Human milk variation is shaped by maternal genetics and impacts the infant gut microbiome

Graphical abstract

Highlights

- Human milk transcriptomes reveal factors shaping the lactating mammary gland

- eQTL analysis identified tissue-shared and milk-specific genetic influences

- Milk eQTLs associated with milk composition and maternal breast cancer risk

- Milk with a signature of inflammation correlated with the infant gut microbiome
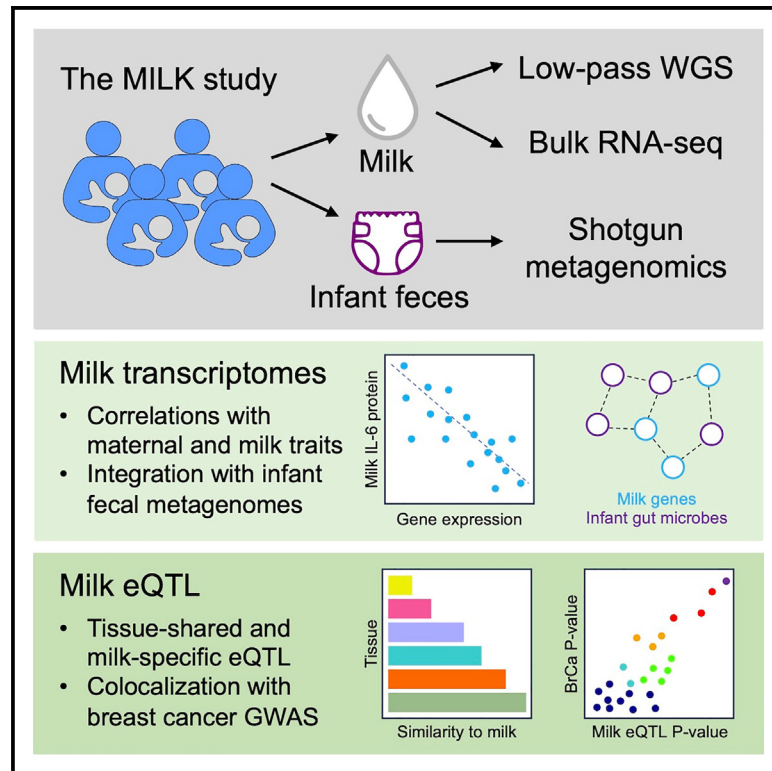
Authors

Kelsey E. Johnson, Timothy Heisel, Mattea Allert, ..., Frank W. Albert, Ellen W. Demerath, Ran Blekhman

Correspondence

kej@umn.edu

In brief

Human milk has positive health impacts for lactating parents and infants, and milk composition varies across individuals. Johnson et al. perform an eQTL study of human milk cells and link the milk transcriptome to maternal traits, milk composition, and the infant gut microbiome.

Resource

Human milk variation is shaped by maternal genetics and impacts the infant gut microbiome

Kelsey E. Johnson,[1,14,*] Timothy Heisel,[2] Mattea Allert,[1] Annalee Fürst,[3] Nikhila Yerabandi,[3] Dan Knights,[4,5] Katherine M. Jacobs,[6] Eric F. Lock,[7] Lars Bode,[3,8] David A. Fields,[9] Michael C. Rudolph,[10] Cheryl A. Gale,[2] Frank W. Albert,[1,13] Ellen W. Demerath,[11,13] and Ran Blekhman[12,13]

[1]Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN, USA
[2]Division of Neonatology, Department of Pediatrics, University of Minnesota Medical School, Minneapolis, MN, USA
[3]Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA
[4]BioTechnology Institute, College of Biological Sciences, University of Minnesota, Minneapolis, MN, USA
[5]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA
[6]Department of Obstetrics, Gynecology and Women's Health, Division of Maternal-Fetal Medicine, University of Minnesota Medical School, Minneapolis, MN, USA
[7]Division of Biostatistics & Health Data Science, University of Minnesota School of Public Health, Minneapolis, MN, USA
[8]Human Milk Institute (HMI) and Mother-Milk-Infant Center of Research Excellence (MOMI CORE), University of California, San Diego, La Jolla, CA, USA
[9]Department of Pediatrics, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA
[10]Harold Hamm Diabetes Center, Department of Physiology, the University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA
[11]Division of Epidemiology and Community Health, University of Minnesota School of Public Health, Minneapolis, MN, USA
[12]Section of Genetic Medicine, Division of Biological Sciences, University of Chicago, Chicago, IL, USA
[13]Senior author
[14]Lead contact
*Correspondence: kej@umn.edu
https://doi.org/10.1016/j.xgen.2024.100638

SUMMARY

Human milk is a complex mix of nutritional and bioactive components that provide complete nourishment for the infant. However, we lack a systematic knowledge of the factors shaping milk composition and how milk variation influences infant health. Here, we characterize relationships between maternal genetics, milk gene expression, milk composition, and the infant fecal microbiome in up to 310 exclusively breastfeeding mother-infant pairs. We identified 482 genetic loci associated with milk gene expression unique to the lactating mammary gland and link these loci to breast cancer risk and human milk oligosaccharide concentration. Integrative analyses uncovered connections between milk gene expression and infant gut microbiome, including an association between the expression of inflammation-related genes with milk interleukin-6 (IL-6) concentration and the abundance of *Bifidobacterium* and *Escherichia* in the infant gut. Our results show how an improved understanding of the genetics and genomics of human milk connects lactation biology with maternal and infant health.

INTRODUCTION

Lactation is a defining trait of mammals and has been essential for our species throughout human evolution.[1] Today, breastfeeding is recommended as the exclusive mode of feeding for infants, given its documented health benefits for both mothers and infants.[2] The nutritional significance of human milk stems from hundreds of milk constituents, including macro- and micro-nutrients, immune factors, hormones, oligosaccharides, and microbes.[3] Maternal factors such as diet, health status, and genetics shape variation in milk composition across lactating women[4,5]; however, the role of maternal genetics in shaping milk composition is particularly understudied. A small number

of studies suggest important relationships between maternal genotype, milk composition, and infant health.[6] For example, maternal secretor status, determined by the *FUT2* gene, is linked to human milk oligosaccharide (HMO) composition.[7] HMOs are sugars in human milk that cannot be digested by the infant but promote the growth of beneficial microbes in the infant gut and may provide additional immunological and metabolic benefits.[8] In addition to HMOs, variation in other milk components, such as fatty acids, has been linked to the infant gut microbiome,[9,10] and breastfeeding (vs. formula feeding) is one of the strongest factors shaping the infant gut microbiome.[11,12] The abundance of certain microbes in the infant gut, particularly *Bifidobacterium*, has been linked to health outcomes in infancy and later
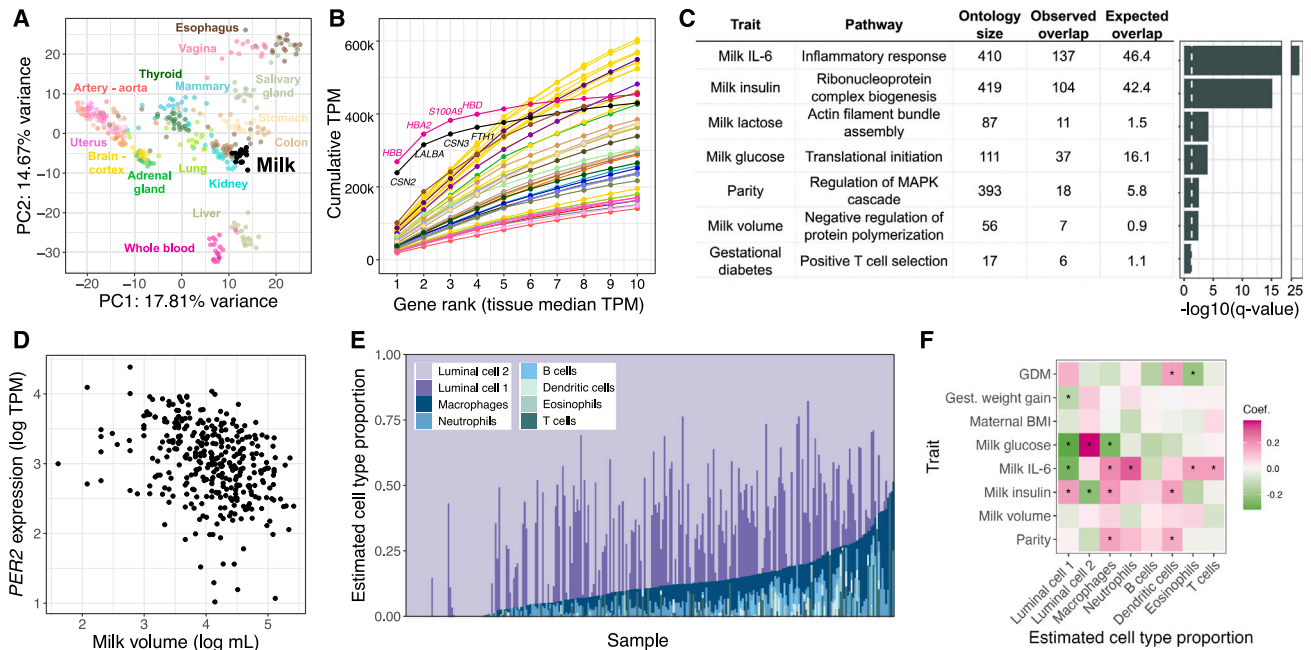
Figure 1. Overview of gene expression in human milk

(A) Principal-component analysis of transcriptomes from a subset of GTEx tissues and milk. 19 random samples were chosen from each tissue. PCs were calculated using the 1,000 most variable genes within GTEx, and then milk samples were projected onto the GTEx samples. An equivalent plot including all GTEx tissues is shown in Figure S5.

(B) Cumulative TPM (transcripts per million) of the top 10 genes by median TPM for milk and GTEx tissues. The color scheme is the same as in (A).

(C) Gene Ontology enrichment of genes with expression correlated to maternal and milk traits. The most significant term for each trait is shown (STAR Methods). The dashed white vertical line denotes a q value of 0.05.

(D) Correlation between milk volume (from standardized electric breast pump expression during a study visit; STAR Methods) and *PER2* gene expression in milk.

(E) Cell type proportion estimates generated using Bisque[27] for transcriptomes from this study with reference milk single-cell RNA-seq from Nyquist et al.[17]

(F) Heatmap of regression coefficients between estimated cell type proportions (x axis) and maternal or milk traits (y axis) from a linear model including technical covariates (STAR Methods). *q < 10%.

See also Figure S1 and Tables S2, S3, S5, and S7.

childhood.[13] Thus, the composition of the infant gut microbiome represents a key outcome through which human milk promotes infant health. Here, we combine maternal clinical and milk composition data with maternal whole-genome sequences, milk transcriptomes, and infant fecal metagenomics to characterize genetic influences on gene regulation in milk and identify pathways linking milk gene expression with milk composition and infant gut health. The results advance our knowledge of the complex molecular and physiological relationships connecting mother, milk, and infant.[14]

RESULTS

Milk gene expression correlates with maternal traits and milk composition in a healthy, successfully lactating cohort

Human milk contains mammary epithelial luminal cells and a variety of immune cell types, including macrophages, lymphocytes, and granulocytes.[15–19] A milk sample provides rich information on immune phenotypes and the biology of milk production, as RNA extracted from milk profiles the milk-producing cells in the lactating mammary gland.[15,16,20,21] To characterize population-level variation in human milk gene expression, we

performed bulk RNA sequencing on cell pellets from 1-month postpartum milk samples from 316 women in the Mothers and Infants Linked for Healthy Growth (MILK) study[22–24] (Figures S1, S2, S3, and S4; Table S1). Comparison to gene expression data from human tissues obtained by the Genotype-Tissue Expression (GTEx) consortium[25] showed that milk expression profiles clustered near other secretory tissues, such as pancreas, kidney, and colon (Figures 1A and S5). The three most highly expressed milk genes (*CSN2*, *LALBA*, and *CSN3*), which comprise a large proportion of milk transcripts,[15] accounted for 34.5% of protein-coding transcripts in milk, reminiscent of the preponderance of hemoglobin transcripts typical in whole blood (Figure 1B).[25] These three genes encode the major milk proteins beta- and kappa-casein (*CSN2* and *CSN3*) and lactalbumin (*LALBA*), an essential protein for lactose and HMO synthesis.[26]

To identify factors associated with the milk transcriptome, we tested for correlations between the expression of 12,006 genes in milk and 13 maternal or milk traits in n = 269 participant's milk samples (or n = 171 for milk macronutrients; Tables S2, S3, and S4; Figures S6, S7, and S8). In this analysis, we used a gene-wise model testing for differences in each gene's expression to maternal or milk traits and technical covariates

(STAR Methods). Milk composition traits were measured from separate aliquots of the same milk samples as used for RNA sequencing (RNA-seq) (STAR Methods). Among maternal traits, gestational diabetes status and parity were correlated with expression of the most genes (gestational diabetes: 784 genes, parity: 172 genes at q < 10%; negative binomial generalized log-linear test; STAR Methods). Genes for which expression correlated with parity were enriched for pathways related to cell communication and the mitogen-activated protein kinase cascade, potentially reflecting persistent differences in mammary gland epigenetic states and remodeling during lactation in participants who had lactated previously[28,29] (Figure 1C). Pre-pregnancy BMI and gestational weight gain, traits associated with delayed lactogenesis and breastfeeding challenges,[30] were correlated with milk expression of just a few genes (<30 genes; Table S3). This weak relationship could be due to our study's inclusion of only women who successfully breastfed for at least 1 month postpartum, thus excluding participants with difficulties initiating breastfeeding related to metabolic health. Milk concentrations of IL-6, glucose, insulin, and lactose and the total single breast milk expression volume produced at the study visit were each correlated with expression of hundreds of genes (q < 10%; Table S3). These milk trait-correlated genes were enriched for processes such as translation (milk insulin) and cytoskeleton organization (milk volume) (Figure 1C; Table S5). There was no significant interaction with maternal obesity status for any gene/trait pair after multiple test correction (STAR Methods; Table S6).

The gene for which expression was most significantly associated with expressed milk volume was the core circadian clock gene *PER2*. Higher *PER2* expression correlated with lower milk volume (log$_2$ fold change = −0.22, q = 9.5 × 10^{-9}; Figure 1D; Table S3). The relationship between *PER2* expression and milk volume was not driven by the time of day of milk expression (F test, *p* = 0.06; Figure S9; STAR Methods). It is notable that we observed this correlation even though milk volume is variable within individuals[31] and was assessed in a single visit (STAR Methods). In addition to *PER2*, the circadian gene *RORC* was also associated with milk volume (log$_2$ fold change = −0.10, q = 0.03). *PER2* plays a role in cell fate and ductal branching in the mammary gland in addition to its circadian function.[32] Our observation suggests that differential expression of circadian clock genes in the mammary gland affects milk production in humans, possibly via regulation of milk production genes or by anatomical changes in the breast during lactogenesis.

Of all milk traits tested, glucose concentration was correlated with expression of the largest number of genes (1,634 genes at q < 10%; Table S3), followed by IL-6 protein and insulin concentrations (1,235 and 1,144 genes at q < 10%, respectively). Genes correlated with insulin and glucose concentrations were both strongly enriched for ribosomal proteins. Genes correlated with milk IL-6 concentration were enriched for immune pathways, with "inflammatory response" the most significantly enriched pathway (q = 4.1 × 10^{-27}, Fisher's exact test; Figure 1C), consistent with IL-6's role as a marker of inflammation in the mammary gland.[33] To estimate the contributions of different cell types to our milk bulk transcriptomes, we performed cell-type deconvolution using a milk single-cell RNA-seq reference panel (Figure 1E; STAR Methods).[17,27] Consistent with previous studies, mammary epithelial cells were estimated to make up the majority of cells.[17–19,34] The estimated proportion of several immune cell types were increased in milk samples with higher IL-6 concentration (e.g., neutrophils: multiple regression coefficient = 0.29, q = 3.4 × 10^{-4}; macrophages: multiple regression coefficient = 0.22, q = 6.2 × 10^{-3}; Figure 1F; Table S7), suggesting that the relationship between IL-6 concentration and immune gene expression is linked to a greater proportion of immune cells in milk.

Genetic influences on gene expression in human milk

Associations between genetic variation and gene expression can illuminate the molecular mechanisms underlying genetic influences on human traits,[35] but this approach has not been applied to human milk. To identify associations between maternal genetic variation and milk gene expression, we generated low-pass whole-genome sequencing data and performed an expression quantitative trait locus (eQTL) scan in 230 unrelated human milk samples (STAR Methods). We identified a local eQTL (q < 5%) at 2,790 genes of 17,302 tested (Table S8; Figures S10, S11, and S12), with 45 genes showing evidence of multiple independent signals in conditional analysis (Table S9). Comparing milk eQTLs to those identified in 45 human tissues in the GTEx project,[25] we partitioned our eQTLs as milk specific (n = 482) or shared with at least one other tissue (n = 2,308) by detecting milk-specific eQTL effects via statistical colocalization[36,37] (Figure 2A; Table S10; STAR Methods). Genes with milk-specific eQTLs highlighted key biological pathways in the lactating mammary gland: production of caseins (e.g., the abundant milk proteins *CSN3* and *CSN1S1*), lactose synthesis (*LALBA*), lipogenesis (e.g., *ACSL1*, *LPL*, *IDH1*, and *LPIN1*), hormonal regulation (*INSR*), and immunity (e.g., *LYZ*, *MUC7*, and *CD68*) (Table S10). In addition, genes with milk-specific eQTLs were twice as likely as genes with eQTLs shared across multiple tissues to overlap genetic associations for milk traits in dairy cattle (odds ratio = 2.0, *p* = 1.7 × 10^{-4}, two-sided Fisher's exact test; Figure 2B; Table S11), a species for which there is far more known about genetic influences on lactation than in humans. This enrichment suggests that genes with milk-specific eQTLs are specifically important for milk biology. Genes with milk-specific eQTLs also tended to have more sequence-level constraint[38] than tissue-shared eQTLs (*p* = 2.4 × 10^{-6}, Wilcoxon rank-sum test; Figure 2C) and were enriched for pathways such as "regulation of ERK1 and ERK2 cascade" (Figure 2D; STAR Methods), which has a key role in mammary morphogenesis.[39]

To identify tissues for which genetic regulation of gene expression is most similar to milk, we estimated the proportion of shared eQTLs between milk and each GTEx tissue using *mash*[40] (STAR Methods; Table S12). Milk shared the largest proportion of eQTLs with secretory tissues (e.g., minor salivary gland, pancreas, and esophagus), with a higher proportion shared than that observed for non-lactating breast tissue (Figures 2E and S13). These comparisons highlight the shared regulation of gene expression across secretory tissues and underscore the insufficiency of non-lactating breast tissue for studying gene expression programs necessary for lactation.

Epidemiological studies describe a complex relationship between lactation and breast cancer risk, with decreased or
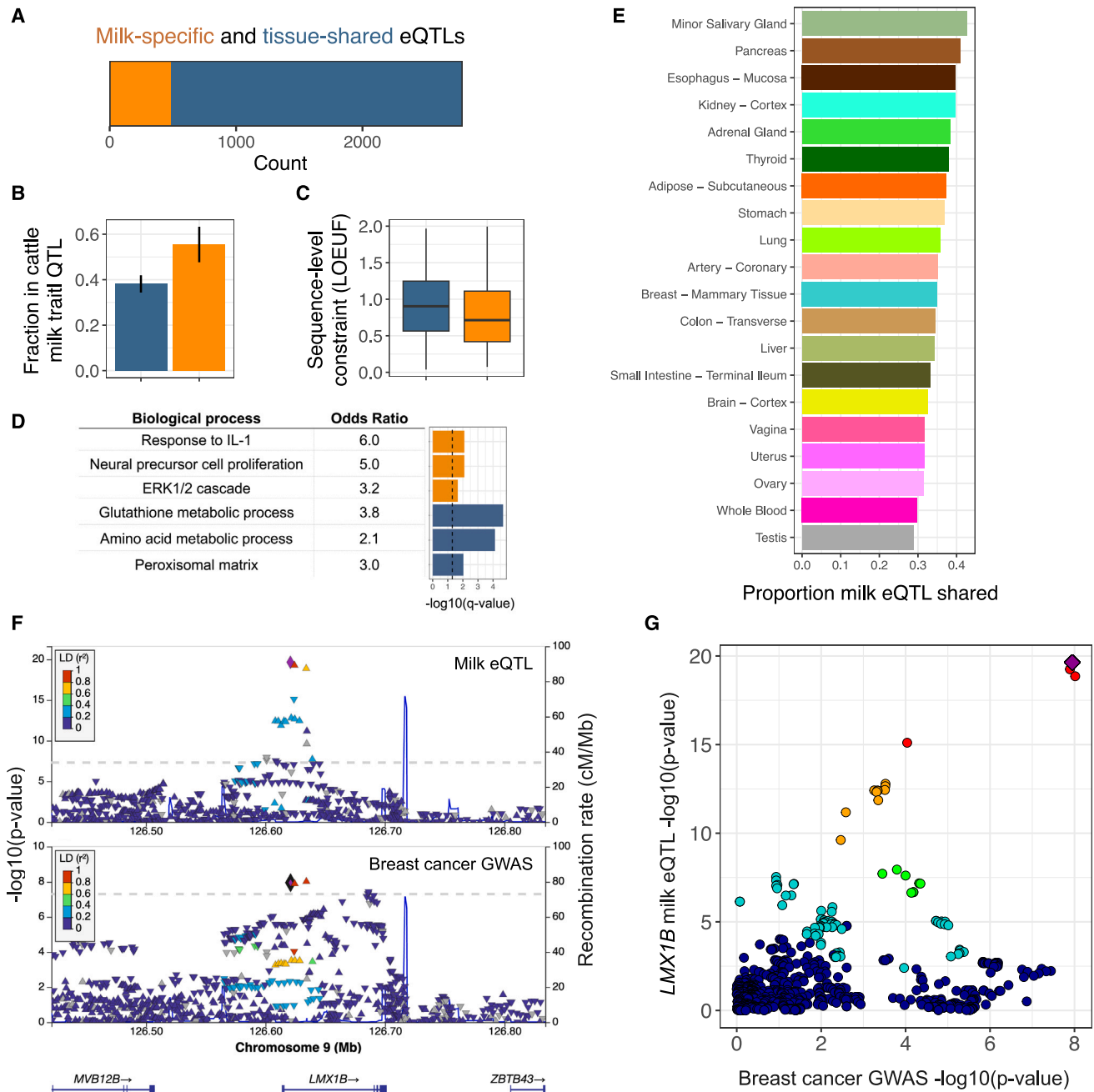
Figure 2. Genetic influences on gene expression in human milk

(A) Counts of genes with milk-specific eQTLs (orange, genes with an eQTL signal that did not colocalize with any GTEx tissue; STAR Methods) vs. tissue-shared eQTLs (blue, genes with all milk eQTL signals colocalized with at least one GTEx tissue).

(B) Fraction of genes in each category that overlapped with a milk trait QTL in the dairy cattle genome. Error bars represent a 95% confidence interval.

(C) Distributions of sequence-level constraint, measured by the loss-of-function observed/expected upper bound fraction statistic.[38]

(D) Enriched Gene Ontologies for genes with milk-specific (orange) or tissue-shared (blue) eQTLs. The dashed vertical line denotes a q value of 5%.

(E) Fraction of shared milk eQTLs with a subset of GTEx tissues, estimated with *mash*.[40]

(F) LocusZoom genetic associations in the *LMX1B* region with milk gene expression (top) and breast cancer risk (bottom). Each data point represents a SNP, plotted by its chromosomal location (x axis) and significance of association (y axis), with colors corresponding to linkage disequilibrium (r^2) to the lead SNP for the milk eQTL, shown as a purple diamond.

(G) Each point is a variant, plotted by the strength of association with milk gene expression (y axis) and breast cancer risk (x axis). Colors are the same as in (F), top, with a purple diamond representing the lead milk eQTL SNP. The pattern of variants in the top right suggests a shared underlying causal variant.

See also Figures S13, S14, S15, S16, S17, S18, S19, and S20 and Tables S8, S9, S10, S11, S12, and S13.

increased risk depending on age at first pregnancy and decreased lifetime risk associated with longer duration of lactation.[41,42] Because the genetics of gene expression in the lactating mammary gland is distinct from that of non-lactating breast (Figure 2E), milk eQTLs provide unique functional annotations to genetic associations with breast cancer. Using colocalization analyses between all milk eQTLs and breast cancer genome-wide association study (GWAS) loci,[43] we identified 7 loci with strong evidence of a shared causal variant (posterior probability of shared causal variant >0.9; Table S13; Figures S14, S15, S16, S17, S18, and S19). Of these milk eQTL-GWAS colocalizations, 4 had been nominated previously as a causal gene for breast cancer,[44–46] and 2 were eQTLs for pseudogenes (Table S13). We identified a novel candidate gene at a breast cancer GWAS locus where a milk eQTL that increased expression of *LMX1B* was associated with increased cancer risk (Figures 2F and 2G). *LMX1B* does have not have a significant GTEx eQTL in mammary tissue.[25] The milk *LMX1B* eQTL colocalized with one GTEx tissue at an eQTL for the tibial nerve (Figure S20). *LMX1B* is a transcription factor essential for normal development of limbs, kidneys, and ears.[47]

Milk gene expression correlates with concentrations of HMOs

Maternal genetics play a strong role in shaping the concentration of HMOs,[7] sugars in milk that are not digested by the infant but promote the growth of beneficial microbes in the infant gut. HMOs are synthesized in the mammary gland by addition of monosaccharides to a lactose molecule, but the glycosyltransferases catalyzing these reactions are largely uncharacterized.[48] Secretor status, determined by the absence of a common nonsense variant in the fucosyltransferase 2 (*FUT2*) gene, strongly predicts the concentration of certain HMOs, with the presence of some HMOs entirely determined by secretor status.[7] Utilizing 310 participants with both milk gene expression and 1-month HMO composition data, we observed distinct HMO profiles between secretors and non-secretors (Figures 3A and S21; see Table S14 for HMO definitions). We hypothesized that, beyond the strong effects of the secretor polymorphism, the expression of *FUT2* in milk would correlate with HMO concentrations within secretor individuals, reflecting variation in milk among women with a functional FUT2 enzyme. We observed nominally significant associations between *FUT2* expression and the concentration of three HMOs: 2′-fucosyllactose (beta = 0.12, p = 0.01; Figure S22), lacto-N-fucopentaose (LNFP)-II (beta = −0.12, p = 0.03; Figure S22), and lacto-N-hexaose (beta = 0.14, p = 0.04; Figure S22). This suggested that milk gene expression data could be useful for identifying critical genes for HMO biosynthesis. We tested for pairwise correlations between gene expression and 19 individual HMOs and the sums of all HMO concentrations, sialylated HMOs, and fucosylated HMOs while controlling for secretor status (STAR Methods). These 22 HMO traits were significantly correlated with expression of between 8 and 1,262 genes (q < 10%; Table S15), including known HMO biosynthesis genes, such as the sialyltransferase *ST6GAL1*,[48] with the HMO sialyl-lacto-N-tetraose c (LSTc) (beta = 0.80, p = 6.6 × 10^{-8}, q = 1.5 × 10^{-4}; Figure 3B). The genes correlated with 6 of the HMO traits were enriched for

pathways related to ribosomes, such as "cytosolic ribosome" enriched in genes correlated with the sum of all HMOs (Figure 3C; Table S16). Genes correlated with the HMO 6′-sialyllactose or the sum of sialylated HMOs were enriched for inflammation-related pathways such as "cytokine activity" (Table S16), consistent with previous evidence that sialylated HMOs were more abundant in women with mastitis compared to healthy women.[49]

HMO biosynthesis represents an ideal system to understand the effects of maternal genetics on milk composition via changes in gene expression, as gene expression from the relevant cell type (mammary epithelial cells) and HMO concentrations can be measured non-invasively in the same milk samples. Among 54 candidate glycosyltransferase genes,[48] seven genes had significant milk eQTLs in our data (q < 5%; Table S17), which we used to test for associations between maternal genotypes at milk eQTL tag SNPs and HMO concentrations in 224 individuals with both data types. For three genes, we observed an association between genotype and between 1 and 13 HMOs (Table S18; q < 10%). These included the known association of *FUT2* with 13 HMOs (e.g., LNFP-I; Figure 3D) and an association between *GCNT3* and fucosyllacto-N-hexaose (FLNH) (Figure 3E). *GCTN3* was also linked to FLNH in our above analysis of correlations between gene expression and HMO concentrations (Table S15; Figure S23). *GCTN3* has been identified previously as the best candidate gene responsible for the addition of a β-1,6-linked N-acetylglucosamine to the lactose core, a step required for the biosynthesis of FLNH.[48] For each eQTL-HMO pair (q < 10%), we then estimated the causal effect of modified gene expression on HMO concentration using a Wald ratio test (Figure 3F; Table S18). These results provide evidence of direct or indirect roles of specific glycosyltransferases in HMO biosynthesis in the lactating mammary gland.

Milk gene expression is associated with the infant gut microbiome

Studies have found correlations between milk composition and variation in the infant gut microbiome.[9,10,50,51] However, it is unclear how these correlations are shaped by maternal genetics and milk gene regulation. We hypothesized that, given milk gene expression reflects milk composition, it could be correlated with the infant gut microbiome. We profiled the fecal microbiome of infants in our study with metagenomic sequencing at 1 and 6 months postpartum (n = 146; Figures 4A and S24) and identified nine correlated sets of genes expressed in milk and microbial taxa or pathways present in the infant gut at 1 or 6 months postpartum using sparse canonical correlation analysis (CCA)[52,53] (STAR Methods; Figure 4B; Table S19). Using pathway enrichment analysis, we identified relevant biological processes in these milk-expressed gene sets correlated with the infant fecal microbiome (Table S20). For example, milk expression of lysosome genes was negatively correlated with the abundance of microbial genetic pathways related to amino acid degradation in the infant gut at 6 months (Figure 4C), and expression of fatty acid metabolism genes in milk was positively correlated with the abundance of species of *Bifidobacterium* in the infant gut at 1 month (Figure 4D). Lysosomes are involved in mammary gland remodeling and involution,[54,55] and human milk fats can act as prebiotics to support growth of commensal
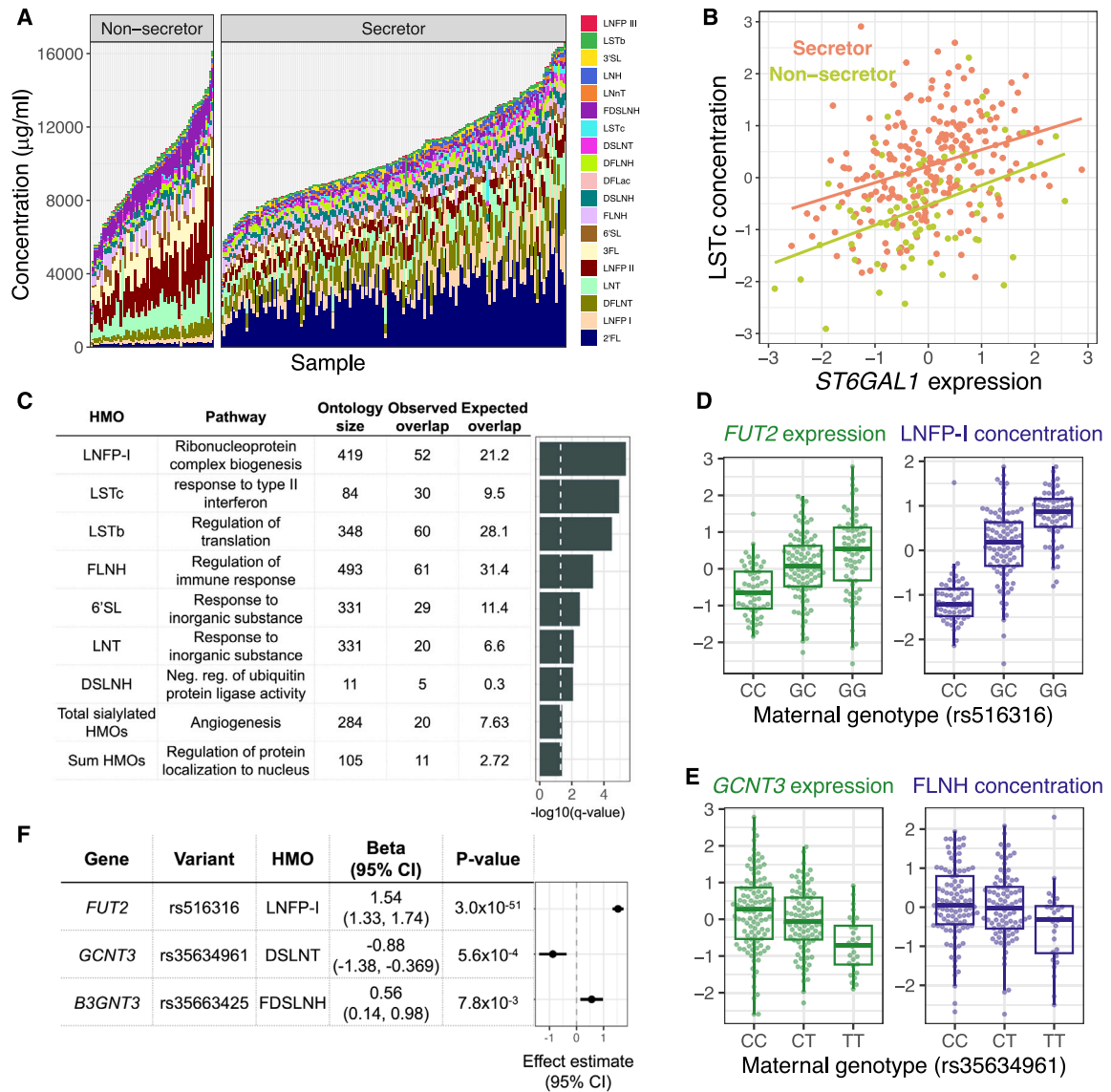
Figure 3. Effects of milk gene expression on HMO composition

(A) HMO concentration (y axis) profiles for milk samples in our study (x axis), grouped by secretor status.

(B) Correlation between *ST6GAL1* gene expression in milk and normalized LSTc concentration, colored by secretor status (log$_2$ fold change = 0.32, p = 6.6 × 10^{-8}, q = 1.5 × 10^{-4}).

(C) Gene Ontology enrichment of genes with expression correlated to a single HMO or HMO category. The most significant term for each HMO is plotted. The dashed vertical line denotes a q value of 5%.

(D) Relationships between genotype at the lead SNP at the *FUT2* eQTL and *FUT2* expression in milk (green) or LNFP-I concentration (purple). LNFP-I concentrations are residuals after correcting for genetic PCs (STAR Methods).

(E) Relationships between genotype at the lead SNP at the *GCNT3* eQTL and *GCNT3* expression in milk (green) or FLNH concentration (purple). FLNH concentrations are residuals after correcting for secretor status and genetic PCs (STAR Methods).

(F) Estimates of the effect of milk gene expression of candidate HMO biosynthesis pathway genes on the abundance of HMOs from a Wald ratio test. Some genes had significant effects on more than one HMO (Table S18). The most significant HMO for each gene is plotted here.

See also Figures S21, S22, and S23 and Tables S14, S15, S16, S17, and S18.

bacteria in the infant gut, including *Bifidobacterium*.[56] These links between milk gene expression and the infant gut microbiome nominate biological pathways through which normal variation in human milk composition may influence the infant gut microbiome.

The sparse CCA algorithm identified species of *Escherichia* at 6 months in the infant gut as negatively correlated with milk-expressed genes in the Janus kinase (JAK)-signal transducer and activator of transcription (STAT) pathway, which is a key regulator of both milk production and mammary inflammation.[57]
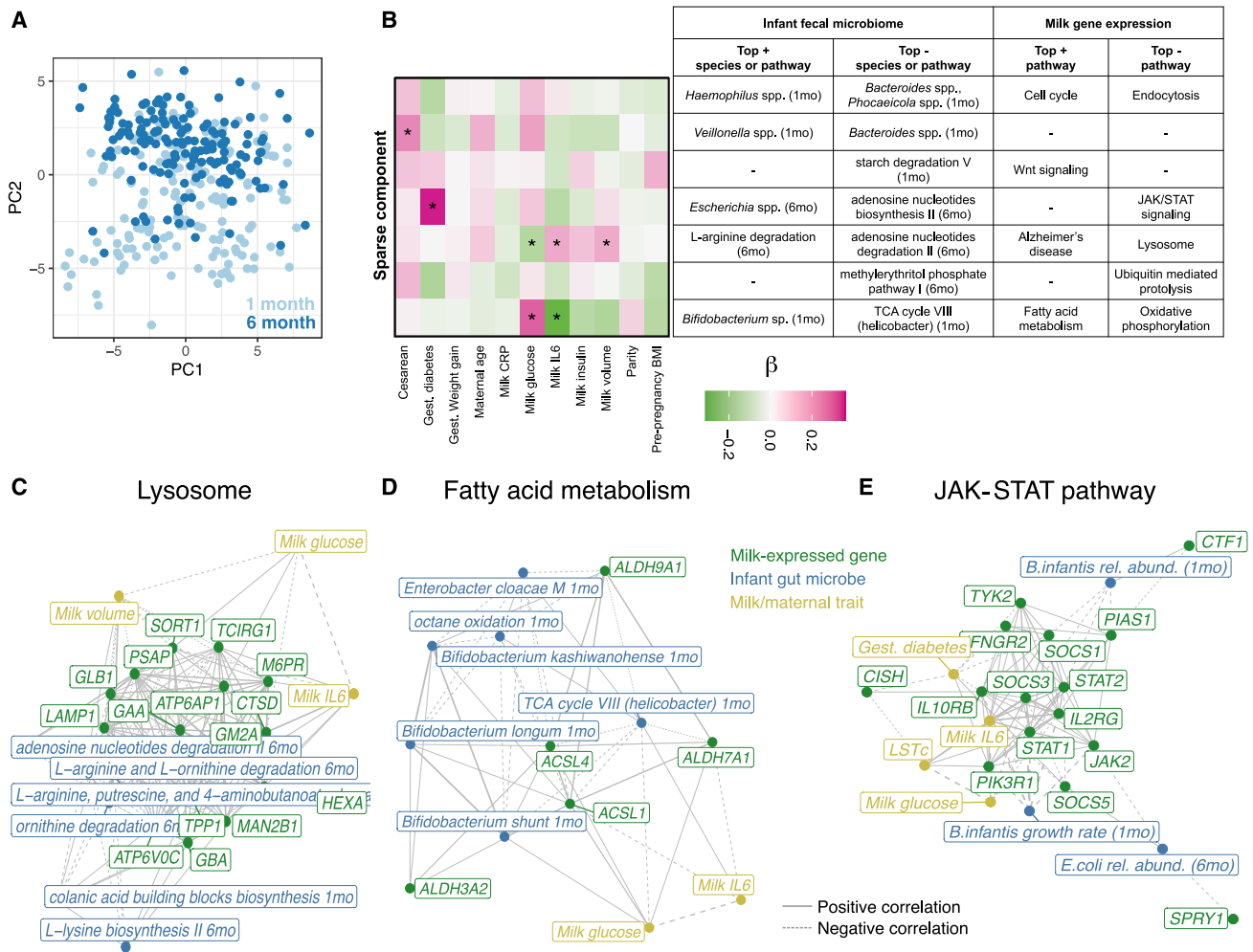
Figure 4. Interactions between milk gene expression and the infant fecal microbiome

(A) Principal-component analysis of infant fecal microbiome metagenomic data, summarized at the taxonomic level, with each point representing a fecal sample and colors representing infant age (light blue, 1 month; dark blue, 6 months).

(B) Sparse CCA integrating milk host gene expression and infant fecal microbial species or microbial genetic pathway relative abundance (at 1 or 6 months of age) identified seven significant sparse components (in rows). The heatmap on the left shows Spearman correlation coefficients between each mother/infant pair score for a given sparse component (rows) and maternal or milk traits (columns). The table lists the most highly weighted microbial taxon or genetic pathway and the most significantly enriched host gene set in milk gene expression. (+) or (−) indicates whether these features were positively or negatively weighted in the sparse component.

(C and D) Network diagrams generated using the correlation matrix of infant fecal microbial species/pathways and milk-expressed host genes within an enriched pathway for two of the sparse components in (B). Line size corresponds to the absolute value of the correlation coefficient, and line type corresponds to negative (dashed) or positive (solid) correlations. Node color signifies milk-expressed host genes (green), infant fecal microbial pathways/taxa (green), or maternal/milk traits (yellow). Plotted edges had correlation $p < 0.05$.

(E) Network diagram displaying correlations between milk IL-6 concentration, LSTc (HMO) concentration, JAK-STAT pathway genes expressed in milk, and *B. infantis* relative abundance and estimated growth rate in the infant gut at 1 month and *Escherichia coli* relative abundance at 6 months. JAK-STAT pathway genes were selected that had a significant correlation with *B. infantis* or *E. coli* abundance after multiple test correction (q < 10%).

See also Figure S24 and Tables S19, S20, and S21.

This sparse component was also correlated with gestational diabetes status (Figure 4B). We noted that the component highlighting abundance of *Bifidobacterium* in infants at 1 month was also enriched for milk-expressed genes in inflammation-related pathways (Table S20) and correlated with milk concentrations of IL-6 and glucose. *Bifidobacterium* spp. are abundant microbes in the breastfed infant gut that promote beneficial health outcomes, particularly *B. infantis*.[58,59] *Escherichia* spp. are abundant in the infant gut, at higher levels in full term (vs. preterm) infants, and increase in abundance after the introduction of solid foods.[60] Given our observation that genes in the JAK-STAT pathway were significantly correlated with milk IL-6 concentration (Table S3), we further examined the relationships between milk expression of JAK-STAT pathway genes, gestational

diabetes, milk composition, and infant fecal *Escherichia* and *Bifidobacterium* levels. Given the well-known relevance of *B. infantis* to infant health, we also computationally inferred *B. infantis* growth rates in samples from 1-month-old infants (STAR Methods), an additional aspect of microbial community dynamics that varies across individuals and is relevant to disease.[61] Both infant fecal *B. infantis* growth rate and relative abundance were negatively correlated with milk expression of JAK-STAT pathway genes, most significantly *SOCS3* (growth rate: Pearson's r = −0.52, $p = 1.4 \times 10^{-4}$; relative abundance: r = −0.19, p = 0.02; Figure 4E; Table S21). *SOCS3* encodes a key element of the mammary anti-inflammatory response to bacterial mastitis[62] and is most highly expressed in the immune cells in milk.[17] Thus, the correlation between increased JAK-STAT signaling in milk and lower *B. infantis* abundance and growth in the infant gut could be related to an immune response to infection of the mammary gland.

DISCUSSION

Here, we generated and integrated multiple omics datasets within a cohort of exclusively breastfeeding mother-infant pairs, leveraging the milk transcriptome as a readout of the biology of milk production. Our results highlight how an improved understanding of the genetics and genomics of human milk reveals connections with maternal and infant health.

A consistent theme across our results was a link between mammary inflammation-related gene expression, milk composition, and the infant gut microbiome. Milk IL-6 concentration was correlated with milk gene expression across hundreds of genes (Table S3). Genes correlated with the concentration of multiple HMOs in milk were enriched for inflammation-related pathways (Figure 3C; Table S16), and expression of inflammation-related genes in milk was inversely correlated with the abundance and growth of *Bifidobacterium* in the infant gut at 1 month and *Escherichia* at 6 months (Figure 4E). All participants in our study were exclusively breastfeeding and did not report symptoms of mastitis (infection of the mammary gland) at the time of milk collection. Subclinical mastitis is prevalent across human populations and is associated with differences in milk composition.[63–66] Thus, our results suggest that mammary inflammation, even when unnoticeable to the lactating individual, is a primary driver of variation in milk composition with potential effects on the infant gut microbiome.

Combining milk gene expression with maternal genetic variation, we identified numerous novel milk-specific eQTLs, which can now be used as targets for investigation of the effects of gene expression on milk production and composition and infant and maternal health. For example, combining our milk eQTLs with breast cancer GWAS summary statistics, we provide the first functional evidence connecting *LMX1B* expression to a nearby breast cancer GWAS locus (Figures 2F and 2G). Functional evidence for this GWAS locus had previously been missing, as this gene does not have an eQTL in GTEx mammary tissue and, thus, may only be detectable during lactation. In an analysis of single-cell RNA-seq across human tissues, *LMX1B* was most highly expressed in salivary and breast glandular cells.[67] In addition, hypomethylation at a CpG island in *LMX1B*

in human milk samples was associated with subsequent diagnosis of breast cancer in an epigenome-wide association study,[68] suggesting higher expression correlated with breast cancer risk, which is concordant with the direction of effect in our results.

The importance of breastfeeding, especially in underdeveloped countries, is widely acknowledged, but the long-term health effects in modern high-income contexts are less concrete.[2] Similarly, the causal effects of differences in milk composition for breastfed infants are underexplored due to the ethical and logistical impediments to performing randomized trials of infant nutrition. The field of human genetics has been hugely successful in identifying genetic effects on molecular and complex traits and has leveraged these associations to improve our understanding of disease pathophysiology, identify drug candidates, and interrogate causal relationships impacting human health. However, traits related to women's health generally have been overlooked by this area of research, and human milk and lactation are glaring examples of this neglect. Fortunately, milk represents an easily obtained non-invasive biospecimen, aiding our ability to close this gap. Our study provides a step toward leveraging modern human genomics techniques to characterize the factors that shape milk composition and understand how this composition impacts infant and maternal health.

Limitations of the study

While our study introduced a framework for integrating multiple and diverse data types in the mother/milk/infant triad, it is limited by sample size, particularly of our milk composition phenotypes and infant fecal microbiome data. Additionally, the MILK study is predominantly composed of participants who self-identify as white and non-Hispanic (∼85%). Thus, our analysis was limited to genetic variants common in participants of European ancestry, and our eQTL results may not be generalizable to other ancestry groups. Last, we studied mature milk collected 1 month postpartum, which did not allow us to assess genetic effects on colostrum or milk produced at other points in lactation.

RESOURCE AVAILABILITY

Lead contact
Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Kelsey Johnson (kej@umn.edu).

Materials availability
This study did not generate new unique reagents.

Data and code availability
- RNA-seq quantifications, infant fecal metagenomic abundances, HMO concentrations, milk eQTL summary statistics, and study metadata are available at figshare and are publicly available as of the date of publication. DOIs are listed in the key resources table.
- Maternal genotypes and raw RNA and DNA sequencing data have been deposited at dbGaP and are available under controlled access in compliance with the study IRB. Use of the data is limited to health/medical/biomedical purposes, including methods development and excluding the study of population origins. Data access is provided by dbGaP (https://www.ncbi.nlm.nih.gov/gap/) for certified investigators and does not require local IRB approval. Accession numbers are listed in the key resources table.

- Raw infant fecal metagenomic sequencing data have been deposited at the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

AUTHOR CONTRIBUTIONS

Conceptualization, K.E.J., F.W.A., E.W.D., and R.B.; formal analysis, K.E.J., T.H., and M.A.; funding acquisition, K.E.J., D.K., K.M.J., E.F.L., L.B., D.A.F., C.A.G., F.W.A., E.W.D., and R.B.; investigation, K.E.J., T.H., E.W.D., K.M.J., D.A.F., A.F., and N.Y.; supervision, K.E.J., L.B., M.C.R., C.A.G., F.W.A., E.W.D., and R.B.; writing – original draft, K.E.J.; writing – review and editing, K.E.J., T.H., E.F.L., L.B., M.C.R., C.A.G., F.W.A., E.W.D., and R.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Human study participants
- METHOD DETAILS
  - MILK study overview
  - Milk composition measurements
  - RNA extraction and sequencing
  - RNA-seq pre-processing and quantification
  - Whole-genome sequencing and quality control
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Comparison of milk transcriptomes to GTEx

- Correlations between milk gene expression and maternal/infant traits
- Examination of PER2 expression and milk traits
- Deconvolution of bulk transcriptomes with bisque
- Milk eQTL analysis
- Colocalization of milk and GTEx eQTLs
- Overlap between milk eGenes and dairy cattle QTL
- Comparison of milk and GTEx eQTL with mash
- Colocalization of milk eQTLs with breast cancer GWAS summary statistics
- Correlations between milk gene expression and oligosaccharides
- Genetic associations at milk eQTLs with milk oligosaccharides
- Processing of infant fecal metagenomes
- Sparse CCA of human milk transcriptomes and infant fecal metagenomes

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100638.

REFERENCES

1. Lefèvre, C.M., Sharp, J.A., and Nicholas, K.R. (2010). Evolution of lactation: ancient origin and extreme adaptations of the lactation system. Annu. Rev. Genomics Hum. Genet. *11*, 219–238. https://doi.org/10.1146/annurev-genom-082509-141806.

2. Victora, C.G., Bahl, R., Barros, A.J.D., França, G.V.A., Horton, S., Krasevec, J., Murch, S., Sankar, M.J., Walker, N., and Rollins, N.C.; Lancet Breastfeeding Series Group (2016). Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. Lancet *387*, 475–490. https://doi.org/10.1016/S0140-6736(15)01024-7.

3. Ballard, O., and Morrow, A.L. (2013). Human milk composition: nutrients and bioactive factors. Pediatr. Clin. North Am. *60*, 49–74. https://doi.org/10.1016/j.pcl.2012.10.002.

4. Andreas, N.J., Kampmann, B., and Mehring Le-Doare, K. (2015). Human breast milk: A review on its composition and bioactivity. Early Hum. Dev. *91*, 629–635. https://doi.org/10.1016/j.earlhumdev.2015.08.013.

5. Christian, P., Smith, E.R., Lee, S.E., Vargas, A.J., Bremer, A.A., and Raiten, D.J. (2021). The need to study human milk as a biological system. Am. J. Clin. Nutr. *113*, 1063–1072. https://doi.org/10.1093/ajcn/nqab075.

6. Golan, Y., and Assaraf, Y.G. (2020). Genetic and Physiological Factors Affecting Human Milk Production and Composition. Nutrients *12*, 1500. https://doi.org/10.3390/nu12051500.

7. Williams, J.E., McGuire, M.K., Meehan, C.L., McGuire, M.A., Brooker, S.L., Kamau-Mbuthia, E.W., Kamundia, E.W., Mbugua, S., Moore, S.E., Prentice, A.M., et al. (2021). Key genetic variants associated with variation of milk oligosaccharides from diverse human populations. Genomics *113*, 1867–1875. https://doi.org/10.1016/j.ygeno.2021.04.004.

8. Bode, L. (2012). Human milk oligosaccharides: every baby needs a sugar mama. Glycobiology *22*, 1147–1162. https://doi.org/10.1093/glycob/cws074.

9. Babakobi, M.D., Reshef, L., Gihaz, S., Belgorodsky, B., Fishman, A., Bujanover, Y., and Gophna, U. (2020). Effect of Maternal Diet and Milk Lipid Composition on the Infant Gut and Maternal Milk Microbiomes. Nutrients *12*, 2539. https://doi.org/10.3390/nu12092539.

10. Pace, R.M., Williams, J.E., Robertson, B., Lackey, K.A., Meehan, C.L., Price, W.J., Foster, J.A., Sellen, D.W., Kamau-Mbuthia, E.W., Kamundia, E.W., et al. (2021). Variation in Human Milk Composition Is Related to

Differences in Milk and Infant Fecal Microbial Communities. Microorganisms *9*, 1153. https://doi.org/10.3390/microorganisms9061153.

11. Stewart, C.J., Ajami, N.J., O'Brien, J.L., Hutchinson, D.S., Smith, D.P., Wong, M.C., Ross, M.C., Lloyd, R.E., Doddapaneni, H., Metcalf, G.A., et al. (2018). Temporal development of the gut microbiome in early childhood from the TEDDY study. Nature *562*, 583–588. https://doi.org/10.1038/s41586-018-0617-x.

12. Fehr, K., Moossavi, S., Sbihi, H., Boutin, R.C.T., Bode, L., Robertson, B., Yonemitsu, C., Field, C.J., Becker, A.B., Mandhane, P.J., et al. (2020). Breastmilk Feeding Practices Are Associated with the Co-Occurrence of Bacteria in Mothers' Milk and the Infant Gut: the CHILD Cohort Study. Cell Host Microbe *28*, 285–297.e4. https://doi.org/10.1016/j.chom.2020.06.009.

13. Milani, C., Duranti, S., Bottacini, F., Casey, E., Turroni, F., Mahony, J., Belzer, C., Delgado Palacio, S., Arboleya Montes, S., Mancabelli, L., et al. (2017). The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. Microbiol. Mol. Biol. Rev. *81*, e00036-17. https://doi.org/10.1128/MMBR.00036-17.

14. Bode, L., Raman, A.S., Murch, S.H., Rollins, N.C., and Gordon, J.I. (2020). Understanding the mother-breastmilk-infant "triad.". Science *367*, 1070–1072. https://doi.org/10.1126/science.aaw6147.

15. Lemay, D.G., Ballard, O.A., Hughes, M.A., Morrow, A.L., Horseman, N.D., and Nommsen-Rivers, L.A. (2013). RNA sequencing of the human milk fat layer transcriptome reveals distinct gene expression profiles at three stages of lactation. PLoS One *8*, e67531. https://doi.org/10.1371/journal.pone.0067531.

16. Lemay, D.G., Hovey, R.C., Hartono, S.R., Hinde, K., Smilowitz, J.T., Ventimiglia, F., Schmidt, K.A., Lee, J.W.S., Islas-Trejo, A., Silva, P.I., et al. (2013). Sequencing the transcriptome of milk production: milk trumps mammary tissue. BMC Genom. *14*, 872. https://doi.org/10.1186/1471-2164-14-872.

17. Nyquist, S.K., Gao, P., Haining, T.K.J., Retchin, M.R., Golan, Y., Drake, R.S., Kolb, K., Mead, B.E., Ahituv, N., Martinez, M.E., et al. (2022). Cellular and transcriptional diversity over the course of human lactation. Proc. Natl. Acad. Sci. USA *119*, e2121720119. https://doi.org/10.1073/pnas.2121720119.

18. Twigger, A.-J., Engelbrecht, L.K., Bach, K., Schultz-Pernice, I., Pensa, S., Stenning, J., Petricca, S., Scheel, C.H., and Khaled, W.T. (2022). Transcriptional changes in the mammary gland during lactation revealed by single cell sequencing of cells from human milk. Nat. Commun. *13*, 562. https://doi.org/10.1038/s41467-021-27895-0.

19. Martin Carli, J.F., Trahan, G.D., Jones, K.L., Hirsch, N., Rolloff, K.P., Dunn, E.Z., Friedman, J.E., Barbour, L.A., Hernandez, T.L., MacLean, P.S., et al. (2020). Single Cell RNA Sequencing of Human Milk-Derived Cells Reveals Sub-Populations of Mammary Epithelial Cells with Molecular Signatures of Progenitor and Mature States: a Novel, Non-invasive Framework for Investigating Human Lactation Physiology. J. Mammary Gland Biol. Neoplasia *25*, 367–387. https://doi.org/10.1007/s10911-020-09466-z.

20. Maningat, P.D., Sen, P., Sunehag, A.L., Hadsell, D.L., and Haymond, M.W. (2007). Regulation of gene expression in human mammary epithelium: effect of breast pumping. J. Endocrinol. *195*, 503–511. https://doi.org/10.1677/JOE-07-0394.

21. Maningat, P.D., Sen, P., Rijnkels, M., Sunehag, A.L., Hadsell, D.L., Bray, M., and Haymond, M.W. (2009). Gene expression in the human mammary epithelium during lactation: the milk fat globule transcriptome. Physiol. Genomics *37*, 12–22. https://doi.org/10.1152/physiolgenomics.90341.2008.

22. Whitaker, K.M., Marino, R.C., Haapala, J.L., Foster, L., Smith, K.D., Teague, A.M., Jacobs, D.R., Fontaine, P.L., McGovern, P.M., Schoenfuss, T.C., et al. (2017). Associations of Maternal Weight Status Before, During, and After Pregnancy with Inflammatory Markers in Breast Milk. Obesity *25*, 2092–2099. https://doi.org/10.1002/oby.22025.

23. Sadr Dadres, G., Whitaker, K.M., Haapala, J.L., Foster, L., Smith, K.D., Teague, A.M., Jacobs, D.R., Jr., Kharbanda, E.O., McGovern, P.M., Schoenfuss, T.C., et al. (2019). Relationship of Maternal Weight Status Before, During, and After Pregnancy with Breast Milk Hormone Concentrations. Obesity *27*, 621–628. https://doi.org/10.1002/oby.22409.

24. Fields, D.A., George, B., Williams, M., Whitaker, K., Allison, D.B., Teague, A., and Demerath, E.W. (2017). Associations between human breast milk hormones and adipocytokines and infant growth and body composition in the first 6 months of life. Pediatr. Obes. *12*, 78–85. https://doi.org/10.1111/ijpo.12182.

25. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. https://doi.org/10.1126/science.aaz1776.

26. Lönnerdal, B. (2003). Nutritional and physiologic significance of human milk proteins. Am. J. Clin. Nutr. *77*, 1537S–1543S. https://doi.org/10.1093/ajcn/77.6.1537S.

27. Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K.M., Sul, J.H., Pietiläinen, K.H., Pajukanta, P., and Halperin, E. (2020). Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nat. Commun. *11*, 1971. https://doi.org/10.1038/s41467-020-15816-6.

28. Dos Santos, C.O., Dolzhenko, E., Hodges, E., Smith, A.D., and Hannon, G.J. (2015). An epigenetic memory of pregnancy in the mouse mammary gland. Cell Rep. *11*, 1102–1109. https://doi.org/10.1016/j.celrep.2015.04.015.

29. Wagner, K.-U., Boulanger, C.A., Henry, M.D., Sgagias, M., Hennighausen, L., and Smith, G.H. (2002). An adjunct mammary epithelial cell population in parous females: its role in functional adaptation and tissue renewal. Development *129*, 1377–1386. https://doi.org/10.1242/dev.129.6.1377.

30. Nommsen-Rivers, L.A., Chantry, C.J., Peerson, J.M., Cohen, R.J., and Dewey, K.G. (2010). Delayed onset of lactogenesis among first-time mothers is related to maternal obesity and factors associated with ineffective breastfeeding. Am. J. Clin. Nutr. *92*, 574–584. https://doi.org/10.3945/ajcn.2010.29192.

31. Kent, J.C., Mitoulas, L.R., Cregan, M.D., Ramsay, D.T., Doherty, D.A., and Hartmann, P.E. (2006). Volume and frequency of breastfeedings and fat content of breast milk throughout the day. Pediatrics *117*, e387–e395. https://doi.org/10.1542/peds.2005-1417.

32. McQueen, C.M., Schmitt, E.E., Sarkar, T.R., Elswood, J., Metz, R.P., Earnest, D., Rijnkels, and Porter, W.W. (2018). PER2 regulation of mammary gland development. Development *145*, dev157966. https://doi.org/10.1242/dev.157966.

33. Garofalo, R. (2010). Cytokines in human milk. J. Pediatr. *156*, S36–S40. https://doi.org/10.1016/j.jpeds.2009.11.019.

34. Gleeson, J.P., Chaudhary, N., Fein, K.C., Doerfler, R., Hredzak-Showalter, P., and Whitehead, K.A. (2022). Profiling of mature-stage human breast milk cells identifies six unique lactocyte subpopulations. Sci. Adv. *8*, eabm6865. https://doi.org/10.1126/sciadv.abm6865.

35. Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. *16*, 197–212. https://doi.org/10.1038/nrg3891.

36. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. J. R. Stat. Soc. Series B Stat. Methodol. *82*, 1273–1300. https://doi.org/10.1111/rssb.12388.

37. Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. PLoS Genet. *17*, e1009440. https://doi.org/10.1371/journal.pgen.1009440.

38. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation

in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.

39. Ender, P., Gagliardi, P.A., Dobrzyński, M., Frismantiene, A., Dessauges, C., Höhener, T., Jacques, M.-A., Cohen, A.R., and Pertz, O. (2022). Spatiotemporal control of ERK pulse frequency coordinates fate decisions during mammary acinar morphogenesis. Dev. Cell *57*, 2153–2167.e6. https://doi.org/10.1016/j.devcel.2022.08.008.

40. Urbut, S.M., Wang, G., Carbonetto, P., and Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat. Genet. *51*, 187–195. https://doi.org/10.1038/s41588-018-0268-8.

41. Migliavacca Zucchetti, B., Peccatori, F.A., and Codacci-Pisanelli, G. (2020). Pregnancy and Lactation: Risk or Protective Factors for Breast Cancer? Adv. Exp. Med. Biol. *1252*, 195–197. https://doi.org/10.1007/978-3-030-41596-9_27.

42. Collaborative Group on Hormonal Factors in Breast Cancer (2002). Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease. Lancet *360*, 187–195. https://doi.org/10.1016/S0140-6736(02)09454-0.

43. Zhang, H., Ahearn, T.U., Lecarpentier, J., Barnes, D., Beesley, J., Qi, G., Jiang, X., O'Mara, T.A., Zhao, N., Bolla, M.K., et al. (2020). Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nat. Genet. *52*, 572–581. https://doi.org/10.1038/s41588-020-0609-2.

44. Fachal, L., Aschard, H., Beesley, J., Barnes, D.R., Allen, J., Kar, S., Pooley, K.A., Dennis, J., Michailidou, K., Turman, C., et al. (2020). Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. Nat. Genet. *52*, 56–73. https://doi.org/10.1038/s41588-019-0537-1.

45. Ferreira, M.A., Gamazon, E.R., Al-Ejeh, F., Aittomäki, K., Andrulis, I.L., Anton-Culver, H., Arason, A., Arndt, V., Aronson, K.J., Arun, B.K., et al. (2019). Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. Nat. Commun. *10*, 1741. https://doi.org/10.1038/s41467-018-08053-5.

46. Beesley, J., Sivakumaran, H., Moradi Marjaneh, M., Shi, W., Hillman, K.M., Kaufmann, S., Hussein, N., Kar, S., Lima, L.G., Ham, S., et al. (2020). eQTL Colocalization Analyses Identify NTN4 as a Candidate Breast Cancer Risk Gene. Am. J. Hum. Genet. *107*, 778–787. https://doi.org/10.1016/j.ajhg.2020.08.006.

47. Harita, Y., Kitanaka, S., Isojima, T., Ashida, A., and Hattori, M. (2017). Spectrum of LMX1B mutations: from nail-patella syndrome to isolated nephropathy. Pediatr. Nephrol. *32*, 1845–1850. https://doi.org/10.1007/s00467-016-3462-x.

48. Kellman, B.P., Richelle, A., Yang, J.-Y., Chapla, D., Chiang, A.W.T., Najera, J.A., Liang, C., Fürst, A., Bao, B., Koga, N., et al. (2022). Elucidating Human Milk Oligosaccharide biosynthetic genes through network-based multi-omics integration. Nat. Commun. *13*, 2455. https://doi.org/10.1038/s41467-022-29867-4.

49. Castro, I., García-Carral, C., Furst, A., Khwajazada, S., García, J., Arroyo, R., Ruiz, L., Rodríguez, J.M., Bode, L., and Fernández, L. (2022). Interactions between human milk oligosaccharides, microbiota and immune factors in milk of women with and without mastitis. Sci. Rep. *12*, 1367. https://doi.org/10.1038/s41598-022-05250-7.

50. Pannaraj, P.S., Li, F., Cerini, C., Bender, J.M., Yang, S., Rollie, A., Adisetiyo, H., Zabih, S., Lincez, P.J., Bittinger, K., et al. (2017). Association Between Breast Milk Bacterial Communities and Establishment and Development of the Infant Gut Microbiome. JAMA Pediatr. *171*, 647–654. https://doi.org/10.1001/jamapediatrics.2017.0378.

51. Kijner, S., Kolodny, O., and Yassour, M. (2022). Human milk oligosaccharides and the infant gut microbiome from an eco-evolutionary perspective. Curr. Opin. Microbiol. *68*, 102156. https://doi.org/10.1016/j.mib.2022.102156.

52. Witten, D.M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics *10*, 515–534. https://doi.org/10.1093/biostatistics/kxp008.

53. Priya, S., Burns, M.B., Ward, T., Mars, R.A.T., Adamowicz, B., Lock, E.F., Kashyap, P.C., Knights, D., and Blekhman, R. (2022). Identification of shared and disease-specific host gene-microbiome associations across human diseases using multi-omic integration. Nat. Microbiol. *7*, 780–795. https://doi.org/10.1038/s41564-022-01121-z.

54. Kreuzaler, P.A., Staniszewska, A.D., Li, W., Omidvar, N., Kedjouar, B., Turkson, J., Poli, V., Flavell, R.A., Clarkson, R.W.E., and Watson, C.J. (2011). Stat3 controls lysosomal-mediated cell death in vivo. Nat. Cell Biol. *13*, 303–309. https://doi.org/10.1038/ncb2171.

55. Sargeant, T.J., Lloyd-Lewis, B., Resemann, H.K., Ramos-Montoya, A., Skepper, J., and Watson, C.J. (2014). Stat3 controls cell death during mammary gland involution by regulating uptake of milk fat globules and lysosomal membrane permeabilization. Nat. Cell Biol. *16*, 1057–1068. https://doi.org/10.1038/ncb3043.

56. Seki, D., Errerd, T., and Hall, L.J. (2023). The role of human milk fats in shaping neonatal development and the early life gut microbiota. Microbiome Res. Rep. *2*, 8. https://doi.org/10.20517/mrr.2023.09.

57. Watson, C.J., and Neoh, K. (2008). The Stat family of transcription factors have diverse roles in mammary gland development. Semin. Cell Dev. Biol. *19*, 401–406. https://doi.org/10.1016/j.semcdb.2008.07.021.

58. Henrick, B.M., Rodriguez, L., Lakshmikanth, T., Pou, C., Henckel, E., Arzoomand, A., Olin, A., Wang, J., Mikes, J., Tan, Z., et al. (2021). Bifidobacteria-mediated immune system imprinting early in life. Cell *184*, 3884–3898.e11. https://doi.org/10.1016/j.cell.2021.05.030.

59. Barratt, M.J., Nuzhat, S., Ahsan, K., Frese, S.A., Arzamasov, A.A., Sarker, S.A., Islam, M.M., Palit, P., Islam, M.R., Hibberd, M.C., et al. (2022). Bifidobacterium infantis treatment promotes weight gain in Bangladeshi infants with severe acute malnutrition. Sci. Transl. Med. *14*, eabk1107. https://doi.org/10.1126/scitranslmed.abk1107.

60. Moore, R.E., and Townsend, S.D. (2019). Temporal development of the infant gut microbiome. Open Biol. *9*, 190128. https://doi.org/10.1098/rsob.190128.

61. Joseph, T.A., Chlenski, P., Litman, A., Korem, T., and Pe'er, I. (2022). Accurate and robust inference of microbial growth dynamics from metagenomic sequencing reveals personalized growth rates. Genome Res. *32*, 558–568. https://doi.org/10.1101/gr.275533.121.

62. Zahoor, A., Yang, Y., Yang, C., Khan, S.B., Reix, C., Anwar, F., Guo, M.-Y., and Deng, G. (2020). MerTK negatively regulates Staphylococcus aureus induced inflammatory response via Toll-like receptor signaling in the mammary gland. Mol. Immunol. *122*, 1–12. https://doi.org/10.1016/j.molimm.2020.03.007.

63. Aryeetey, R.N.O., Marquis, G.S., Timms, L., Lartey, A., and Brakohiapa, L. (2008). Subclinical mastitis is common among Ghanaian women lactating 3 to 4 months postpartum. J. Hum. Lact. *24*, 263–267. https://doi.org/10.1177/0890334408316077.

64. Pace, R.M., Pace, C.D.W., Fehrenkamp, B.D., Price, W.J., Lewis, M., Williams, J.E., McGuire, M.A., and McGuire, M.K. (2022). Sodium and Potassium Concentrations and Somatic Cell Count of Human Milk Produced in the First Six Weeks Postpartum and Their Suitability as Biomarkers of Clinical and Subclinical Mastitis. Nutrients *14*, 4708. https://doi.org/10.3390/nu14224708.

65. Samuel, T.M., De Castro, C.A., Dubascoux, S., Affolter, M., Giuffrida, F., Billeaud, C., Picaud, J.-C., Agosti, M., Al-Jashi, I., Pereira, A.B., et al. (2019). Subclinical Mastitis in a European Multicenter Cohort: Prevalence, Impact on Human Milk (HM) Composition, and Association with Infant HM Intake and Growth. Nutrients *12*, 105. https://doi.org/10.3390/nu12010105.

66. Tuaillon, E., Viljoen, J., Dujols, P., Cambonie, G., Rubbo, P.-A., Nagot, N., Bland, R.M., Badiou, S., Newell, M.-L., and Van de Perre, P. (2017). Subclinical mastitis occurs frequently in association with dramatic changes in inflammatory/anti-inflammatory breast milk components. Pediatr. Res. *81*, 556–564. https://doi.org/10.1038/pr.2016.220.

67. Karlsson, M., Zhang, C., Méar, L., Zhong, W., Digre, A., Katona, B., Sjöstedt, E., Butler, L., Odeberg, J., Dusart, P., et al. (2021). A single-cell type transcriptomics map of human tissues. Sci. Adv. 7, eabh2169. https://doi.org/10.1126/sciadv.abh2169.

68. Salas, L.A., Lundgren, S.N., Browne, E.P., Punska, E.C., Anderton, D.L., Karagas, M.R., Arcaro, K.F., and Christensen, B.C. (2020). Prediagnostic breast milk DNA methylation alterations in women who develop breast cancer. Hum. Mol. Genet. 29, 662–673. https://doi.org/10.1093/hmg/ddz301.

69. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

70. DeLuca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C., Reich, M., Winckler, W., and Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics 28, 1530–1532. https://doi.org/10.1093/bioinformatics/bts196.

71. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. https://doi.org/10.1186/s13059-014-0550-8.

72. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience 10, giab008. https://doi.org/10.1093/gigascience/giab008.

73. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575. https://doi.org/10.1086/519795.

74. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. https://doi.org/10.1093/bioinformatics/btp616.

75. Alexa, A., and Rahnenfuhrer, J. (2022). topGO: Enrichment Analysis for Gene Ontology. https://doi.org/10.18129/B9.bioc.topGO.

76. Quick, C., Guan, L., Li, Z., Li, X., Dey, R., Liu, Y., Scott, L., and Lin, X. (2020). A versatile toolkit for molecular QTL mapping and meta-analysis at scale. bioRxiv. https://doi.org/10.1101/2020.12.18.423490.

77. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 10, e1004383. https://doi.org/10.1371/journal.pgen.1004383.

78. Al-Ghalith, G., and Knights, D. (2020). BURST enables mathematically optimal short-read alignment for big data. bioRxiv. https://doi.org/10.1101/2020.09.08.287128.

79. Beghini, F., McIver, L.J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A.M., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife 10, e65088. https://doi.org/10.7554/eLife.65088.

80. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell 185, 3426–3440.e19. https://doi.org/10.1016/j.cell.2022.08.004.

81. Heisel, T., Johnson, A.J., Gonia, S., Dillon, A., Skalla, E., Haapala, J., Jacobs, K.M., Nagel, E., Pierce, S., Fields, D., et al. (2022). Bacterial, fungal, and interkingdom microbiome features of exclusively breastfeeding dyads are associated with infant age, antibiotic exposure, and birth mode. Front. Microbiol. 13, 1050574. https://doi.org/10.3389/fmicb.2022.1050574.

82. Seferovic, M.D., Mohammad, M., Pace, R.M., Engevik, M., Versalovic, J., Bode, L., Haymond, M., and Aagaard, K.M. (2020). Maternal diet alters human milk oligosaccharide composition with implications for the milk metagenome. Sci. Rep. 10, 22092. https://doi.org/10.1038/s41598-020-79022-6.

83. Fields, D.A., and Demerath, E.W. (2012). Relationship of insulin, glucose, leptin, IL-6 and TNF-α in human breast milk with infant growth and body composition. Pediatr. Obes. 7, 304–312. https://doi.org/10.1111/j.2047-6310.2012.00059.x.

84. Casadio, Y.S., Williams, T.M., Lai, C.T., Olsson, S.E., Hepworth, A.R., and Hartmann, P.E. (2010). Evaluation of a mid-infrared analyzer for the determination of the macronutrient composition of human milk. J. Hum. Lact. 26, 376–383. https://doi.org/10.1177/0890334410376948.

85. Billard, H., Simon, L., Desnots, E., Sochard, A., Boscher, C., Riaublanc, A., Alexandre-Gouabau, M.-C., and Boquien, C.-Y. (2016). Calibration Adjustment of the Mid-infrared Analyzer for an Accurate Determination of the Macronutrient Composition of Human Milk. J. Hum. Lact. 32, NP19. https://doi.org/10.1177/0890334415588513.

86. Harrell, F.E., Jr. (2022). Hmisc: Harrell Miscellaneous. https://cran.r-project.org/package=Hmisc.

87. Li, J.H., Mazur, C.A., Berisa, T., and Pickrell, J.K. (2021). Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. Genome Res. 31, 529–537. https://doi.org/10.1101/gr.266486.120.

88. Wasik, K., Berisa, T., Pickrell, J.K., Li, J.H., Fraser, D.J., King, K., and Cox, C. (2021). Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. BMC Genom. 22, 197. https://doi.org/10.1186/s12864-021-07508-2.

89. Lund, S.P., Nettleton, D., McCarthy, D.J., and Smyth, G.K. (2012). Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. Stat. Appl. Genet. Mol. Biol. 11. https://doi.org/10.1515/1544-6115.1826.

90. Lun, A.T.L., and Smyth, G.K. (2017). No counts, no variance: allowing for loss of degrees of freedom when assessing biological variability from RNA-seq data. Stat. Appl. Genet. Mol. Biol. 16, 83–93. https://doi.org/10.1515/sagmb-2017-0010.

91. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

92. Grolemund, G., and Wickham, H. (2011). Dates and Times Made Easy with lubridate. J. Stat. Softw. 40, 1–25. https://doi.org/10.18637/jss.v040.i03.

93. Lüdecke, D., Ben-Shachar, M.S., Patil, I., Waggoner, P., and Makowski, D. (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. J. Open Source Softw. 6, 3139. https://doi.org/10.21105/joss.03139.

94. McCaw, Z. (2022). RNOmni: Rank Normal Transformation Omnibus Test. https://CRAN.R-project.org/package=RNOmni.

95. Arvanitis, M., Tayeb, K., Strober, B.J., and Battle, A. (2022). Redefining tissue specificity of genetic regulation of gene expression in the presence of allelic heterogeneity. Am. J. Hum. Genet. 109, 223–239. https://doi.org/10.1016/j.ajhg.2022.01.002.

96. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16, 284–287. https://doi.org/10.1089/omi.2011.0118.

97. Wallace, C. (2020). Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. PLoS Genet. 16, e1008720. https://doi.org/10.1371/journal.pgen.1008720.

98. Burgess, S., Small, D.S., and Thompson, S.G. (2017). A review of instrumental variable estimators for Mendelian randomization. Stat. Methods Med. Res. 26, 2333–2355. https://doi.org/10.1177/0962280215597579.

99. Al-Ghalith, G.A., Hillmann, B., Ang, K., Shields-Cutler, R., and Knights, D. (2018). SHI7 Is a Self-Learning Pipeline for Multipurpose Short-Read DNA Quality Control. mSystems *3*, e00202-17. https://doi.org/10.1128/mSystems.00202-17.

100. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

101. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60. https://doi.org/10.1038/nmeth.3176.

102. Ye, Y., and Doak, T.G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS Comput. Biol. *5*, e1000465. https://doi.org/10.1371/journal.pcbi.1000465.

103. Briatte, F. (2021). ggnetwork: Geometries to Plot Networks with "ggplot2.". https://CRAN.R-project.org/package=ggnetwork.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
STAR v2.7.1a	Dobin et al.[69]	https://github.com/alexdobin/STAR
RNA-SeQC73 v2.3.4	DeLuca et al.[70]	https://github.com/francois-a/rnaseqc
R package: DESeq2 v1.30.0	Love et al.[71]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
BCFtools v1.6	Danecek et al.[72]	https://www.htslib.org/download/
PLINK v1.90b6.10	Purcell et al.[73]	https://www.cog-genomics.org/plink/
R package: edgeR v3.32.1	Robinson et al.[74]	https://bioconductor.org/packages/release/bioc/html/edgeR.html
R package: topGO	Alexa et al.[75]	https://bioconductor.org/packages/release/bioc/html/topGO.html
BisqueRNA R package	Jew et al.[27]	https://github.com/cran/BisqueRNA
APEX toolkit	Quick et al.[76]	https://github.com/corbinq/apex
R package: coloc	Giambartolomei et al.[77]	https://cloud.r-project.org/web/packages/coloc/index.html
R package: mashR	Urbut et al.[40]	https://github.com/stephenslab/mashr
BURST version 0.99.7f96	Al-Ghalith et al.[78]	https://github.com/knights-lab/BURST
MetaPhlAn v3.0.7	Beghini et al.[79]	https://huttenhower.sph.harvard.edu/metaphlan/
Sparse canonical components analysis code	Priya et al.[53]	https://github.com/blekhmanlab/host_gene_microbiome_interactions
CoPTR	Joseph et al.[61]	https://github.com/tyjo/coptr
Deposited data		
Milk RNA-sequencing data	This paper	dbGaP: phs003408.v1.p1
Milk DNA-sequencing data and genotypes	This paper	dbGaP: phs003408.v1.p1
Infant fecal metagenomic sequencing data	This paper	SRA: PRJNA1019702
Milk transcriptome quantifications, infant fecal metagenome abundances, milk eQTL summary statistics, HMO concentrations, additional metadata	This paper	https://figshare.com/collections/Johnson_et_al_human_milk_multi-omics/7371256
GTEx RNA-sequencing quantifications and eQTL summary statistics	GTEx Portal[25]	https://gtexportal.org/home/downloads/adult-gtex/overview
1000 Genomes genotypes	Byrska-Bishop et al.[80]	https://www.internationalgenome.org/data-portal/data-collection/30x-grch38
Single-cell human milk RNA-seq data	Nyquist et al.[17]	https://singlecell.broadinstitute.org/single_cell/study/SCP1671/cellular-and-transcriptional-diversity-over-the-course-of-human-lactation
Breast cancer GWAS summary statistics	Zhang et al.[43]	http://bcac.ccge.medschl.cam.ac.uk/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Human study participants

This observational study comprised female adults recruited prenatally in the United States and their infants. Individual level demographic information and covariates are available in supplementary tables and on figshare (see key resources table). The Institutional Review Boards of the University of Oklahoma, the University of Minnesota, and the HealthPartners Institute approved this study (STUDY00009021). This study has been registered with ClinicalTrials.gov (identifier NCT03301753).

METHOD DETAILS

MILK study overview

Participant recruitment, clinical data, and milk sample collection for the Mothers and Infants LinKed for health (MILK) study have been described previously.[22–24,81] Briefly, participants who intended to exclusively breastfeed were enrolled prenatally during healthy, uncomplicated pregnancies at the University of Minnesota in collaboration with HealthPartners Institute (Minneapolis, MN) or the University of Oklahoma Health Sciences Center. Recruited mothers were 21–45 years old, non-smokers, non-diabetic, and delivered singleton infants at full term (37 0/7–41 6/7 weeks gestation) with 10th–90th percentile birth weight on the WHO growth chart. No participants reported symptoms of mastitis or breast infection at the time of milk sample collection. Clinical data for each mother-infant dyad was collected from the delivering hospitals' electronic health record and from electronic questionnaires at study visits at 1 and 6 months postpartum. Clinical study data were managed using REDCap electronic data capture tools hosted at the University of Minnesota. REDCap (Research Electronic Data Capture) is a secure, web-based software platform designed to support data capture for research studies. The data described in this manuscript comes from a subset of MILK Study mother/infant pairs who consented to maternal whole-genome sequencing, milk RNA sequencing, and microbiome assessment of infant fecal samples.

Gestational diabetes diagnosis

Gestational diabetes screening occurred between the 26th and 28th weeks of gestation by a 1-h blood glucose concentration after a 50 g oral glucose challenge test (OGCT). Women with OGCT levels greater than 130 g/dL then received a 3-h 100 g oral glucose tolerance test to confirm gestational diabetes. Gestational diabetes was diagnosed if a minimum of two out of four glucose level time point assessments were met or exceeded: 95 mg/dL (fasting), 180 mg/dL (1 h), 155 mg/dL (2 h), or 140 mg/dL (3 h).

Milk sample collection

Milk samples were collected at study visits at approximately 1 month postpartum, and infant fecal samples were collected at study visits at 1 and 6 months. Upon study visit arrival, participants fed their infants *ad libitum* from one or both breasts until infants were satisfied. Two hours following this feeding, milk was collected from the right breast using a hospital-grade electric breast pump (Medela Symphony; Medela, Inc., Zug, Switzerland), with expression ceasing when milk stopped flowing. Expressed milk volume and weight was recorded, milk was gently mixed, aliquots were made, and then stored at −80°C within 20 min of collection and kept at −80°C until thawed for RNA/DNA extraction.

Milk composition measurements

Human milk oligosaccharides

Concentrations of HMOs were quantified from 2 mL previously frozen whole milk aliquots as previously described.[82] 19 HMOs were identified and quantified: 2′-fucosyllactose (2′FL), 3-fucosyllactose (3′FL), 3′-sialyllactose (3′SL), 6′-sialyllactose (6′SL), difucosyllactose (DFLac), difucosyllacto-N-hexaose (DFLNH), difucosyllacto-N-tetrose (DFLNT), disialyllacto-N-hexaose (DSLNH), disialyllacto-N-tetraose (DSLNT), fucodisialyllacto-N-hexaose (FDSLNH), fucosyllacto-N-hexaose (FLNH), lacto-N-fucopentaose (LNFP) I, LNFP II, LNFP III, lacto-N-hexaose (LNH), lacto-N-neotetraose (LNnT), lacto-N-tetrose (LNT), sialyl-lacto-N-tetraose b (LSTb), and sialyl-lacto-N-tetraose c (LSTc). Secretor milk was defined as having a 2′FL concentration that was greater than a natural, very low break in the data (Figure 3A). Weight-based concentrations were used for all statistical analyses (micrograms per milliliter). The sum of HMO concentrations was calculated as the total concentrations of the 19 measured HMOs. HMO concentrations were estimated over two batches, and HMO batch was included as a covariate in all analyses of HMO data.

Milk cytokines/nutrients/hormones

Milk fat was separated from the aqueous phase by centrifugation, and skim milk was assayed using commercially available immunoassay kits for insulin, glucose, leptin, CRP, and IL6 as previously described.[22,24,83] These milk component assays were processed in 2–5 batches depending on the assay. Batch effects were corrected using an analysis of variance model with formula:

$$\log(\text{assay value}) \sim \text{factor}(\text{batch})$$

using the 'aov' command in R. The residuals from this model, representing the batch-corrected values, were used in all downstream data analyses. There were not sample replicates across batches; original and corrected values are plotted in Figure S7.

Milk fat and lactose

Milk fat and lactose concentrations were assessed using mid-infrared spectrophotometry (Calais Milk Analyzer, North American Instruments, LLC, Lake Oswego, OR).[84,85] Human milk samples were gradually thawed and then diluted with deionized water in a 1:1 dilution. Breastmilk control samples with standard macronutrient content were run prior to study sample testing to confirm instrument calibration. Samples were heated in a water bath until the samples reached 40°C and were mixed by gentle hand inversion for 2 min prior to analysis, per manufacturer instructions. Milk fat percent reliability was assessed in a random subset of 34 samples (17 duplicate samples) with an intraclass correlation coefficient (ICC) of 0.99, $p < 0.001$. Validity was assessed in a random subset of 30 samples against the gold standard Mojonnier method[83] yielding a high cross-method ICC of 0.936, $p < 0.001$.

RNA extraction and sequencing

We extracted RNA from whole milk cell pellets to capture gene expression from both mammary epithelial cells and immune cells in milk. Previous studies that have performed bulk RNA-sequencing from human milk have used RNA extracted from the milk fat layer.[15] This procedure enriches for milk fat globule RNA, which originates from mammary epithelial cells.[15,16] Our approach allowed us to computationally estimate the contribution of different cell types to the milk transcriptomes, and explore genetic influences on gene expression that could be specific to the immune cells in milk, in addition to mammary epithelial cells.

Nucleic acid extractions and RNA-seq library preparation and sequencing was performed at the University of Minnesota Genomics Center (UMGC) in two batches (Table S1). In the first batch, frozen 2 mL whole milk aliquots from 245 milk samples were thawed and split in two, with each 1 mL half used for either RNA or DNA extraction. In the second batch, frozen 2 mL whole milk aliquots from 106 milk samples were thawed and the entire sample was used for RNA extraction. RNA was extracted from the cell pellet using the RNeasy Plus Universal HTP following the manufacturer's instructions. We used the TakaraBio SMARTer Stranded Total RNA-seq Kit v2 - Pico Input Mammalian for RNA-seq library preparation. RNA libraries were sequenced on an Illumina NovaSeq 6000 S2 flow cell with 2 × 150 paired-end reads to a median depth of 36.8 million reads per sample. Sample-level details of RNA extraction and sequencing are in Table S1.

RNA-seq pre-processing and quantification

RNA-seq reads were trimmed with Trimmomatic and aligned with STAR[69] v2.7.1a to the GRCh38 human reference genome. Gene-level quantification was performed with RNA-SeQC[70] v2.3.4 using a Gencode v36 gene model annotation that was collapsed to a single transcript model per gene using a script provided by GTEx ("collapse_annotation.py" from https://github.com/broadinstitute/gtex-pipeline/tree/master/gene_model).

To assess the gene-level quantifications, TPM spearman correlations were calculated between each pair of samples with the 'rcorr' function from the 'Hmisc' R package.[86] The first RNA-seq batch was sequenced in two pools (Table S1). Two samples that had poor quality in the first RNA-seq batch were re-run in the second RNA-seq batch (using an additional aliquot from the same original milk sample). We included only the replicate from the second batch for downstream analyses (Table S1). Samples with fewer than 10,000 genes detected were removed. There were five participants that contributed two milk samples, from two separate pregnancies. We included only one milk sample from each of these participants in our analyses, leaving 316 milk transcriptomes from 316 different participants (Table S1).

To explore technical sources of variation in our gene expression data, we performed a principal-component analysis of all 316 milk transcriptomes (Figure S1). We used the thinCounts function in edgeR to downsample each milk sample to 3,491,080 reads (the fewest reads in any one sample). We took the resulting count matrix as a DESeq2 object and performed a variance stabilizing transformation (VST). We then selected the 1000 most variable genes from the VST matrix, and performed principal-component analysis in R with the 'prcomp' function. Examining correlations between the PCs and quality control metrics of RNA extraction, library preparation, or sequencing, we selected five covariates to include in our differential gene expression analysis (below): batch, RIN, RNA concentration, number of genes detected, and mean 3′ bias (Figure S1). The 'batch' categorical variable had 3 levels representing the two sequencing pools of batch 1 and the single pool of batch 2 (Table S1; Figure S1).

Whole-genome sequencing and quality control

DNA was extracted from the cell pellet using the QIAamp 96 DNA Blood Kit at UMGC following the manufacturer's instructions. Low-pass whole genome sequencing (WGS) at ~1x and genotype imputation was performed by Gencove. Gencove's low-pass WGS and imputation provides comparable or improved accuracy and variant discovery to array-based genotyping.[87,88] 173 milk samples successfully underwent WGS and imputation from the original 1 mL aliquot DNA extraction. 72 samples had insufficient DNA extracted from the initial 1 mL sample, or failed Gencove's quality control. Of these 72 samples, 62 had an additional 15 mL frozen aliquot that was shipped to Gencove and DNA was extracted using a mag Nucleic Acid Purification Kit (Biosearch Technologies), and ~1x low-pass WGS was performed. 11 of these samples failed Gencove's quality control and 51 samples successfully underwent WGS and imputation, resulting in 224 samples with genotype information. Finally, we submitted a third batch of 38 additional samples with 15 mL frozen aliquots to Gencove for DNA extraction and WGS as with the 15 mL aliquots above. 35 of these passed Gencove's QC pipeline, resulting in a total of 259 samples with genotype information. Of the 19 total samples that failed Gencove's QC pipeline, 1 failed the minimum bases sequenced and 18 failed the contamination metric (i.e., contamination by DNA from another sample of the same species, likely due to cross-sample contamination upstream of sequencing). 8 participants contributed 2 milk samples (from 2 separate pregnancies), and we included only one sample per participant in our analyses, leaving 251 unique individuals with genotype information. Sample-level details of extraction and sequencing are in Table S1. BCFtools[72] was used to combine all VCFs into a BCF file for all individuals, filtering for minor allele frequency >1% and maximum missing genotypes of 5%. A genetic relatedness matrix was generated with the PLINK[73] (v1.90b6.10) '–make-rel' command, and one individual from pairs with relatedness coefficient >0.05 were pruned, leaving 230 individuals for genetic analyses.

To compare our genotypes to a well-defined population sample, we utilized the 1000 Genomes (1KG) 30x coverage whole genome sequencing dataset.[80] VCF files containing genotypes for 2,504 participants were downloaded from https://www.internationalgenome.org/data-portal/data-collection/30x-grch38. We used BCFtools to combine all 1KG VCFs into a single BCF file, filtering for minor allele frequency >1% and maximum missing genotypes of 5%. We then used the BCFtools command 'merge'

to create a single BCF file containing both the 1KG and milk study genotypes, filtering for genotypes missing in >5% of samples, thus removing variants absent in our milk study which comprised ∼8% of samples in the combined dataset. Genetic principal components (PCs) were calculated with PLINK using 902,579 SNPs with minor allele frequency >1% after pruning for linkage disequilibrium (PLINK command '—indep-pairwise 200 100 0.5'). The milk study participants mainly clustered with the European ancestry 1KG samples (Figure S2), in agreement with the genetic ancestry proportion estimates provided by Gencove, with only 19 of 230 individuals with estimated European ancestry <95% (Figure S3). We selected the first 3 genetic PCs to use as covariates in eQTL mapping.

We checked for sample swaps by performing genotype calling from RNA-seq reads aligned to chromosome 2 using 'bcftools mpileup', and using 'bcftools gtcheck' to compare genotypes from RNA-seq to Gencove variant calls from low-pass WGS.[72] We did not detect any sample swaps: for all samples included in eQTL analysis, the DNA sample with matching sample ID had the lowest average concordance, compared to all DNA samples with a different sample ID (Figure S4).

QUANTIFICATION AND STATISTICAL ANALYSIS

Comparison of milk transcriptomes to GTEx
We downloaded gene-level counts for GTEx samples from the GTEx portal (dataset GTEx_Analysis_2017-06-05_v8_RNA-SeQCv1.1.9_gene_reads.gct.gz). We filtered to only female GTEx samples, removed tissues with fewer than 19 remaining samples, and then selected 19 random samples for each tissue. We filtered to genes that were detected in both datasets after filtering genes with count 0 across all GTEx & milk samples, leaving 30,468 genes. We then used the thinCounts function in edgeR to downsample each GTEx and milk sample to 5 million read counts. We took the resulting count matrix as a DESeq2 object and performed variance stabilizing transformation (VST). We then took the VST matrix of only GTEx samples, selected the 1000 most variable genes, and performed principal-component analysis in R with the 'prcomp' function. We then projected the milk samples onto the PCA scatterplots by calculating 19 random milk sample's values from the GTEx-only PCA to generate Figures 1A and S5.

To compare TPM values across milk and GTEx samples (Figure 1B), we downloaded gene-level TPM values from the GTEx portal (GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz). We filtered to include only female GTEx samples and filtered to protein-coding genes (as annotated in EnsDb.Hsapiens.v86) and removed histone genes. Our RNA library preparation kit (TakaraBio SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian) did not include polyA selection and histone gene mRNAs are not polyadenylated, resulting in higher detection of histone mRNAs in our data than in GTEx. We then rescaled the TPM for each GTEx and milk sample to again sum to 1 million and calculated each gene's median TPM across a tissue type.

Correlations between milk gene expression and maternal/infant traits
We used edgeR[74] to test for correlations between milk gene expression and maternal/milk traits, including all tested traits and technical covariates. Included traits were: Milk CRP concentration, milk glucose concentration, milk IL-6 protein concentration, milk insulin concentration, milk leptin concentration, milk volume expressed, gestational diabetes status, gestational weight gain, maternal pre-pregnancy BMI, maternal age, and parity ($N = 269$ milk samples with no missing data that were included in this analysis; Table S2). We also performed differential gene expression for two macronutrient traits (milk fat % and milk lactose %) separately, as these traits had the smallest sample size, and no individuals with gestational diabetes also had these measurements. Thus, we tested for gene expression for these traits including all other traits except gestational diabetes status as covariates on a smaller sample size ($N = 171$). We scaled each trait to a mean of zero and standard deviation of one, except binary traits (gestational diabetes status) and parity, for which we use the integer number of previous births. The count matrix and metadata were loaded into an edgeR object and the "filterByExpr" was used to remove lowly expressed genes, leaving 12,006 genes (or 12,332 genes for milk fat/lactose). We then used the 'estimateDisp' function on a design matrix regressing gene expression across all traits. This model accounted for potential confounding technical effects, including batch, RIN, RNA concentration, number of genes detected, and mean 3′ bias, by including them as covariates. We then used 'glmQLFit' to fit a quasi-likelihood negative binomial generalized log-linear model to the count matrix and design model, and 'glmQLFTest' to perform a quasi-likelihood F-test testing for the relationship between each gene against each tested trait. This model was selected for its handling of the over-dispersion in RNA-seq count data and type I error control.[89,90] We used Benjamin-Hochberg correction of p values across all 12,006 genes (or 12,332 for fat/lactose) by 13 traits.

To assess the impact of RNA quality (as measured by RIN) on our differential gene expression results, we ran the same analysis on each trait in the top and bottom half of samples separately. Gestational diabetes status was excluded from this analysis because few samples with gestational diabetes were in the bottom half by RIN (only $N = 5$ samples with GDM). For the five traits with at least ten differentially expressed genes identified in the low RIN score subset (q value < 10%; milk glucose, IL-6, lactose, volume expressed, and parity), we tested for a correlation between the log fold-change estimates between the low and high RIN sample subsets for those genes. For all five traits there was a significant correlation ($p < 0.01$, r > 0.8 [except lactose]; Figure S8). Considering all genes, not just those significantly differentially expressed, there was a significant positive correlation between the top and bottom RIN subsets for all traits that had at least 50 differentially expressed genes in the full sample (Table S4). Thus, we moved forward with gene ontology enrichment for those traits with at least 50 differentially expressed genes.

We tested for gene ontology enrichment of significant genes (q value < 10%) for each trait using the R package topGO,[75] with all tested genes as the background gene list. We used the 'resultFisher' function to run a classic Fisher's exact test for each gene ontology, and used a Benjamini-Hochberg correction[91] for all ontologies ($N = 14,119$) across the 7 traits with at least 50 significant

genes (milk glucose, milk IL-6, milk insulin, milk volume expressed, gestational diabetes status, milk lactose %, parity; 98,833 tests). We report pathways with q value < 10%, fewer than 500 annotated genes, and an overlap of more than 5 genes with the significantly associated gene list for each trait (Table S5). All 7 traits had enriched ontologies that met these criteria.

To test for in interaction between maternal obesity status and the 6 traits with at least 50 significant differentially expressed genes (milk IL-6, milk glucose, milk insulin, milk lactose, parity, milk volume) with their association with milk gene expression, we filtered the 269 participants included in differential gene expression above into two categories based on pre-pregnancy BMI: 'normal weight' ($N = 121$, $18.5 \leq$ BMI < 25) or 'obese' ($N = 69$, BMI ≥ 30). For milk lactose, after filtering individuals with missing data as described above, there were $N = 78$ 'normal weight' and $N = 38$ 'obese'. Gestational diabetes was excluded from the interaction analysis because there were only 3 individuals with gestational diabetes in the 'normal weight' category. We then repeated the analysis as with the gene-wise model in the full sample above, but replacing BMI with this normal/obese categorical variable and including an interaction term between obesity status and the milk composition trait. Only gene/trait pairs with a significant correlation in the original analysis without an interaction term (q value < 10%) were tested. The interaction term p values were corrected across all included gene/trait pairs (4,525 tests) using a Benjamin-Hochberg correction (Table S6).

Examination of PER2 expression and milk traits
Circadian rhythm genes were defined as those in KEGG pathway 'hsa04710'. To test if the time of day of the milk sample collection study visit explained the relationship between *PER2* expression and expressed milk volume, we transformed the time of the study visit into a quantitative variable with the R package 'lubridate'.[92] *PER2* expression values from a variance-stabilizing transformation of the sample-by-gene count matrix in DESeq2[71] were used, including sample RNA mass and RIN as covariates. Regression models were calculated with 'lm' in R. Study time of day was correlated with *PER2* expression in a linear regression ($p = 0.02$), but not with milk volume expressed ($B = -0.03$, $p = 0.4$) We then ran the following linear models:

PER2 expression $\sim$ milk volume + [technical covariates].
PER2 expression $\sim$ milk volume + time of study visit + [technical covariates].

The same technical covariates included in differential gene expression testing were included here (batch, number of genes detected, RIN, RNA concentration, mean 3′ bias). These two linear models were compared by an F-test via the 'anova' command in R to test if adding the time of study visit term to the model provided a better fit to the data. This test ($p = 0.06$) suggested that adding the time of study visit variable did not provide a substantially better fit to the data. We used the 'check_model' function from R package 'performance'[93] to ensure that these models fit the linear regression model assumptions (Figure S8).

Deconvolution of bulk transcriptomes with bisque
Raw gene counts (MIT_Milk_Study_Raw_counts.txt.gz) and metadata (MIT_milk_study_metadata.csv.gz) were downloaded for the Nyquist et al. study[17] from the Broad Insitute Single Cell Portal (https://singlecell.broadinstitute.org/single_cell/study/SCP1671/cellular-and-transcriptional-diversity-over-the-course-of-human-lactation) on 6/3/2022. Count data was filtered to keep just one sample per participant, requiring samples to have been collected >14 days and <3 months postpartum, leaving 10 samples. The count matrix and associated metadata was then formatted as a Bioconductor 'ExpressionSet' object, combining the two macrophage cell type annotations from Nyquist et al. into one cell type called just "Macrophage" and resulting in 8 cell type annotations. The milk gene-level count data was then loaded into an ExpressionSet object and Cell type deconvolution was run with the R package "BisqueRNA" and the function 'ReferenceBasedDecomposition', with parameters "markers = NULL" and "use.overlap = F". Bisque[27] used 19,387 genes present in both the bulk and single-cell expression sets. To generate the heatmap in Figure 1F, for each of the 8 cell types, sample cell type proportion estimates were regressed against all 8 traits (gestational diabetes status, gestational weight gain, maternal pre-pregnancy BMI, milk glucose, milk IL-6, milk insulin, milk volume expressed, parity) and technical covariates (RNA concentration, RIN, sequencing batch, number of genes detected, and mean 3′ bias) using the 'glm' function in R. The coefficients plotted are the regression coefficients for each trait for a given cell type from this multiple regression model.

Milk eQTL analysis
Gene-level quantifications were filtered for the 230 unrelated individuals with RNA-seq and genotype data. Genes were filtered to retain those with ≥ 6 counts and and TPM >0.1 in at least 20% of samples, leaving 17,672 genes of the original 45,473. TPM quantifications were then rank-normalized with the 'RankNorm' function in R package RNOmni,[94] and gene coordinates were added using annotations from R package 'EnsDb.Hsapiens.v86'. Genes without coordinate annotations, mitochondrial, and Y chromosome genes were removed, leaving 17,302 genes used in eQTL analyses.

The APEX toolkit was used for *cis*-eQTL analysis (https://corbinq.github.io/apex/doc/).[76] First, 50 latent factors from the gene expression matrix were calculated using command 'apex factor' with 10 iterations. cis eQTL analysis was run with the command 'apex cis' with 3 genetic PCs (calculated with 1000 Genomes samples, described above) and 45 gene expression latent factors as covariates. The 45 latent factors were correlated with batch and other quality control metrics of the RNA-seq data (Figure S10). We used APEX's linear mixed model with a genetic relatedness matrix calculated as above in PLINK, and with distance to start site weighting for eGene p values (ACAT-dTSS). SNPs with minor allele frequency >1%, missing genotype information <5%, and within 1 Mb of the gene transcription start site were included. The command used was as follows:

apex cis –bcf [genotypes bcf file] –bed [gene expression bed file] –cov [genetic PCs + gene expr. LFs covariate file] –grm [genetic relatedness matrix] –prefix [output file prefix] –long –dtss-weight 0.00001.

APEX uses an aggregated Cauchy association test to calculate a gene-level p value, and can use the distance to TSS weighting to improve discovery power (parameter '–dtss-weight' in the command above). eGene p values were adjusted for multiple tests using a Benjamini-Hochberg correction.[91]

To assess the impact of RNA quality (as measured by RIN) on our eQTL results, we ran the eQTL scan on the top and bottom half of samples by RIN separately, as well as a random subset of the same size ($N = 115$). eGene p values were strongly concordant across all pairs of subsets and the entire $N = 230$ sample ($p < 2 \times 10^{-16}$; Figure S11), but with larger p values in the sample subsets reflecting the reduced power of a smaller sample size. Thus, we concluded that the lower RIN score samples in our eQTL analysis improved our power and should be included.

Conditional analysis of milk eQTLs were also run in APEX using the same covariates (3 genetic PCs, 45 gene expression latent factors) and the '–stepwise' flag:

apex cis –bcf [genotypes bcf file] –bed [gene expression bed file] –cov [genetic PCs + gene expr. LFs covariate file] –prefix [output file prefix] –long –dtss-weight 0.00001 –stepwise.

Colocalization of milk and GTEx eQTLs
eQTL summary statistics for single tissues (*.v8.allpairs.txt.gz), and gene eQTL summary (*.v8.egenes.txt.gz) were downloaded from the GTEx portal (https://gtexportal.org/). For each gene with an eQTL in milk at q value < 5%, each GTEx tissue with a significant eQTL (q value < 5%) was identified, and colocalization between the milk and GTEx tissue performed with the *coloc* R package[37,77]: *cis*-eQTL summary statistics for milk and each GTEx tissue with an eGene were filtered for those present in both milk and GTEx, within 200 kilobases of a top SNP of any tissue, and effect estimates harmonized so the reference/alternative alleles matched. LD matrices for these SNPs were generated using PLINK's '–r square' function with our genotyping data and using the European ancestry subset of the 1000 Genomes dataset ($N = 503$). eQTL signals for each tissue were fine-mapped using the 'runsusie' command, using the milk study LD reference for milk eQTLs and the 1000 Genomes LD reference for GTEx tissues. Colocalization was run between milk and each GTEx tissue with the command 'coloc.susie'[36] with a prior probability of colocalization of $p_{12} = 3.5 \times 10^{-5}$. This prior was chosen to require a lower burden of evidence for colocalization than the default value in coloc ($p_{12} = 1 \times 10^{-5}$), as here we are most interested in identifying milk-specific eQTLs and analyses of the GTEx project has demonstrated that most eQTLs are shared across tissues.[95] Coloc.susie tests for colocalization between each pair of fine-mapped signals between the two tissues, and thus there will be multiple tests if fine-mapping identifies more than one signal for a particular tissue/gene pair. Each colocalization test was designated as 'colocalized' if the ratio PP.H4/(PP.H4+PP.H3) > 0.8; as 'not-colocalized' if the ratio PP.H3/(PP.H4+PP.H3) > 0.8; and 'ambiguous' otherwise.

Each fine-mapped milk eQTL signal was designated as milk-specific if either of these criteria were met: (1) there were no GTEx tissues with a significant eQTL for the gene (q value < 5%), or (2) there were no tissues with an eQTL signal that colocalized with the milk signal, and at least 75% of tested tissues' eQTLs were categorized as not-colocalized. Of the 2,790 milk eGenes, 18 did not have an eQTL in any GTEx tissue, 401 failed at fine-mapping either the milk or GTEx signals, 1,907 had all eQTL signals colocalize with a GTEx eQTL, and 464 had at least one milk-specific eQTL signal. Enrichment analysis of genes with milk-specific eQTLs ($N = 482$) or tissue-shared eQTLs was performed with the 'enrichGO' command from the R package 'clusterProfiler',[96] using a background gene list of all tested milk genes (17,302 genes) with a minimum gene set size of 10 and maximum size of 250.

Overlap between milk eGenes and dairy cattle QTL
Cattle gene coordinates for ARS_UCD1.2 genome were downloaded from https://bovinegenome.elsiklab.missouri.edu/downloads/ARS-UCD1.2, filtered for mRNAs, and for each gene with multiple entries the entry with the largest region was retained. Dairy cattle QTL were downloaded from the animalQTLdb (https://www.animalgenome.org/cgi-bin/QTLdb/index) by selecting "All data by bp (on ARS_UCD1.2 in bed format)".

For each of 4 milk-related traits, we selected QTL with the following trait labels: milk yield (Milk yield, 305-day milk yield, Average daily milk yield), milk somatic cell count (Somatic cell score, Somatic cell count), milk protein (Milk protein percentage, Milk protein yield, Milk protein content), and milk fat (Milk fat percentage, Milk fat yield, Milk fat content). To identify a smaller list of genes identified in QTL from multiple studies, as some of these traits' QTL overlapped thousands of genes, we identified genes that overlapped at least 1 QTL for all 4 dairy cattle milk traits ($N = 1,035$ genes, Table S11).

To test for enrichment of milk-specific vs. tissue-shared eQTL genes in these lists, we filtered milk eGenes for those that were present in the dairy cattle genome annotation above and that had a milk-specific eQTL ($N = 146$) vs. only tissue-shared eQTLs ($N = 591$). We performed a two-sided Fisher's exact test where the 2×2 contingency table axes were: (A) milk-specific vs. tissue-shared eGenes (from our human milk eQTL analysis), and (B) cattle QTL overlapping genes vs. cattle QTL nonoverlapping (from the gene lists identified above), using the 'fisher.test' command in R.

Comparison of milk and GTEx eQTL with mash
We applied Multivariate Adaptive Shrinkage (*mash*) using the mashR package[40] to assess patterns of eQTL sharing across milk and GTEx eQTLs. *mash* is an empirical Bayesian method that utilizes the covariance structure across conditions (in this application,

tissues) to identify tissue shared or unique eQTL. We first identified the 13,593 genes that had eQTL summary statistics across all GTEx tissues and milk, as summary statistics from all tissues are required to run *mash*. Then, following the analysis outlined at https://stephenslab.github.io/mashr/articles/eQTL_outline.html, we extracted a 'random' matrix of summary statistics for 48 GTEx tissues and milk for 10,000 random gene/variant pairs. The 'strong' matrix was defined as the variant effects from all tissues for (1) the variants with the lowest milk eQTL *p* value for the 2,261 milk eGenes in this dataset; and (2) for each GTEx tissue, the variants with the lowest *p* value for 1000 random eGenes for that tissue. In total the 'strong' matrix contains summary statistics for 42,677 gene/variant pairs across 48 GTEx tissues and milk. From these input data we (1) estimate correlation structure from the 'random' matrix; (2) estimate data-driven covariances from the 'strong' matrix; (3) fit the *mash* model on the 'random' matrix using the data-driven and canonical covariances; and (4) estimate posterior summaries for the 'strong' matrix, i.e., re-calibrated effect estimates and statistical significance for each gene/variant pair in each tissue (Table S12). Using the output posterior summaries, we then calculated the fraction of milk eQTL effects that were shared with each GTEx tissue using the default criteria in *mashR*: local false sign rate <0.05, same direction of effect, and effect estimates within a factor of 2. This proportion of shared milk eQTL is plotted for a subset of GTEx tissues in Figure 2E. These tissues were chosen to represent the full range of similarity/dissimilarity to milk while not displaying all tissues for clarity of presentation. Results for all tissues are shown in Figure S13.

Colocalization of milk eQTLs with breast cancer GWAS summary statistics

GWAS summary statistics from Zhang et al.[43] (icogs_onco_gwas_meta_overall_breast_cancer_summary_level_statistics.txt.gz) were downloaded from the BCAC website (http://bcac.ccge.medschl.cam.ac.uk/). Coordinates were converted to hg38 with LiftOver, and the meta-analysis summary statistics for all breast cancers were used (column names 'Beta.Meta', 'p.meta', etc.). For each milk eGene, colocalization was performed if there was a breast cancer GWAS hit of $p < 5 \times 10^{-8}$ within the eQTL window (within 1 Mb of gene TSS). Breast cancer GWAS and milk eQTL summary statistics were filtered to variants within 200 kb of the smallest milk eQTL *p* value, and statistics harmonized so the reference/alternative alleles matched. An LD matrix for these variants was calculated using (1) our milk study data and (2) the European ancestry subset of the 1000 Genomes European reference ($N = 503$). The milk and breast cancer GWAS signals were fine-mapped using 'runsusie' in the *coloc* R package,[36,37,77] using the milk LD reference for the milk eQTLs and the 1000 Genomes LD reference for the breast cancer signals. Colocalization was run with the command 'coloc.susie' with a prior probability of colocalization of $p_{12} = 5 \times 10^{-6}$. We chose this prior based on the recommendation in Wallace.[97]

Correlations between milk gene expression and oligosaccharides

HMOs were rank normalized within the 310 individuals with both gene expression and HMO data, using the 'RankNorm' function from R package 'RNOmni'.[94] For HMOs absent in non-secretors (2′FL and DFLac; Figure S21), we included only secretor individuals ($N = 231$). The following HMO categories were also calculated: the sum of all HMO concentrations, the sum of all sialylated HMO concentrations (DSLNH, DSLNT, FDSLNH, LSTb, LSTc, 3′SL, 6′SL), and the sum of all fucosylated HMO concentrations (DFLNH, DFLNT, FDSLNH, FLNH, LNFP-I, LNFP-II, LNFP-III, 3′FL, DFLac, 2′FL). These HMO category sums were rank normalized across all individuals.

We used edgeR[74] to test for correlations between milk gene expression and HMO concentrations. The count matrix and metadata were loaded into an edgeR object and "filterByExpr" was used to remove lowly expressed genes, leaving 11,780 genes (or 11,695 genes for secretors only). For each HMO, we then used the 'estimateDisp' function on a design matrix regressing gene expression across HMO concentration, secretor status (except for when only secretors were included, i.e., 2′FL and DFLac), HMO batch, sequencing batch, RIN, RNA concentration, number of genes detected, and mean 3′ bias. We then used 'glmQLFit' to fit a quasi-likelihood negative binomial generalized log-linear model to the count matrix and design model, and 'glmQLFTest' to perform a quasi-likelihood F-test of each gene against each tested HMO. We used Benjamin-Hochberg[91] correction of *p* values across all HMO-gene pairs.

We tested for gene ontology enrichment of significant genes (q value < 10%) for each trait using the R package topGO, with all tested genes as the background gene list. We used the 'resultFisher' function to run a Fisher's exact test for each gene ontology, and used a Benjamini-Hochberg correction[80] for all ontologies ($N = 14,034$) across the 15 HMOs/HMO categories with at least 50 significant genes (Table S16).

Genetic associations at milk eQTLs with milk oligosaccharides

The list of candidate genes to test for effects of milk eQTLs on HMO concentrations was downloaded from Supplementary Dataset 2 in Kellman et al.[48] From this gene list, we identified 7 genes with significant eQTLs in our dataset (q value < 5%). To test for genetic associations between the lead variant identified by fine-mapping above (all 7 genes had only one signal detected) at each milk eGene and HMO concentrations using rank-normalized HMO concentrations. For 2′FL and DFLac, which were absent in non-secretors (Figure S21), we rank-normalized the concentrations within secretors and scaled concentrations in non-secretors to have mean −3 and s.d. 0.1, to avoid introducing variation that did not exist in non-secretors. We used 'glm' in R to fit a model with HMO concentrations as the outcome, including genotype, secretor status, HMO batch, and the first three genetic PCs as covariates:

HMO ∼ genotype + secretorStatus + HMO batch + PC1 + PC2 + PC3.

For models of HMOs vs. *FUT2* eQTL genotype, we excluded the secretor status term. Genotype vs. HMO concentration plots in Figures 3D and 3E show the residual HMO concentration after regressing out HMO batch and the first 3 genetic PCs. For Figure 3E, secretor status was also regressed out of the plotted FLNH concentrations.

To estimate the effect of modified milk gene expression on HMO concentrations, we used a Wald Ratio, which estimates the causal effect between an exposure (milk gene expression) and outcome (HMO concentration) by dividing a single genetic variant's effect on outcome by the genetic effect on the exposure.[98]

Processing of infant fecal metagenomes

Infant fecal collection and storage, and metagenomic DNA extraction were described previously.[81] Briefly, feces were collected from diapers either during study visits and frozen at −80°C immediately, or collected at home, stored in 2 mL cryovials with 600 μL RNALater (Ambion/Invitrogen, Carlsbad, CA), and stored at −80°C after shipping to the lab at the University of Minnesota. DNA was extracted using the PowerSoil kit (QIAGEN, Germantown, MD), eluted with 100 μL of the provided elution solution, and stored in microfuge tubes at −80°C.

Extracted DNA was used to construct libraries for metagenomics sequencing using the Illumina Nextera XT kit (Illumina, San Diego, CA, United States). Metagenomics libraries were then sequenced on an Illumina NovaSeq system (Illumina, San Diego, CA) using the S4 flow cell with the 2 × 150 bp paired end V4 chemistry kit by the University of Minnesota Genomics Center, achieving a sequencing depth of ∼4.5 million reads per sample.

Microbial taxon abundances were generated by first processing metagenomic fastq files with Shi7 version 1.0.1,[99] which learns optimal quality control parameters from the data. Sequences were then trimmed, filtered by quality scores, and stitched per the learned parameters in Shi7. Sequences from all samples were multiplexed into a single fasta file for downstream processing. Processed sequences were aligned to reference databases using BURST version 0.99.7f,[78] using a reference genome database generated from GTDB r95 (https://gtdb.ecogenomic.org/stats/r95). A 95% identity cutoff and forward/reverse complement flag were used. Resulting.b6 files were converted to reference and taxonomy tables using embalmulate[78] with 'GGtrim' activated. To generate microbial pathway abundances, metagenomic sequences were run through the MetaPhlAn[79] version 3.0.7 pipeline, with BowTie2[100] version 2.4.2 64-bit, DIAMOND[101] version 0.9.24, and MinPath[102] version 1.5.

To generate the PCA of infant metagenomes in Figure 4A, data were filtered to include only taxa with relative abundance >0.001 in at least 10% of 1-month or 6-month samples. A centered log-ratio transformation was performed on the relative abundances of each sample, and principal components were calculated with the 'prcomp' command in R.

Sparse CCA of human milk transcriptomes and infant fecal metagenomes

Input datasets were prepared as follows.

Milk gene expression

To prepare gene expression data for this analysis, the sample-by-gene count matrix was loaded into DESeq2,[71] filtered to keep only protein-coding genes with count > 0 in at least half the participants (14,905 genes), and transformed using the variance stabilizing transformation. After this transformation, the variance of each gene was calculated across all samples and genes in the lowest 25% variance were removed, leaving 9,421 genes.

Infant fecal metagenomes

Taxon abundances and pathway abundances from 1- and 6-month infant fecal samples were processed separately. The taxon relative abundance matrix was filtered to retain species-level taxa only, keeping only species with a relative abundance $>1 \times 10^{-3}$ in at least 10% of samples (92 species for 1-month and 82 species for 6-month samples). A centered log-ratio transformation was then performed on each sample's relative abundances. For microbial pathways, species-specific and unclassified pathways were removed, leaving 241 pathways for 1-month and 216 pathways for 6-month samples. The species and pathway level information from both timepoints was then combined into one matrix.

Each dataset was filtered for the 146 individuals with both 1- and 6-month infant fecal metagenomes and 1-month milk gene expression. Sparse canonical correlation analysis (sparse CCA), to identify sparse components maximizing correlation between the milk gene expression and infant fecal metagenome datasets, and enrichment analyses of genes in each sparse component, were performed as previously described,[53] using k = 15 components. Code was downloaded from https://github.com/blekhmanlab/host_gene_microbiome_interactions. Significance of the sparse components was calculated with leave-one-out cross-validation, and 12 components were retained at Benjamini-Hochberg q value < 10%. Pruning significant components whose scores across mother-infant pairs were correlated at Pearson's r > 0.5 left 7 remaining sparse components (Figure 4B). Pathway enrichment was performed separately on positively weighted and negatively weighted genes for each component.

To generate network interaction plots between milk-expressed genes and infant fecal microbes identified in the sparse CCA analysis, for each significantly enriched pathway (q value < 10%) in a component, we (1) filtered for overlapping genes between the component and pathway; (2) generated a pairwise correlation matrix of mother-infant pairs' trait values for those genes, the top 3 microbiome traits in the component with positive weights, and top 3 microbiome traits with negative weights; (3) pruned for correlations with Pearson's r > 0.3 and p < 0.05; (4) generated a network plot from the pairwise correlation matrix using the 'ggnetwork' package in R.[103]

B. infantis growth rates were estimated using Compute PTR (CoPTR).[61] We aligned the infant gut metagenomic shotgun reads to the *B. infantis* ATCC 15697 reference genome, downloaded from NCBI, using bowtie2 v2.2.4.[100] We then used CoPTR to get coverage information for each mapped sample, filtering for samples with at least 75% coverage and at least 3000 mapped reads to the *B. infantis* genome. For samples that passed these filters, CoPTR was used to estimate the peak-to-trough ratio (PTR) from the coverage information, an estimate of the bacterial growth rate.