cognitive development. PubMed searches identified SNPs described in prior GWAS of related phenotypes. The 11 500 unique SNPs identified were supplemented with tagging SNPs for 162 candidate genes. After designability review, the resulting malnutrition SNP array (dubbed the ‘MalChip’) contained probes for 33 588 SNPs. Four-digit resolution human leucocyte antigen (HLA) genotyping was performed on the 700 PROVIDE infant DNA samples and converted to 159 pseudo-SNPs which were merged into the MalChip genotype file. Six hundred and forty infant DNA samples were genotyped on the MalChip at the Center for Public Health Genomics, University of Virginia, USA on an Affymetrix GeneTitan machine. More details are available in online supplementary methods.

Illumina GWA genotyping—both studies

Infant DNA samples from the PROVIDE (n=576) and CRYPTO (n=672) studies were also genotyped under the VacGene consortium programme (VacGene: Characterising the Genetic Determinants of Vaccine Response in Children from the Developing World) on precommercial versions of the Illumina multieithnic genotyping array (MEGA) at the Wellcome Trust Sanger Institute. The second version of the array contained 1 522 034 SNPs and was used for the PROVIDE samples. Subsequent genotyping for the CRYPTO samples was performed on a later array version containing 1 635 469 SNPs.

Imputation of maternal genotypes

Using obligate Mendelian inheritance and assuming Hardy-Weinberg equilibrium, we calculated the probability that a maternal genotype is either AA, AB or BB conditional on the genotype of the infant. Statistical details are described in online supplementary methods. For SNPs directly typed in the infant, the imputation into mother is the only source of uncertainty, but for untyped, imputed infant SNPs, the maternal genotype is the convolution of the two sources of uncertainty. The dual loss of information means that properly calibrated statistical inference from de novo ranking of maternal association results in imputed SNPs is not possible, so the initial GWAS results were filtered to include only genotyped or perfectly imputed SNPs. A custom script converted the infant genotype data to maternal genotype data in Oxford Statistics format, preserving the maternal imputation uncertainty.

SNP imputation and GWA analysis

Imputation and GWA analysis of the two studies was performed separately, at different times, as genotyping data became available. The samples in both study data sets were separately phased using SHAPEIT V2.r8372 and then imputed using IMPUTE V2.3.2 and Oct2014 Phase3 1000 Genomes panels (more details in online supplementary methods). SNP association analysis was performed using PLINK1.9 for the MalChip infant genotype data and SNPTEST2.5.24 for the maternal genotype data imputed from the infant MalChip data, without reducing the maternal genotypes to an expected non-integer allele dosage. The score test in SNPTEST incorporates the maternal imputation at typed SNPs analogously to the uncertainty with SNP
imputation. Because of the maternal genotype imputation, the initial genomewide scan was restricted to SNPs that were either directly genotyped in the infants or perfectly imputed with info=1 in both PROVIDE and CRYPTO to avoid double loss of information through untyped and maternal imputation. Selected SNPs of interest with info<1 were tested as described.

Phenotypes and statistical models
The quantitative trait outcomes in these analyses were the 26 FA concentrations, expressed as %wt/wt, bounded in composition range [0,100%], summing to 100% per mother. Seven derived major summary fraction FA measures, computed were added as phenotypes: SFA (total saturated), PUFA3 (total cis-polymunsaturated omega-3), PUFA6 (total cis-polymunsaturated, omega-6), PUFA (total PUFA6+PUFA3), PUFA6/PUFA3 ratio, MFA (total monounsaturated) and TFA (total trans). All were log transformed to stabilise variance, except PUFA6/PUFA3 (square root). Multiple linear regression models were run within each study using a score test to test the marginal additive association of each SNP with each FA phenotype, and the summary results were meta-analysed using a fixed-effects model using METAL.35 The minimal GWAS screening model included adjustments for maternal age, infant age at breast milk sample, infant sex and adjustment for study site (CRYPTO study only). Principal components were not included in the association screening model but were included in robustness tests of selected results (described further in online supplementary methods). Conditional models for AA and for the leading SNP also included log(%AA) or lead SNP as covariates. The compositional FA phenotypes had a range of intercorrelation coefficient magnitudes 0–0.998 resulting in 12.0 effective independent phenotypes calculated from the intercorrelation coefficient magnitudes. The compositional FA phenotypes had a range [0,100%], summing to 100% per mother. Seven derived major summary fraction FA measures, computed were added as phenotypes: SFA (total saturated), PUFA3 (total cis-polymunsaturated omega-3), PUFA6 (total cis-polymunsaturated, omega-6), PUFA (total PUFA6+PUFA3), PUFA6/PUFA3 ratio, MFA (total monounsaturated) and TFA (total trans). All were log transformed to stabilise variance, except PUFA6/PUFA3 (square root). Multiple linear regression models were run within each study using a score test to test the marginal additive association of each SNP with each FA phenotype, and the summary results were meta-analysed using a fixed-effects model using METAL.35 The minimal GWAS screening model included adjustments for maternal age, infant age at breast milk sample, infant sex and adjustment for study site (CRYPTO study only). Principal components were not included in the association screening model but were included in robustness tests of selected results (described further in online supplementary methods). Conditional models for AA and for the leading SNP also included log(%AA) or lead SNP as covariates. The compositional FA phenotypes had a range of intercorrelation coefficient magnitudes 0–0.998 resulting in 12.0 effective independent phenotypes calculated from the eigenvalues of the FA correlation matrix using the method of Li,36 yielding an experiment-wise Bonferroni-adjusted significance rate of 5×10−6.

RESULTS
Demographic and clinical characteristics of the study populations
See online supplementary table S1 which summarises the demographic and clinical characteristics of the Bangladeshi families, and the participant/sample flow and loss to follow-up is shown in online supplementary figure S1. After study dropout and sample quality control (QC), n=532 (PROVIDE) and n=402 (CRYPTO) families from Mirzapur and n=208 (CRYPTO) from Mirzapur were retained with both breast milk FA concentrations and infant GWAS data for analysis. The Affymetrix replication ‘MalChip’ PROVIDE study cohort retained 616 families, because genotyping was conducted prior to the GWAS and total DNA collection from neonatal infants was limited. The breast milk samples were collected at mean days of lactation (infant age) of 5.8–10.9 days with range 3–43 days. A summary of the breast milk FA concentrations is shown in table 1 and the FA correlation structure in figure 1. There were significant differences between the cohorts in the breast milk FA composition, with higher PUFA6 and PUFA3 in the CRYPTO cohorts compared with PROVIDE, with lower SFA/PUFA ratio and higher concentration of AA and DHA.

Genotyping results
The genotyping QC results are shown in online supplementary table S2, and more detailed discussion is in online supplementary results. For the custom Affymetrix 30K MalChip, post-QC, 626 (640 total) PROVIDE samples and 20 908 SNPs (33 588 total) remained. After merging the 159 HLA pseudo-SNPs, 16 688 (21067=79.2%) were common in this population with minor allele frequency (MAF) >0.05. For the Illumina preproduction MEGA array used, 776 921 SNPs (52.8%) were polymorphic in PROVIDE and 970 928 SNPs (60.3%) in CRYPTO. Filtering the set of SNPs to those with info=1 in both study groups and MAF≥0.05 (SNP passed QC in both studies) or MAF>0.1 (SNP passed QC in only one study), there were approximately 932K SNPs available.

Ancestry of the Bangladeshi population
Figure 2 shows the first three principal components of the PROVIDE and CRYPTO study samples, 541+404=945 Bangladeshi in Dhaka (dark green glyphs) and 226 Bangladeshi in Mirzapur (light green glyphs) projected onto components defined using the 1000 genomes populations (online supplementary table S3). The samples lie between the European and South/ East Asian clusters and overlapped the 1000 genomes Indian subcontinent samples, particularly Bengalis in Bangladesh. The samples comprise a reasonably tight cluster which explained the low genomic control inflation and are genetically more similar to the European population (60%) than to East Asians (40%), stemming from ancient migration and admixture events.

Genomewide scans identified the FADS1/2/3 region as a locus for breast milk AA and three other FA loci
The GWAS results are shown in table 2 and online supplementary table S4 for 1142 families with complete data, filtered from 932K frequent SNPs (MAF>0.05 both studies; or MAF>0.1 present in one study) with info=1, that is, SNPs genotyped or perfectly imputed. We found four distinct loci that met the maternal experiment-wise significance threshold of 4.2×10−6 and another six that met the less stringent traditional threshold of 5×10−8. The chromosome 11 FADS1/2/3 region was one of the four, but only for breast milk AA fraction, and 24 SNPs met the experiment-wise significance threshold consistent with the known strong LD through this region in Europeans. The FADS1/2/3 locus results are discussed in more detail below. The top SNP association (rs7198559) was with the ratio of total PUFA6/PUFA3 and was located in intron 18 of the sorting nexin 29 (SNX29) gene. The second SNP (rs34440628), associated with eicosenoic acid (EIC, C20:1n9), was found in an intergenic region with no immediately obvious link to function. The fourth locus contained eight SNPs in perfect LD associated with capric acid (C10:0) levels. The SNP listed in table 2 (rs12583793) was within an intron of component of oligomeric Golgi complex 3 (COG3) intron and was closest or within epigenetic marks (H3K4Me1 and H3K27Ac). This SNP was found to be a cis-eQTL for COG3 transcript in transformed fibroblast cells and tibial artery tissue, SLC25A30 in transformed fibroblasts and FAM194B in subcutaneous adipose (Gene–Tissue Expression Project (GTEx) portal, https://www.gtexportal.org). Six other distinct loci were significant at the usual 5×10−8 significance level but not experiment-wise, five of which were located within a gene locus.

AA is the only FA unconditionally associated at FADS1/2/3
We performed a detailed search of the FADS1/2/3 region using more permissive SNP search filters to capture a wider range of possible FA associations. We defined the region as hg19 chr11:61 547 000–61 673 000 encompassing the two previously identified LD blocks in Europeans.13 37 38 Expanding the SNP
filter to info>0.9 in both studies and MAF>0.05 yielded 180 SNPs in the region versus only 85 used in the initial GWAS scan. The results are shown graphically in figure 3A, and the detailed top regional SNPs with maternal P value <1×10^{-9} are in table 3 (full FADS1/2/3 region results are in online supplementary table S5). The clustering of the significant associated SNPs recapitulates the LD structure previously described in Europeans, seen here in the Bangladeshi populations. The top-ranked 58 SNPs were all associated with the AA phenotype, whereas the first non-AA SNP association was rank 59 for DPA6 (docosapentaenoic acid-6, C22:5n6), an omega-6 metabolite of AA after two-carbon elongation and desaturation, P value=5.5×10^{-9}. These results suggested that only AA was associated at FADS1/2/3 in unconditional univariate analyses. We replicated the association at FADS1/2/3 in the PROVIDE study in the separate Affymetrix Axiom MalChip data set (45 SNPs in this region), online supplementary table S6 and supplementary results.

No evidence for common variant association with other, non-AA FAs at FADS1/2/3

The results in table 3 and its extended version suggested that AA was the only FA associated with common variants in the FADS1/2/3 region at experiment-wise or genomewide significance, but the AA association could have masked additional underlying independent FA associations. We repeated the
Mychaleckyj JC, et al. J Med Genet 2018;0:1–10. doi:10.1136/jmedgenet-2017-105134

Complex traits

FADS1/2/3 regional analysis for the remaining 32 phenotypes with log(AA) fraction included as a conditioning covariate in the log-additive linear models. The results are shown in figure 3B. We found no evidence of a second independently associated locus for any of the 33 phenotypes at experiment-wise, genomewide or lesser (1×10^-5) significance. There was no evidence of a secondary associated common variant locus for AA.

rs174556 and the next eight SNPs are the leading FADS1/2/3 SNPs associated with breast milk AA fraction

Our lead SNP, rs174556, lies within intron 2 of the FADS1 gene, less than 100 bp 3' to exon 2, within a region of H3K4Me1 and H3K27Ac marks. FADS1 codes for the omega-6 (and omega-3) pathway enzyme with delta-5 desaturase activity that converts DGLA (20:3n6) to AA (20:4n6) and also catalyses distal down-pathway steps. The next eight SNPs ranked by P value after the top SNP had identical effect sizes and r^2 ≥ 0.98 with rs174556, hence there was insufficient statistical evidence to choose the most likely causative among them by association. We estimated that the major allele of rs174556 led to a 17% increase in fractional AA content per allele and explained 9.6% of the variation in log(AA) (table 3), an extraordinary genetic effect size, but comparable to results seen previously for association of lipid traits in this region. rs174556 was in strong LD with rs174546 which was recently identified as a candidate selected SNP in a GWA of 230 ancient Eurasian genomes. SNP rs174556 was one of the 28 SNPs in high LD that discriminated derived and ancestral European haplotypes spanning 38.9 kb from FADS1 through FADS2.
Literature-based replication from prior studies of selected SNPs in breast milk composition

We have conducted the first GWAS of breast milk composition using a comprehensive panel of 26 individual and 7 derived FA compositions to give tests across 33 correlated compositional phenotypes, which we ranked by SNP and FA in a hypothesis-neutral approach. With results from only 1142 families, we were able to establish experiment-wise significant results for a cluster of SNPs in the chromosome 11 FADS1/2/3 region and that AA was the primary FA fraction in breast milk influenced by maternal genetic variation at this locus, specifically, non-coding SNP rs1746556 localised within the FADS1 gene. A second analysis on a separate array platform recapitulated the signal with only 616 families. For breast milk AA composition, our results suggest that there is only a single independent common variant association signal at FADS1/2/3, and that the nine top SNPs, in near-perfect LD $r^2=0.98\sim1$, are the primary candidates at this locus. This result was consistent with, and replicated, four previous FADS1/2/3 candidate SNP breast milk composition studies, but our results emphasise that the primary breast milk fatty acid component under genetic influence is AA. Using recent GTEx cis-eQTL results, we showed that despite being physically located within the FADS1 gene, the lead SNP rs1746556 was associated with increased FADS2 and/or decreased FADS1 gene expression in various tissues, and because of the very high $r^2$ among these nine SNPs, the same eQTL results broadly apply to them as a cluster. Finally, we identified three other new significant loci for specific breast milk FA fractions, and six others that met the traditional GWAS significance threshold but not the adjusted multiplex phenotype testing threshold hence were considered suggestive. Of additional interest is the fact that this is one of few GWAS performed using samples and data from Bangladeshi populations.

This is the first GWAS of breast milk composition, and compared with other lipid traits and phenotypes, breast milk has been the subject of little previous genetic work. Genetic analysis of breast milk composition has lagged behind other plasma lipid traits because breast milk composition is not a standard diagnostic or risk factor entered into medical records and requires more complex and costly chromatographic analysis of samples from women during limited periods of lactation. The importance of breast milk has lagged behind other plasma lipid traits and phenotypes, breast milk has been the subject of little previous genetic work. Genetic analysis of breast milk composition has lagged behind other plasma lipid traits because breast milk composition is not a standard diagnostic or risk factor entered into medical records and requires more complex and costly chromatographic analysis of samples from women during limited periods of lactation. The importance...
Complex traits

of the work is that the variants may modify infant growth and cognitive development during breastfeeding, particularly in comparing breast-fed versus formula-fed infants. We found three new loci in our multiplex GWAS scan, and there are prior published functional data that suggest a link to lipid metabolism. rs7198595, in an intron of SNX29, was associated with the ratio of total PUFA6/PUFA3. Fox found that an intronic SNP in this gene (rs1641895) was associated with subcutaneous (P value=0.003) and visceral (P value=0.01) adipose tissues in women although our lead SNX29 SNP is not in LD with rs1641895. In a porcine model of mammary gland development during late gestation, SNX29 was one of 68 genes that was upregulated more than twofold in the near-maturity gland (4 days prepartum) versus immature gland (>30 days prepartum). Also, in a GWAS of 13 bovine udder traits, SNPs in the coding region of SNX29 were associated with rear udder height. The fourth most significant SNP (rs12583793) was associated with CAP (C10:0) and is located in an exon of COG3. Besides its energy content, CAP has bactericidal and antifungal properties and may help to protect the infant from infection.

Figure 3  Regional genome association plots for the GWAS results of the FADS1/2/3 region of chromosome 11. The horizontal red dashed line in each plot shows the −log10 (experiment-wise significance rate). Panel A shows the unconditioned additive genetic model results for all tested SNPs (85)×all 33 breast milk fatty acid phenotypes in this region plotted as chromosome map position (bp) versus −log(P value) of the score test result from SNPTTEST. Arachidonic acid phenotype results are highlighted in blue and the lead SNP in cyan. Panel B shows the same SNPs and genetic model, but with the addition of log(AA) as a covariate adjustment to condition on AA concentration in breast milk (AA phenotype results not shown). Panel C shows the same analyses but with the SNP genetic models conditioned on the lead SNP (rs174556) as a covariate. Panel D shows the hg19 local genome physical map with overlaid epigenetic tracks (UCSC browser). GWAS, genome wide association studies.
from pathogens such as *Escherichia coli* and *Candida albicans* on mucosal and skin surfaces.46–48

Our association analyses with breast milk employed a relatively small sample by modern genetic association standards, only 1142 for Illumina GWAS analysis and 616 for our custom MalChip. But even using testing that properly allowed for the imputed loss of information in mothers from infant-genotyped SNPs (SNPTTEST), we were still able to declare statistical significance in all of the 83 SNPs, maternal P values < 0.018. This was due to the extraordinarily large effects on lipid concentrations and metabolites resulting from variation in the FADS genes, seen not just in this study, but also previously.9, 11 The other new significant loci also have large per SNP variance explained but need to be validated in replication cohorts to adjust the winner’s curse inflation of effect sizes.

Previous work has shown that present-day humans have two common and distinct FADS1/2 haplotypes consisting of 28 SNPs that are associated with levels of synthesis of LCPUFA, and several of the top nine SNPs associated with breast milk AA in the FADS1/2/3 region were present in these haplotypes.27 There is evidence of positive selection of the haplotype that enhances the ability to produce AA and DHA in African and European populations.38–40 Our results for breast milk AA content extend the possible lipid-based phenotypes that may have been subject to ancient selection pressure, especially since AA levels directly influence growth and immune function at the very earliest ages, and AA is a critical structural component of brain and neuronal tissue. This could have had important consequences for individual fitness and survival through adulthood.

Our study has limitations. Absent maternal genotyping data we had to impute the maternal genotypes which led to a loss of power. It is highly likely that there are additional GWAS-significant loci for breast milk FA fractions that we would have been able to identify with maternal genotyping data. Breast milk composition adapts as the infant develops over the first few months, but we collected only a single specimen per mother so were not able to test the effect of genetic variation on longitudinal composition changes, and our results represent averaged association over colostrum, transitional and mature milk stages. Our sample size was limited, hence we were only able to identify FA fractions with relatively large variance explained by genetic variation. There are undoubtedly other loci with smaller effects waiting to be detected. Our study was conducted in cohorts where food insecurity and nutrition may be suboptimal for the infant, which may affect reproducibility of the findings in higher income cohorts, although the population mean %AA in both the PROVIDE and the CRYPTO cohorts was higher than published mean values from a global meta-analysis of 84 studies and was higher than several included studies from high-income countries.50 Finally, we used publicly available cis-eQTL data that were generated from adult tissue specimens not derived from lactating women.

In summary, we have performed the first GWAS on breast milk composition, identified FA fractions influenced by genetic variation and made some important observations about the effect of genetic variation at the critical FADS1/2/3 locus. Our results confirm the idea that breast milk composition is influenced by maternal genetics as well as diet and body composition and
extends the list of phenotypes possibly under selection at the FADS1/2/3 locus.

Acknowledgements
We thank all of the PROVIDE and Cryptoripidiosis Study families for their continuing support of our research. We acknowledge Kathryn Auckland, Deepthi Gursadan, Manj Sandhu and Adrian Hill for work on VacciGene consortium design, methods and funding. We acknowledge two reviewers whose comments improved the discussion of results in this manuscript.

Contributors
JCM conceptualised the study, performed statistical genetic analyses and wrote the paper. UN contributed to paper drafts. UN, ERC, DZ and SA performed data curation, interpretation and analysis. AJM supervised genotyping and quality control. WAPJ, ASGF, RH and BDK administered the study cohort, obtained funding and reviewed the paper. MA and TA oversaw the study clinical operations and data collection. TC performed GWAS genotyping quality control analysis.

Funding
This work was supported by the Bill and Melinda Gates Foundation (OPP1017093 and OPP1100514 to WAPJ), the National Institutes of Health (AI043596 to WAPJ) and the Wellcome Trust (106289/Z/14/Z Clinical Research Training Grant to AJM and 098051 Sanger Institute Core Funding Grant).

Competing interests
None declared.

Patient consent
Obtained.

Ethics approval
Ethical Review Committee for human subjects protection and Research Review Committee for scientific merit at the International Centre for Diarrhoeal Diseases Research, Bangladesh (ICDDR,B) and the Institutional Review Boards at the University of Virginia and University of Vermont.

Provenance and peer review
Not commissioned; externally peer reviewed.

Data sharing statement
The authors are happy to assist with reasonable efforts to replicate the results in this manuscript. The authors are willing to share a limited number of individual SNP summary results from these analyses that are not present in the manuscript or supplementary tables but due to identifiability concerns, complete GWAS summary statistics will require IRB approval and a data use agreement. Submission to dbGaP is in process for the PROVIDE study genotype data, accession phs001478. Other requests for individual level genetic or phenotypic data will require IRB approval and a data use agreement for a research proposal that is consistent with the consents of the individual studies.

Open Access
This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: http://creativecommons.org/licenses/by/4.0/

REFERENCES
1 Ueasy R, Mena P, Rojas C. Essential fatty acids in early life: structural and functional role. Proc Nutr Soc 2000;59:3–15.
2 Sauerwald TU, Hachey DL, Jensen CL, Chen H, Anderson RE, Heid WC. Intermediates in endogenous synthesis of C22:6 omega 3 and C20:4 omega 6 by term and preterm infants. Pediatr Res 1997;41:183–7.
3 Makrides M, Neumann M, Simmer K, Pater J, Gibson R. Are long-chain polyunsaturated fatty acids essential nutrients in infancy? Lancet 1995;345:1463–8.
4 Lane E, Demmelmaier H, Kolektzko B. Perinatal supply and metabolism of long-chain polyunsaturated fatty acids: Importance for the early development of the nervous system. Am J Perinatol 2002;19:299–310.
5 Lin YH, Llanos A, Mena R, Ueasy R, Salem N, Pawlosky RJ. Compartamental analyses of 2HS-alpha-linolenic acid and C-18-eicosapentaenoic acid toward synthesis of plasma labeled 22:6n-3 in newborn term infants. Am J Clin Nutr 2010;92:284–93.

9
Complex traits

Wolters, M., Dering, C., Siani, A., Russo, P., Kaprio, J., Risé, P., Moreno, L.A., De Henauw, S., Mehlig, K., Veidebaum, T., Molnár, D., Tomaritis, M., Iacovelli, L., Ptsiladis, Y., Galli, C., Foraita, R., Börnhorst, C., IDEFS-IC and the Family consortia. The role of a FADS1 polymorphism in the association of fatty acid blood levels, BMI and blood pressure in young children: Analyses based on path models. PloS One 2017;12:e0161485.

Martinielli, N., Girelli, D., Malegerba, G., Guarrini, P., Illig, T., Trabetti, E.,Sandri, M., Friso, S., Pizzoli, E., Schaeffer, L., Heinrich, J., Pignatti, P.F., Crocchietti, R., Olivier, O. FADS genotypes and desaturation activity estimated by the ratio of arachidonic acid to linoleic acid are associated with inflammation and coronary artery disease. Am J Clin Nutr 2008;88:941–9.

Roche, K., Fullston, J.C., Abdelmagid, S., Nielsen, D., Badawi, A., El-Sohemy, A., Ma, D.W., Mitch, D.M. Variation in the FADS5/12 gene cluster alters plasma n-6 PUFA and is weakly associated with hsCRP levels in healthy adults. Prostaglandins Leukot Essent Fatty Acids 2013;89:257–63.

Zietemann, V., Kröger, J., Enzenbach, C., Jansen, E., Fritsche, A., Weikert, C., Boeing, H., Schulze MB. Genetic variation of the FADS1 FADS2 gene cluster and n-6 PUFA composition in erythrocyte membranes in the European Prospective Investigation into Cancer and Nutrition-Potsdam study. Br J Nutr 2010;104:1748–59.

Kie L., Innis, S.M. Genetic variants of the FADS1 FADS2 gene cluster are associated with altered (n-6) and (n-3) essential fatty acids in plasma and erythrocyte phospholipids in women during pregnancy and in breast milk during lactation. J Nutr 2008;138:2222–8.

Caspi, A., Williams, B., Kim-Cohen, J., Craig, I.W., Milne, B.J., Poulton, R., Schalkwyk, L.C., Taylor, A., Werts, H., Moffitt, T.E. Moderation of breastfeeding effects on the IQ by genetic variation in fatty acid metabolism. Proc Natl Acad Sci U S A 2007;104:18860–5.

Morales, E., Bustamante, M., Gonzalez, J.R., Guexen, M., Torrent, M., Mendez, M., Garcia-Estebaranz, R., Julvez, J., Vrijheid, M., Molto-Puigmarti, C., Lopez-Sabater, C., Evestil, X. and Sunyer, J. Genetic variants of the FADS gene cluster and ELOVL gene family, colocalise LC-PUFA levels, breastfeeding, and child cognition. PLoS One 2011;6:e17181.

Kirkpatrick, B.D., Colgate, E.R., Mychaleckyj, J.C., Haque, R., Dickson, D.M., Carmelli, M.P., Nayak, U., Taniuchi, M., Nayak, U., Taniuchi, M., Dickson, D.W., Wender, W., Zaman, K., Hippe, E., Alm, M., Kirkpatrick, B.D., Petri, W.A. PROVIDE Study Team. The “performance of Rotavirus and Oral Polio Vaccines in Developing Countries” (PROVIDE) study: description of methods of an interventional study designed to explore complex biologic problems. Am J Trop Med Hyg 2015;92:744–51.

Mychaleckyj, J.C., Haque, R., Carmelli, M., Zhang, D., Colgate, E., Nayak, U., Taniuchi, M., Dickson, D.W., Wender, W., Zaman, K., Hippe, E., Alm, M., Kirkpatrick, B.D., Petri, W.A. Effect of substituting IPV for tOPV on immunity to poliovirus in Bangladesh infants: An open-label randomized controlled trial. Vaccine 2016;34:358–66.

Nayak, U., Kanungo, S., Zhang, D., Ross, C., Colgate, E., Carmelli, M.P., Dey, A., Alm, M., Manna, B., Nandy, R.K., Kim, D.R., Paul, D.K., Choudhury, S., Sahu, S., Harris, W.S., Wierzb, T.F., Ahmed, T., Kirkpatrick, B.D., Haque, R., Petri, W.A, Mychaleckyj, J.C. Influence of maternal and socioeconomic factors on breast milk fatty acid composition in urban, low-income families. Matern Child Nutr 2017;13:e12423.

Harris, W.S., Potalla, J.V., Vasan, R.S., Larson, M.G., Robins, S.J. Changes in erythrocyte membrane trans and marine fatty acids between 1999 and 2006 in older Americans. J Nutr 2012;142:1297–303.

Jackson, KH., Poliak, J., Sarabon, L., Chaima, D., Harris, W.S. Analysis of breast milk fatty acid composition using dried milk samples. Int Breastfeed J 2016;11:11:1.

Delanoue, A., Zagury, J.F., Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 2013;10:5–6.

Howie, B.N., Donnelly, P., Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PloS Genet 2009;5:e1000529.

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