SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
GENE EXPRESSION ANALYSIS
DIAGNOSTIC MARKERS
COLON
BIOINFORMATICS

Received
28 March 2014

Accepted
9 June 2014

Published
26 June 2014

Correspondence and requests for materials should be addressed to R.S.C. (r-chapkin@tamu.edu)

* Current address: Ayasdi, 4400 Bohannon Drive, Suite #200, Menlo Park, CA 94025

Non-invasive analysis of intestinal development in preterm and term infants using RNA-Sequencing

Jason M. Knight[1,3], Laurie A. Davidson[2,3], Damir Herman[6]*, Camilia R. Martin[7], Jennifer S. Goldsby[2,3], Ivan V. Ivanov[5], Sharon M. Donovan[8] & Robert S. Chapkin[2,3,4]

[1]Department of Electrical Engineering, Texas A&M University, College Station, TX, [2]Department of Nutrition & Food Science, Texas A&M University, College Station, TX, [3]Center for Translational Environmental Health Research, Texas A&M University, College Station, TX, [4]Department of Veterinary Integrated Biosciences, Texas A&M University, College Station, TX, [5]Department of Veterinary Physiology and Pharmacology, Texas A&M University, College Station, TX, [6]Winthrop P. Rockefeller Cancer Institute, University of Arkansas for Medical Sciences, Little Rock, AR, [7]Department of Neonatology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, [8]Department of Food Science & Human Nutrition, University of Illinois, Urbana, IL.

The state and development of the intestinal epithelium is vital for infant health, and increased understanding in this area has been limited by an inability to directly assess epithelial cell biology in the healthy newborn intestine. To that end, we have developed a novel, noninvasive, molecular approach that utilizes next generation RNA sequencing on stool samples containing intact epithelial cells for the purpose of quantifying intestinal gene expression. We then applied this technique to compare host gene expression in healthy term and extremely preterm infants. Bioinformatic analyses demonstrate repeatable detection of human mRNA expression, and network analysis shows immune cell function and inflammation pathways to be up-regulated in preterm infants. This study provides incontrovertible evidence that whole-genome sequencing of stool-derived RNA can be used to examine the neonatal host epithelial transcriptome in infants, which opens up opportunities for sequential monitoring of gut gene expression in response to dietary or therapeutic interventions.

Growth and maturation of the gastrointestinal tract occurs on a set developmental ontogeny[1,2]. Although some aspects of intestinal development appear to be hardwired and do not emerge until a specific gestational age[3], others are influenced by dietary intake[1] and microbial colonization[4]. Accordingly, the early neonatal period is a critical phase for both intestinal digestive development as well as establishment (colonization) of the intestinal microbiota[5]. Functional immaturity of the gut compromises nutrient absorption and utilization in the preterm[3]. This digestive immaturity, coupled with immature mucosal barrier function, immune response and inappropriate bacterial colonization, makes premature neonates particularly susceptible to intestinal inflammation and injury and the development of necrotizing enterocolitis (NEC), a potentially lethal bowel disorder[3,6–8].

The intestine is lined by epithelial cells that process nutrients and provide the first line of defense against pathogens. Because colonization of the intestine with non-pathogenic, or commensal, bacteria is vital for infant health, it is important to understand how epithelial cells and the microbial ecosystem are modulated by diet and disease[9]. Therefore, our on-going efforts are directed at understanding the regulation of neonatal development by components present in human milk, as it is the gold standard for infant nutrition[10–12]. Advancing knowledge in this area has been limited by an inability to directly assess epithelial cell biology in the healthy newborn intestine. Previously, to that end, we developed non-invasive high throughput microarray techniques to examine intestinal gene expression in stool samples[11] and to probe the cross-talk between host gene expression and the microbiota[12]. This methodology has the advantage of using exfoliated cell mRNA directly isolated from feces, which contains sloughed small intestinal and colon cells, and therefore does not require invasive procedures or discomfort to the subject.

In this study, we applied our non-invasive methodology, which has been optimized in the term infant, to the extremely preterm infant (24–30 weeks gestational age). With respect to neonatal health, it is well known that extremely preterm infants are a vulnerable population whose intestinal and immune development is modulated early in the postnatal period by multiple environmental factors, such as medications, indigenous microorganisms

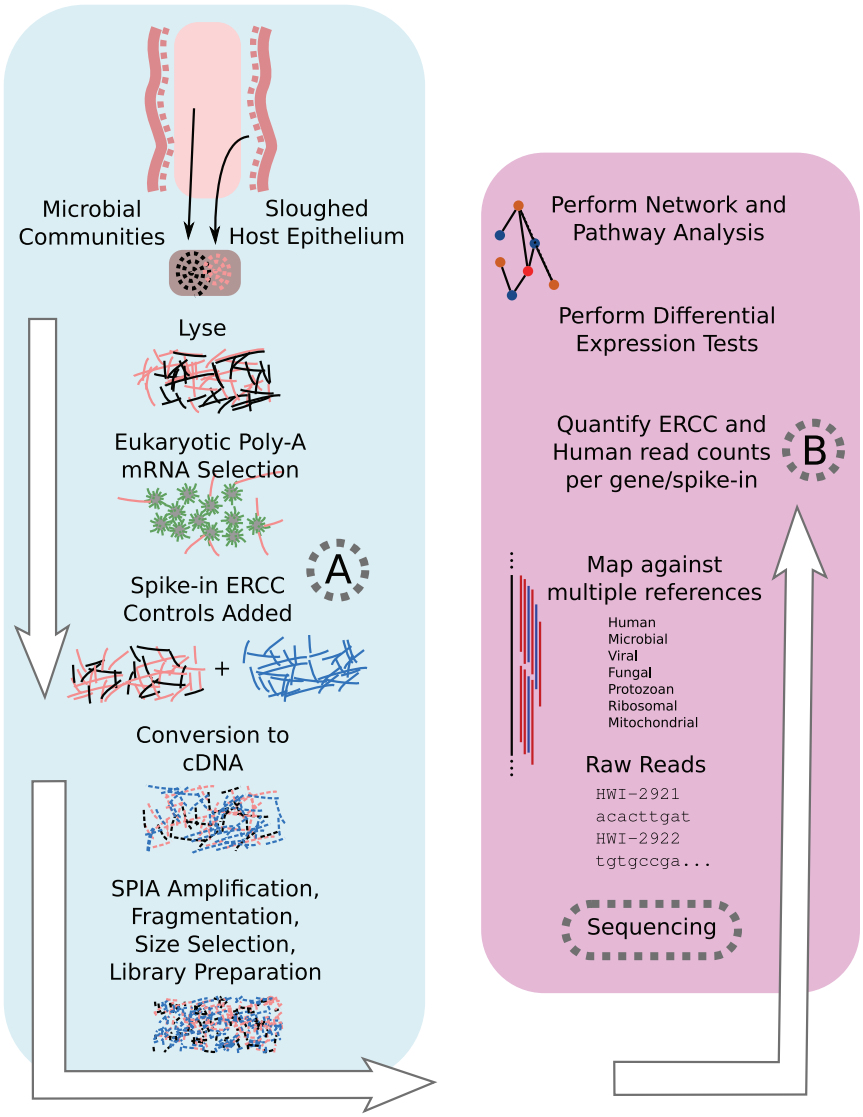

Figure 1 | Fecal samples were processed to enrich eukaryotic mRNA, develop libraries, and assess bioinformatic sequencing content. The steps in cyan were applied to mRNA processing and the magenta steps were applied to sequenced read data. Step A, ERCC controls were spiked-in to determine processing efficacy and reproducibility up to and including step B.

in the intensive care unit, and limited enteral diet[13]. Using global transcriptome RNA Sequencing (RNA-Seq) profiling, we compared host transcript abundance and alternative splicing in healthy term and extremely preterm infants. This novel application of RNA-Seq to measure host gene expression has, for the first time, provided insight into host responses to dietary and environmental influences in the early neonatal period.

Results

Fecal samples were obtained from three term and three preterm infants and enriched for host mRNA transcripts through poly A^+ RNA selection. We also considered the effects of pooling by sequencing a sample consisting of multiple term infants. Approximately 50 M reads were sequenced for each individual and 30 M reads for the pooled sample, and on average 5075 human genes were detected above an FPKM threshold of one despite 55–90% of the reads mapping to a representative set of microbial genomes (Supplementary Table 1). Complementary analyses were conducted using ERCC spike-in RNA controls, sample correlations, and qPCR in order to validate the host mRNA transcript data.

In general, transcripts consisted of hundreds of reads that accumulated in a narrow region in or close to the 3′ UTR. A typical example is RefSeq id NM_000482, representing apolipoprotein A-IV (Supplementary Figure 1). Since bacterial RNA in stool is much more abundant than human RNA[14], the first objective was to ascertain the efficacy of host poly A^+ RNA enrichment, library preparation, and sequencing procedures. We note that poly A^+ RNA enrichment precludes the use of these data to probe microbiome diversity or transcript levels due to the bias this procedure might place on the microbial transcripts. ERCC spike-in controls added at step A in Figure 1 were used to verify the integrity of sample handling from steps A to B. The ERCC controls reflect a diverse set of sequence content and length, have low homology with eukaryotic transcripts and span a large range of concentrations. The amount of observed reads from the six individual samples, which mapped to known ERCC transcripts, is plotted in Figure 2. The observed ERCC reads were highly correlated between samples (all Pearson correlation coefficients > 0.998, all Spearman correlation coefficients > 0.992). Additionally, each sample correlated strongly to the known concentrations (all Pearson correlation coefficients $>$ 0.88, all Spearman correlation coefficients > 0.9, see methods for

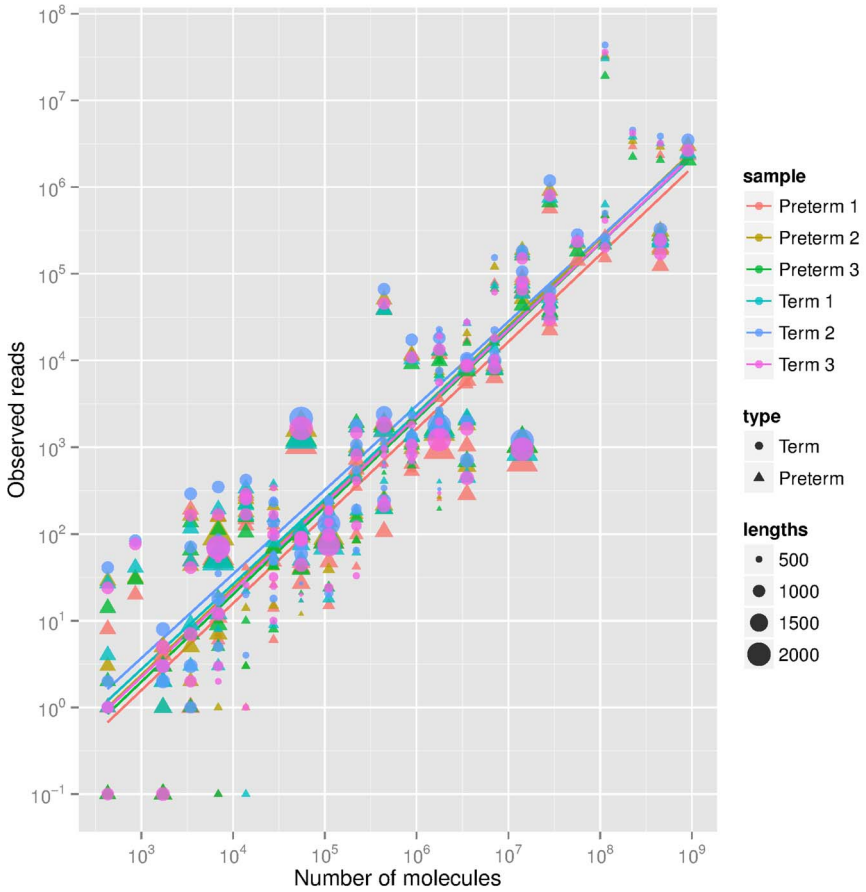

Figure 2 | Observed reads were mapped against known ERCC reference sequences, and read counts were compared against known amounts added in the spike-in control. High correlations between samples (all Pearson correlation coefficients > 0.998, all Spearman correlation coefficients > 0.992) and to known concentrations (all Pearson correlation coefficients > 0.88, all Spearman correlation coefficients > 0.9, see methods for details) indicate that the sequencing and mapping procedures are effective and reproducible across a variety of transcript lengths.

details). Furthermore, using known quantities and concentrations of exogenous ERCC transcripts, we observed reads from mRNA species present at amounts as low as 800 molecules in our samples.

We subsequently validated RNA-Seq against qPCR as in[15]. For this purpose, eleven differentially-expressed genes, as detected by RNA-Seq, were selected based on the fact that in the RNA-Seq data set expression of each gene in all term infants was either higher or lower than each gene in all preterm infants. In addition, we selected genes with FPKM greater than 10 since genes expressed at a lower level are typically difficult to detect by qPCR. Nine of the eleven genes measured (Figure 3) demonstrated fold-changes of similar magnitude and direction by qPCR and RNA-Seq (Supplementary Table 2), while SLC2A1 and RPS16 differed in directionality of differential expression. By comparison, a similar RNA-Seq validation using qPCR in the literature observed four out of five DE genes replicated[16]. In addition, the average Spearman correlation coefficient between qPCR and RNA-Seq fold changes calculated was 0.586, and average Pearson correlation coefficients were 0.570 with an average line of best fit with slope near unity at 0.869 (Supplementary Figure 2).

To obtain a benchmark for comparison, the average quantified gene expression for three preterm and three term infants were compared. Overall, global transcriptional profiles compared favorably with high correlation for highly expressed genes (Figure 4). In addition, expression correlation between individual samples verified the presence and detection of human epithelial mRNA transcripts in the fecal samples (Supplementary Figure 3, Figure 5a). Examples of genes associated with specific intestinal cell types included absorptive enterocytes (lactase, 7.4-fold higher expression in preterm than term and sucrase – isomaltase, 1.7-fold higher expression in term than preterm), Goblet cells (muc-2, 1.3-fold higher expression in preterm than term), enteroendocrine cells (chromogranin A, 12-fold higher expression in term than preterm) and Paneth cells (lysozyme, 1.6-fold higher expression in preterm than term). In some cases where small volumes of stool are available, pooling of samples may be necessary. Therefore, the effect of sequencing pooled samples versus individual samples from term infants was assessed. Individual samples and their pooled counterparts exhibited homogeneity as determined by Spearman correlation coefficients plotted in a heatmap (Figure 5a). Interestingly, the preterm individuals appeared more heterogeneous amongst themselves and as compared to the term samples. This is potentially a result of the larger clinical variation of the preterm samples in terms of nutrition and gut health. In addition, when considering genes expressed at an arbitrary cutoff threshold (>10 Fragments per Kilobase per Million [FPKM]), the number of genes expressed (Figure 5b) in common among the four term infant samples (three individual and one pooled) was larger than the number of genes expressed in common by any subset of these samples. Therefore, the term individual samples and the pooled sample were more similar than dissimilar.

Functional gene set enrichment analysis was performed using Ingenuity Pathway Analysis (IPA) pathway profiling[17] to probe the biological relevance of the differentially expressed genes between preterm and term infants. For this purpose, differentially expressed genes with a p value cutoff <0.05 between groups were associated with canonical gene networks using the Ingenuity Knowledge base (Figure 6). Broad differences in genes associated with lipid metabol-

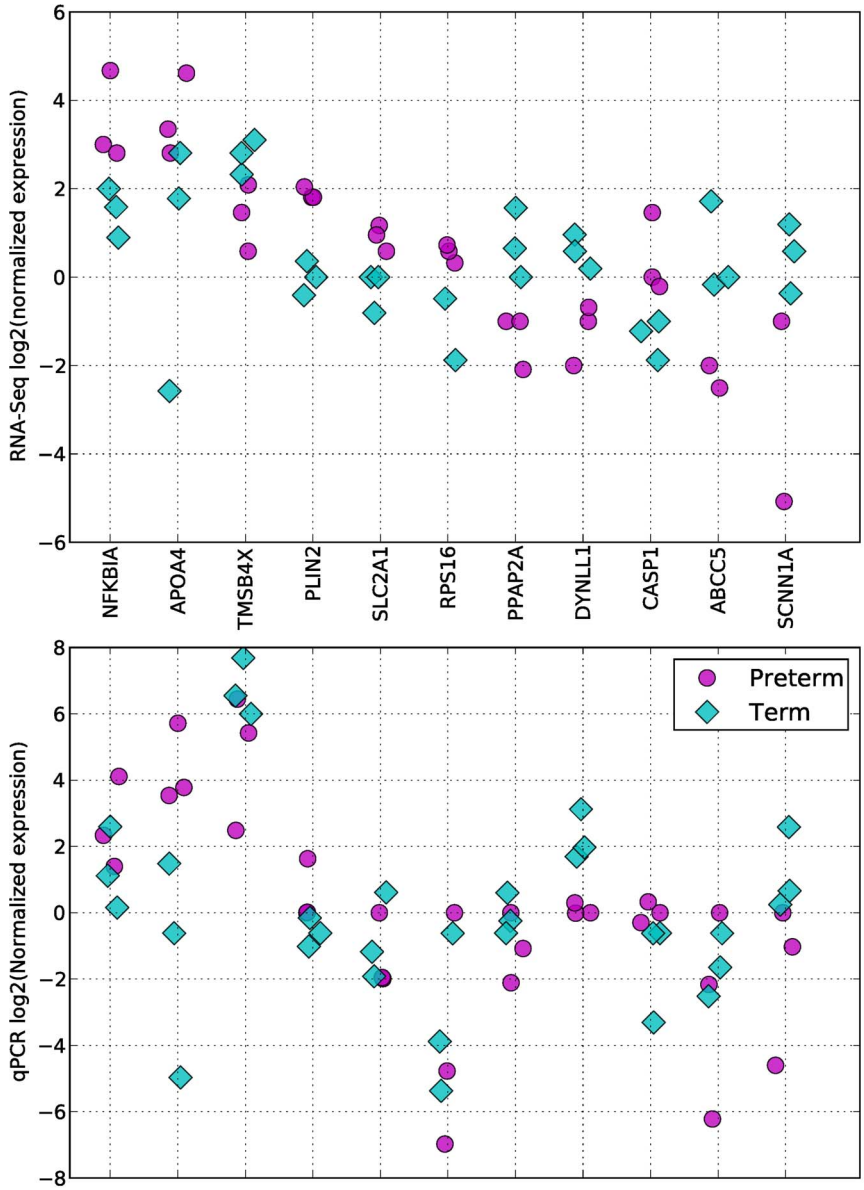

Figure 3 | Eleven differentially expressed genes were selected for validation using qPCR. Nine of the eleven genes exhibited average-fold changes in the same direction of change as qPCR, with SLC2A1 and RPS16 being the only exceptions. By comparison, a similar RNA-Seq validation using qPCR in the literature observed four out of five DE genes replicated[36]. Additionally, correlation analysis (method described in Figure 2 from[16]) (Supplementary Table 1 and Supplementary Figure 2) generated average Pearson correlation coefficients of 0.57 and average Spearman correlation coefficients of 0.59 with an average best fit slope line of 0.87.

ism, molecular transport, organismal injury, infectious disease and cellular development were observed. This is further highlighted in Table 1 (preterm > term expression) and Table 2 (term > preterm gene expression), where differentially-expressed genes associated with these networks are documented. preterm infants expressed numerous genes associated with immune cell function (e.g., CASP1, IL-1beta, IL-33, NFkB1A, S100-A9, SOCS3, and TREM-1) and lipid metabolism (e.g., ApoA1, ApoA4, ASAH1, MTM1 and PLIN2) (Table 1). In contrast, term infants exhibited highly up-regulated expression of genes associated with regulation of cell growth/cell cycle (e.g., CDKN2B, ESRRA, INSR, KREMEN2, MTRNR2L6, PDPK1 and TRIM36) (Table 2).

Discussion

Nutritional regulation of intestinal development begins *in utero* with exposure to protein-rich amniotic fluid and continues after birth with human milk and/or infant formula[1,10]. These developmental processes are essential for continued cellular differentiation of the gut and development of mucosal immunity[18]. In the healthy term infant, the continuum of enteral stimulation is continued postnatally, whereas the preterm infant is typically supported on parenteral nutrition with limited enteral stimulation in the first few weeks of life[3]. In addition, postnatal exposure to environmental organisms in the neonatal intensive care unit and the routine use of antibiotics can lead to aberrant intestinal development, microbial colonization and risk of intestinal disease in the preterm infant[3,6]. Hence, it is imperative to understand the transcriptional responses of the preterm gut so that specific nutritional practices can be employed in order to optimize intestinal development.

Sensitive noninvasive tests will become critical tools in tailoring nutritional interventions, including pre- and probiotics, in order to promote intestinal development and maturation in the growing infant. As part of this effort, our laboratory has developed a molecular methodology that utilizes stool samples containing intact

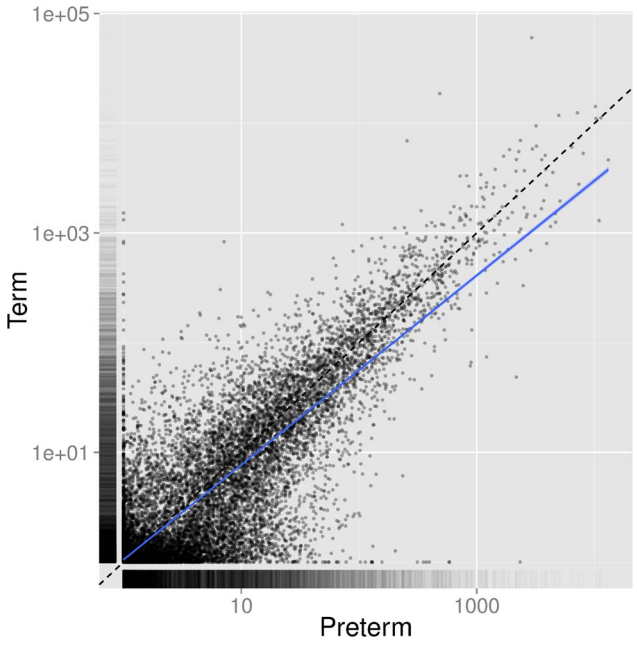

Figure 4 | Expression profiles of reads to mapped human genes show good between-group correlation on average. This indicates that the detection of similar expression profiles is likely from the similar tissue types present in both sets of samples.

sloughed epithelial cells in order to noninvasively quantify intestinal gene expression profiles in both the human infant[11,12] and adult[19]. Systems biology approaches, such as computational linear discriminant analysis (LDA), were used to identify the best single genes and two- to three-gene combinations for distinguishing term breast-fed versus formula-fed groups[11]. In addition, putative "Master" regulatory genes were identified using coefficient of determination (CoD) analysis[11]. Collectively, these approaches can be used to identify mechanistic pathways of intestinal development in the first few months of life, and to assess the impact of nutrition and other environmental exposures on the microbiome in the developing gut[12]. In this study, we have extended upon our initial observations by unraveling previously inaccessible complexities in the term vs preterm infant intestinal transcriptome by non-invasively interrogating the infant intestine using RNA-Seq, rather than gene microarray[11].

In order to develop a more comprehensive understanding of the complexity of transcriptome profiles in the intestine, we utilized neonatal stool samples containing intact sloughed epithelial cells and generated large-scale RNA-Seq genome profiles. For this purpose, poly A^+ mRNAs were first copied into DNA sequences, randomly sheared, attached to linkers and directly sequenced. Sequences were compared with the reference human genome, and the density of corresponding reads determined. Furthermore, using this form of global digital transcriptome profiling, we documented the host transcript abundance and alternative splicing in healthy term infants at 12-weeks postnatal age and extremely preterm infants (24–30 weeks gestational age) at 2–5 weeks postpartum. Although the precise origin of exfoliated cells is not known, results from our previous study[11], and reported herein, indicate that genes associated with discrete epithelial cell types (absorptive enterocytes, goblet cells, enteroendocrine cells, and Paneth cells) are detectable. Thus, it is likely that transcriptome signatures of both the small and large intestine can be monitored over time.

The examination of global alterations in gene expression offers insight into the effects of premature birth and the resultant influence of environmental exposures uniquely experienced by the preterm infant (e.g. antibiotics, other medications, prolonged period of parenteral nutrition) on intestinal mRNA profiles. RNA-Seq (validated by qPCR) revealed that following an enrichment process, reads from stool derived RNA are of human origin. Unlike RNA extracted from human cell cultures or surgical specimens, where the quality and quantity of RNA is usually high and RNA degradation can be controlled with tissue handling[20], infant stool represents a unique biological sample. Typically, expressed host transcripts consist of a narrow stretch of RNA that is rarely longer than several hundred bp. While exon-exon junctions in principle can be detected, we noted that less than 5% of transcripts exhibited their splice variants. In this respect, non-invasive gene expression analysis using infant stool

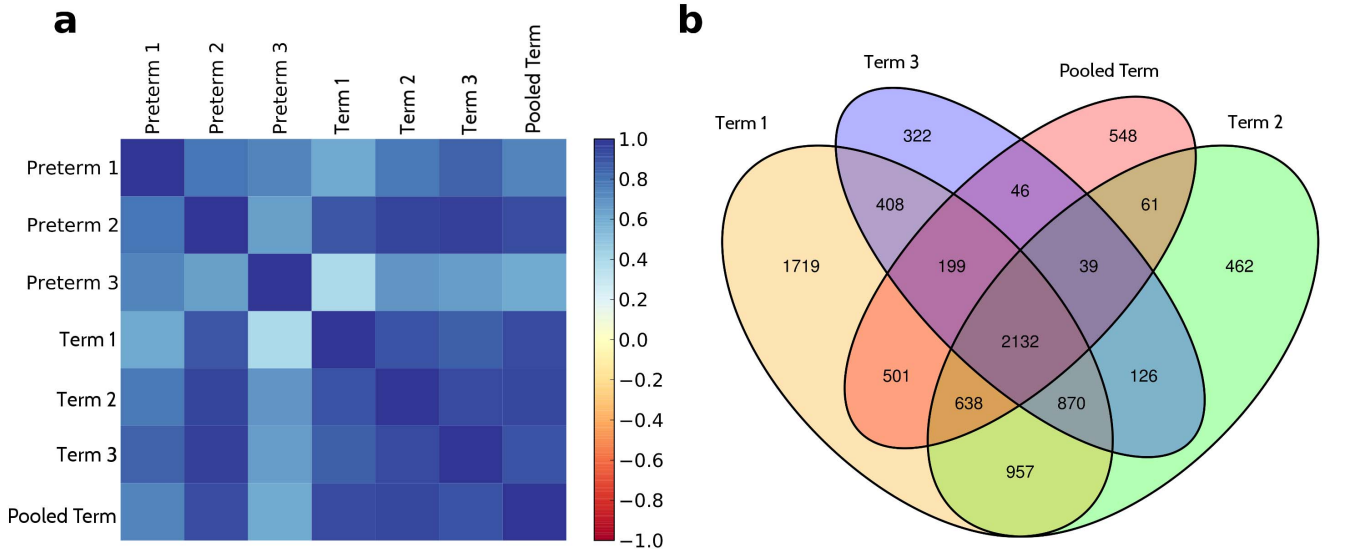

Figure 5 | (a) Pearson correlation coefficients among normalized mapped read counts (see Methods for details) for three preterm, three term individuals and a pooled term sample. Term samples are visibly correlated amongst each other, whereas higher heterogeneity amongst the preterm population is expected given their differing treatment regimens and developmental stages. A similar correlation heatmap using Spearman correlation coefficients is shown in Supplementary Figure 6 for completeness. (b) A four-way Venn diagram shows the number of expressed genes (>10 FPKM) among three individual term samples and a sequenced pooled sample. In the center of the diagram, 2132 genes are expressed in the pooled sample and all three individuals at greater than 10 FPKM. This large number of shared genes indicates that the sequencing procedure is consistent across sets of samples.

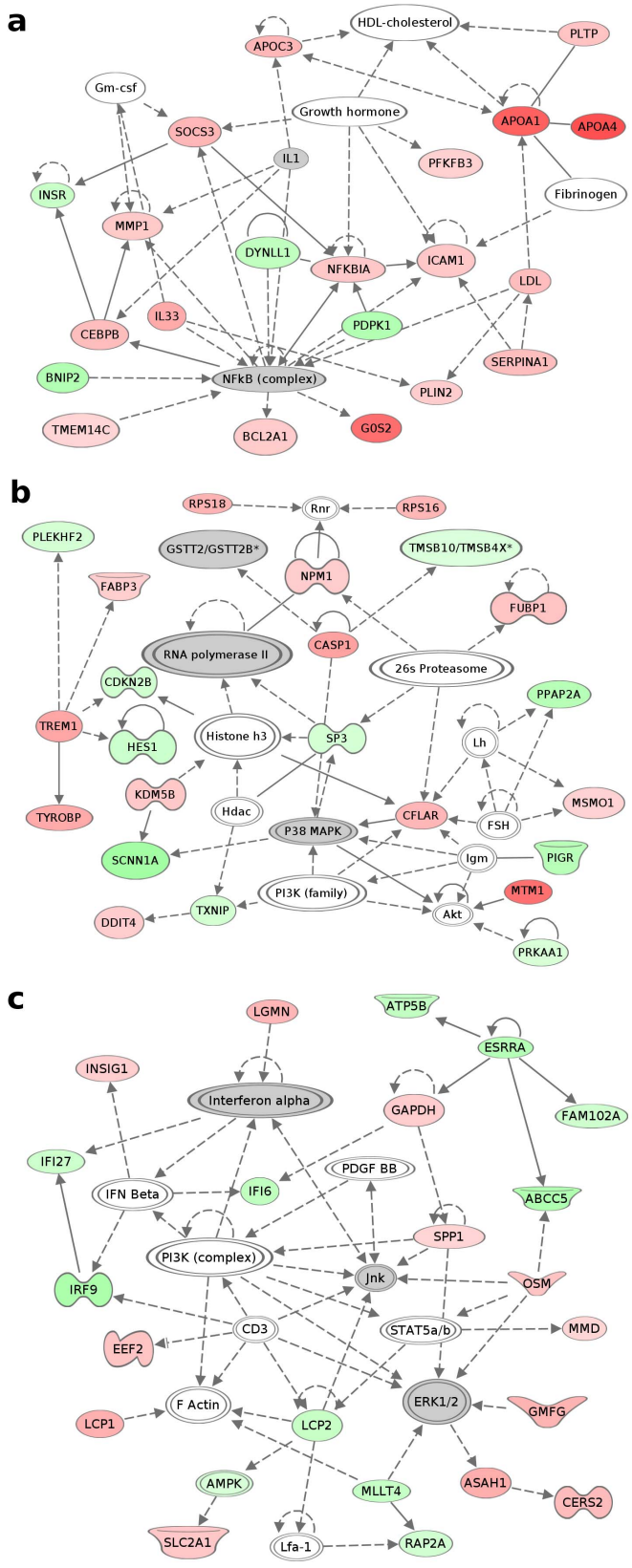

Figure 6 | Network enrichment analysis (Ingenuity IPA software) was performed using differentially expressed genes. Three networks of interest (a) Lipid metabolism, molecular transport, small molecule biochemistry; (b) Neurological disease, organismal injury and abnormalities, and infection disease; and (c) Cellular development, tissue development, lipid metabolism were generated. Red indicates higher expression in preterm infants and green indicates higher expression in term infants.

appears to be more challenging than analysis of formalin fixed, paraffin embedded (FFPE) tissue[21–23]. Unlike FFPE tissue blocks that can yield large quantities of fragmented DNA and RNA sufficient to explore complete topologies of expressed genes and local properties of DNA, infant stool requires careful enrichment of human RNA. Our experience with gene expression in the infant gut indicates that next generation sequencing provides a robust non-invasive glimpse into the host transcriptome. We expect that the development of novel methodologies for library preparation will allow us to further elucidate the physiology of the developing infant intestine.

For obvious reasons, directly examining the host epithelium in the human preterm infant is unlikely, as intestinal biopsies are not routinely performed unless medically indicated. Therefore, noninvasive methodologies currently provide us the best snap shot of infant gene expression[11] and host-microbe dynamic interactions[12] in an *in vivo* setting. Although future well-controlled studies are needed to evaluate environmental/dietary exposures, this study highlights the potential of using the described noninvasive technology. We evaluated genes that were over-expressed in preterm vs. term (Table 1) or term vs. preterm intestine (Table 2). Although none of the infants were clinically ill at the time the stool samples were collected, one of the major categories of genes overexpressed in preterm vs. term samples was immune function. Because of the vast number of exfoliated epithelial cells shed from the lining of the intestine on a daily basis, it is unlikely that changes in cell composition, e.g., contribution of inflammatory cells from the submucosa, directly contributed to alterations in gene expression. Several cytokines, including IL1α and IL-33 were up-regulated in preterm vs. term. In addition, several genes that regulate the expression of cytokines and other immune genes were expressed at 3- (NFKB1α) to 6-fold (CASP1) higher levels in preterm vs. term infant exfoliated cells. Previous studies have shown that immortalized cells isolated from fetuses (H4 cells) or tissue explants from fetuses mount a more robust proinflammatory cytokine response (IL-8) after inflammatory stimulation with lipopolysaccharide or IL1β than cells from adult tissue (Caco-2) or explants from older children[24]. The excessive inflammatory response of the immature intestine is in part due to a developmental underexpression of IkB[25] as well as overexpression of the NFkB/MyD88 innate inflammatory genes (TLR2, TLR4, MyD88, TRAF-6, NFkB1 and IL-8) and under-expression of negative regulator genes (SIGIRR, IRAK-M, A-20 and TOLLIP) in fetal intestine relative to older children[8]. Thus, it appears that activation status of the intestinal innate immune response may contribute to excessive inflammation in the immature intestine in response to colonizing bacteria, which is a hallmark of NEC[8].

In exfoliated cells of term infants, up-regulated immune genes were associated with balancing the immune system, e.g. promoting T-cell development (LCP2; 3.6-fold greater in term than preterm), while inhibiting macrophage activation (LENG9; 16-fold greater in term than preterm). The majority of genes were involved in cell turnover, by regulating proliferation and apoptosis. One of the most highly differentially-expressed genes was an anti-apoptotic factor (MTRNR2L6), which was 5-fold higher in term than preterm. Another notable gene is SP3 (~2-fold higher in term than preterm), which is a transcription factor that can be regulated through short-chain fatty acid - acetylation, potentially supporting the role of these products of microbial metabolism in regulating normal gut growth in term infants[26].

In summary, we provide incontrovertible evidence that whole-genome sequencing of stool-derived RNA can be used to generate a global transcriptome gene expression signature in neonates. We have also compared for the first time, the intestinal global transcriptome in individual term and preterm and pooled term infants. Our findings provide insight into the global patterns of gene expression that vary in exfoliated epithelial cells of term and preterm infants. We anticipate that the described noninvasive RNA sequencing-based

Table 1 | Representative differentially expressed genes that were significantly higher in preterm infants versus term infants

Category and Gene ID	Term (average FPKM)	Preterm (average FPKM)	Preterm/ Term	p value (uncorrected)	Description
Lipid Metabolism					
ApoA4	483.85	18651.40	38.55	0.0001	Lipid metabolism, chylomicron and VLDL secretion
ApoA1	254.23	6906.41	27.17	0.0001	Lipid metabolism, reverse cholesterol transport
ASAH1	14.90	80.51	5.40	0.0064	Acid ceramidase, mediates cell growth arrest, differentiation and apoptosis
CERS2	92.13	371.37	4.03	0.0061	Ceramide synthase 2, regulates cell growth
MTM1	3.10	59.97	19.36	0.0037	Lipid phosphatase that negatively regulates EGFR degradation
PLIN2	127.72	359.13	2.81	0.0149	Controls the levels of lipid droplets, regulation of colonic cell growth
Immune					
CASP1	128.56	888.21	6.91	0.0030	Inflammasomes mediate activation of caspase 1 and promote secretion of IL-1beta and IL-18
IL1B	2343.32	6974.19	2.98	0.0926	Linked to intestinal disorders of inflammation in neonates
ICAM1	34.94	112.40	3.22	0.0280	Leukocyte trafficking across the endothelium
IL33	10.52	61.12	5.81	0.0144	Innate immune cell modulation, mediates Th2 immune response (parasite infections)
LCP1	14.06	72.14	5.13	0.0081	Associated with active Crohn's disease
NFKBIA	373.98	1164.21	3.11	0.0112	Regulates many epithelial and immune cell functions
PFKFB3	237.61	621.85	2.62	0.0271	Gene target of PPARgamma, regulates diet induced intestine inflammation response
S100-A9	71.14	1192.90	16.77	0.0043	Calcium and zinc binding protein that regulates inflammatory and immune responses; functions extracellularly as an antimicrobial
SOCS3	164.05	748.63	4.56	0.0017	Suppressor of cytokine signaling, generally anti-inflammatory that limits IL-6 induction of STAT3
TREM-1	15.29	96.78	6.33	0.0172	Expressed on myeloid cells (PMNs, monocytes/macrophages), increases with inflammatory diseases, linked to macrophage amplification of chronic inflammation
TYROBP	153.23	902.38	5.89	0.0067	An adaptor protein associated with multiple cell surface activating receptors expressed on both lymphoid and myeloid lineages
Cellular Function					
CEBPB	75.39	290.56	3.85	0.0098	Transcription factor regulating keratin synthesis
CFLAR	30.29	139.71	4.61	0.0451	Apoptosis regulatory protein, tissue homeostasis
FAM65B	9.00	169.30	18.81	0.0007	Promotes myogenic differentiation and cytoskeletal rearrangement
FUBP1	31.78	110.84	3.49	0.0128	DNA binding protein that promotes c-myc expression
G0S2	2927.30	60194.00	20.56	0.0003	Promotes apoptosis, prevents bcl2-bax heterodimers
KDM5B	32.12	98.30	3.06	0.0281	Histone demethylase, favors cell proliferation
MMP1	53.25	159.15	2.99	0.0392	Matrix metalloproteinase, potential regulator of intestinal homeostasis
NPM1	59.07	165.49	2.80	0.0484	Nucleolar acidic protein, plays a positive role in cell proliferation and growth
SERPINA1	136.74	534.73	3.91	0.0012	Serine protease inhibitor, proteolytic activity towards insulin, protects lower respiratory tract

approach will enable elucidation of how the bedside clinical management of an extremely preterm infant population influences intestinal gene expression. With this understanding, dietary and medical practices can be evaluated that optimally promote intestinal development and, ideally, identify those clinical practices that approximate as closely as possible the development of the healthy, term breast fed infant. The possible uses of non-invasive high-throughput RNA sequencing data are vast and include early detection (screening), monitoring disease progression, risk assessment, and diet-dependent interaction between gut microbiota and host epithelium. We propose that stool samples containing exfoliated cells have the potential for generating comprehensive, diagnostic gene sets for the noninvasive identification/prediction of different intestinal phenotypes in infants.

Methods

Subject recruitment. The experimental human protocol for term infants was approved by the University of Illinois and Texas A&M Institutional Review Boards and for preterm infants by the Beth Israel Deaconess Medical Center and Texas A&M Institutional Review Boards. Informed consent was obtained from parents prior to participation in the study, and all experiments were performed in accordance with relevant guidelines and regulations. Details of the study admission criteria and protocols for term infants have been previously described[11]. Briefly, healthy, term exclusively BF or FF infants were eligible for enrollment into the study. All infants were considered term and gestational age was similar for BF (39.7 ± .08 weeks; range: 38 5/7 to 41 3/7 weeks) and FF (39.7 ± 0.4 weeks; range: 38 5/7 to 40 4/7 weeks) infants. For each term infant, a stool sample was collected at three months-of-age. In addition, stool samples were collected from six extremely preterm (24–30 weeks gestation) infants admitted to the Beth Israel Deaconess Medical Center neonatal intensive care unit (NICU). The six preterm infants had a gestational age ranging from 24 1/7 weeks of gestation to 30 0/7 weeks. Four infants under 30 weeks gestational age had a stool sample collected at 4–5 weeks of life, while two preterm infants at 30 weeks gestational age had a stool sampled at 2 weeks of life. All six preterm infants were on full enteral feedings (1 formula; 2 breast milk; 3 mixed formula and breast milk). For full metadata for term and preterm infants, please consult Supplementary Tables 3 and 4.

Sample preparation and sequencing. PolyA$^+$ RNA was isolated from stool samples from term and preterm infants as previously described[27]. ERCC was diluted 1 : 10 and 0.5 mL was added to each sample. Pooled polyA$^+$ samples from 18 term infants or 6 preterm infants were processed with the NuGEN Ovation 3′-DGE kit (San Carlos, CA) to convert RNA into cDNA followed by NuGEN Encore NGS Library System I kit to create Illumina libraries, as per manufacturer's instructions. Sequencing on Illumina GAIIx and HiSeq 2000 platforms (San Diego, CA) were carried out using standard Illumina protocols on the Texas A&M University campus. Briefly, 70 ng of each sample were used to synthesize first and second strand cDNA, which was purified using Agencourt RNAClean XP beads (Brea, CA) included in the kit. The cDNA was linearly amplified using the NuGEN SPIA primer, and cDNA quality and quantity were determined using an Agilent 2100 Bioanalyzer and Nanodrop spectrophotometer. Three micrograms of cDNA was fragmented using a Covaris S2 sonicator with the following settings: duty cycle 10%, intensity 5, cycles/burst 100, time 5 min. Fragmented samples were concentrated using the QIAquick PCR purification kit (Qiagen, Venlo, Netherlands) as per manufacturer's instructions. Samples were quantified using the Quant-iT kit (Invitrogen, Carlsbad, CA) and evaluated for proper fragmentation of 150 to 200 bp on an Agilent Bioanalyzer DNA 1000 chip. Following cDNA fragment repair and purification, Illumina adaptors were ligated onto fragment ends followed by amplification to produce the final library. Libraries were quantified using Quant-iT (Invitrogen) and run on an Agilent DNA

Table 2 | Representative differentially expressed genes that were significantly higher in term versus preterm infants

Category and Gene ID	Term (average FPKM)	Preterm (average FPKM)	Term/ Preterm	p value (uncorrected)	Description
Immune					
CD177	197.49	19.186	10.29	0.01215	A glycosylphosphatidylinositol (GPI)-anchored membrane protein is a potential receptor for PR3, the preferred target of antineutrophil cytoplasmic antibodies (ANCAs) in Wegener's granulomatosis. Involved in neutrophil transmigration.
IFI27	4643.67	1503.71	3.09	0.0256	Mediates IFN-induced apoptosis
IRF9	293.777	45.0231	6.53	0.0048	Transcription regulatory factor that mediates type I interferons
LENG9	1209.33	73.0508	16.55	0.00145	Negative regulators of macrophage activation
LCP2	86.2228	24.4284	3.53	0.0378	SLP76, a substrate for T cell ZAP70, promotes T cell development
Cellular Function					
ABCC5	122.498	20.6639	5.93	0.00545	ATP binding cassette transporter
ATP5B	162.509	41.1475	3.95	0.01605	ATP synthase
BNIP2	44.0329	7.76199	5.67	0.0191	bcl-2 binding protein, responsive to estrogen; regulates cell dynamics by interacting with cdc42
CDKN2B	243.227	86.5265	2.81	0.02475	Cyclin dependent kinase inhibitor, controls cell cycle G1 progression
DYNLL1	528.599	129.298	4.09	0.01345	Intracellular transport and motility
ESRRA	435.007	83.0017	5.24	0.01245	Estrogen receptor related alpha, cell growth and maintenance
INSR	49.2393	14.5811	3.38	0.00755	Insulin receptor, regulates cell growth
KREMEN2	63.2595	0	NA	0.0284	Receptor for Dickkopf protein, cooperates with Dickkopf to block Wnt signaling.
MLLT4	53.2364	14.4052	3.70	0.0059	Organization of cell junctions, belongs to the cell adhesion system
MTRNR2L6	340.597	36.8941	9.23	0.00025	Antiapoptotic factor
PDPK1	54.8395	9.62117	5.70	0.0045	Master lipid kinase regulating PI-3 kinase pathway/Akt, apical endosome trafficking
RAP2A	81.8369	29.7633	2.75	0.0357	Belongs to the ras oncogene family, regulates cytoskeletal rearrangements
SCNN1A	132.419	16.8286	7.87	0.0077	Sodium mediated non voltage ion channel. Mediates diffusion of luminal sodium and water through the apical membrane
SP3	56.516	21.7481	2.60	0.0271	A transcription factor that can be regulated by acetylation, e.g., SCFA, can repress insulin like growth factor action.
TRIM36	50.8464	5.88888	8.63	0.0004	Mediates ubiquitination and proteosomal degradation; chromosome segregation and cell cycle regulation

Chip 1000 to confirm appropriate sizing and the exclusion of adapter dimers. Approximately 32 million, 36 bp reads were sequenced in each lane for the pooled samples on an Illumina GAIIx [SRA:SRR626229] and approximately 300 million, 50 bp single end reads on an Illumina HiSeq for the six individual samples [SRA:PRJNA182262]. Pooled and individual sample collection and processing is depicted visually in Supplementary Figure 4. Raw, de-multiplexed FASTQ files were examined using FastQC and determined to be of sufficient read quality, and sequence, nucleotide, and k-mer diversity. The number of ERCC molecules present in the samples was calculated using known concentrations of ERCC transcripts as published by Ambion and the quantity of RNA solution that was used for library preparation.

Next generation sequencing read alignment. A wide array of reference genomes were assembled to identify the rough composition of detected RNA reads (see Supplementary Methods). The UCSC hg19 human reference genome was used with annotation from the Illumina iGenomes resource[28]. Microbial, mitochondrial, fungal, viral, and protozoan nucleotide databases were obtained from RefSeq version 59 which resulted in 18.5 Gb, 100 Mb, 2.3 Gb, and 1.7 Gb genomes, respectively. 580 Mb of ribosomal sequences were obtained from SILVA release 111[29], using both small and large subunit repositories. BWA[30] was used to align reads against the bacterial reference genome due to its size. SNAP[31] was used to align to mitochondrial, ribosomal, and viral references, and STAR[32] was used for protozoan, fungal, human, and ERCC reference genomes due to the possible presence of splice junctions. All alignment was performed using default tolerance parameters and considered acceptable based on the agreement with known and observed ERCC controls. We also compared our alignments against the slower TopHat aligner[33] and obtained similar results. Read mapping results were visualized in Integrative Genomics Viewer, IGV[34]. Additional details regarding reference genomes and automated source code for analysis are available in the Supplementary Materials.

Gene expression quantification. As a genome-guided assembly method, Cufflinks was used to estimate gene expression for the human reference and HTSeq was used for all other references[35,36]. Due to the 3′ selection based enrichment and subsequent observed 3′ bias in read alignment, normalized read counts for expression correlation results were also used. Read counts were obtained using HTSeq and normalized by the mean expression of 3804 "house-keeping" mRNA species as determined by[37]. Correlation of ERCC measurements between samples and known concentrations were performed following log transformation and addition of 0.001 to the counts to avoid unbounded values. This was utilized for the Spearman correlation coefficient because of the logarithmic nature of the concentrations provided in ERCC spike-in kits. Differential expression testing was performed using Cuffdiff as available from the Cufflinks package. Genes with P-values < 0.05 were input into the Ingenuity IPA analysis tools (Ingenuity Systems Inc., Redwood City, CA) to assess pathway, biological function, and upstream activity[14] and are visualized in a volcano plot in Supplementary Figure 5 as produced by the cummeRbund package.

Real time PCR was used for gene expression confirmation. For this purpose, cDNA was prepared from fecal RNA from three term and three preterm infants using SuperScript II (LifeTechnologies, Carlsbad, CA) followed by PCR on an ABI 7900HT Real Time PCR System. TaqMan assays were purchased for ABCC5, APOA4, CASP1, DYNLL1, NFKBIA, PLIN2, PPAP2A, RPS16, SCNN1A, SLC2A1, and TMSB4X (LifeTechnologies). These genes were selected from a larger set of differentially expressed genes in term vs. preterm infants as identified in the IPA analysis.

Ingenuity pathway analyses. "Functional enrichment" analysis was performed using Ingenuity Pathway Analysis (IPA) version 2.0 software. To perform IPA analysis, all differentially expressed genes ($P < 0.05$) in the preterm or term infants were uploaded into three columns for the purpose of generating Illumina probe ID, t-value (fold change) and P-value data. P-values were uncorrected for multiple testing owing to the number of human fecal samples. By default, during IPA analysis, only molecules from the data set associated with the Ingenuity Knowledge Base repository (Ingenuity Systems Inc.) were considered. Functional Analysis identified the biological functions and/or diseases that were most significant to the data set. The significance of the association between the data set and the specific pathways of interest was determined in three ways: (a) as a ratio of the number of molecules from the data set that mapped to the pathway divided by the total number of molecules that mapped to the Ingenuity Knowledge Base pathway, (b) Fisher's exact test was used to calculate a *P* value determining the probability that the association between the genes in the data set and the pathway of interest could be explained by chance alone, and (c) activation state ("increased" or "decreased") was inferred by the activation z-score. The derivations of the z-scores are based on relationships in the molecular network that represent experimentally observed causal associations between genes and those functions.

"Canonical pathway" analysis was used to identify networks from the IPA library that were most significantly modulated across subject groups. Significance of the association between each data set and the canonical pathway was measured in 2 ways: (1) as a ratio of the number of molecules from the data set that mapped to the pathway

divided by the total number of molecules that mapped to the canonical pathway, and (2) Fisher's exact test was used to calculate p-values determining the probability that the association between genes in the dataset and each canonical pathway was explained by chance alone.

1. Commare, C. E. & Tappenden, K. A. Development of the infant intestine: implications for nutrition support. *Nutr Clin Pract* **22**, 159–173 (2007).
2. Tremblay, E. *et al.* Gene-expression profile analysis in the mid-gestation human intestine discloses greater functional immaturity of the colon as compared with the ileum. *J Pediatr Gastroenterol Nutr* **52**, 670–678, doi:10.1097/MPG.0b013e3182078370 (2011).
3. Neu, J. Gastrointestinal development and meeting the nutritional needs of premature infants. *Am J Clin Nutr* **85**, 629S–634S (2007).
4. Chowdhury, S. R. *et al.* Transcriptome profiling of the small intestinal epithelium in germfree versus conventional piglets. *BMC genomics* **8**, 215; doi:10.1186/1471-2164-8-215 (2007).
5. Van den Abbeele, P., Van de Wiele, T., Verstraete, W. & Possemiers, S. The host selects mucosal and luminal associations of coevolved gut microorganisms: a novel concept. *FEMS Microbiol Rev* **35**, 681–704, doi:10.1111/j.1574-6976.2011.00270.x (2011).
6. Claud, E. C. & Walker, W. A. Bacterial colonization, probiotics, and necrotizing enterocolitis. *J Clin Gastroenterol* **42 Suppl 2**, S46–52, doi:10.1097/MCG.0b013e31815a57a8 (2008).
7. Diehl-Jones, W. L. & Askin, D. F. Nutritional modulation of neonatal outcomes. *AACN Clin Issues* **15**, 83–96 (2004).
8. Nanthakumar, N. *et al.* The mechanism of excessive intestinal inflammation in necrotizing enterocolitis: an immature innate immune response. *PLoS one* **6**, e17776; doi:10.1371/journal.pone.0017776 (2011).
9. Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L. & Gordon, J. I. Human nutrition, the gut microbiome and the immune system. *Nature* **474**, 327–336, doi:10.1038/nature10213 (2011).
10. Donovan, S. M. *et al.* Host-microbe interactions in the neonatal intestine: role of human milk oligosaccharides. *Adv Nutr* **3**, 450S–455S, doi:10.3945/an.112.001859 (2012).
11. Chapkin, R. S. *et al.* Noninvasive stool-based detection of infant gastrointestinal development using gene expression profiles from exfoliated epithelial cells. *Am J Physiol Gastrointest Liver Physiol* **298**, G582–589, doi:10.1152/ajpgi.00004.2010 (2010).
12. Schwartz, S. *et al.* A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response. *Genome Biol* **13**, r32; doi:10.1186/gb-2012-13-4-r32 (2012).
13. Arboleya, S. *et al.* Establishment and development of intestinal microbiota in preterm neonates. *FEMS Microbiol Ecol* **79**, 763–772, doi:10.1111/j.1574-6941.2011.01261.x (2012).
14. Triff, K. *et al.* Genome-wide analysis of the rat colon reveals proximal-distal differences in histone modifications and proto-oncogene expression. *Physiol Genomics* **45**, 1229–1243 (2013).
15. Davidson, L. A. *et al.* Identification of actively translated mRNA transcripts in a rat model of early-stage colon carcinogenesis. *Cancer Prev Res (Phila)* **2**, 984–994, doi:10.1158/1940-6207.CAPR-09-0144 (2009).
16. Marioni, J. C. *et al.* "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." *Genome Res* **18**, 1509–1517 (2008).
17. Davidson, L. A., Jiang, Y. H., Lupton, J. R. & Chapkin, R. S. Noninvasive detection of putative biomarkers for colon cancer using fecal messenger RNA. *Cancer Epidemiol Biomarkers Prev* **4**, 643–647 (1995).
18. Donovan, S. M. Role of human milk components in gastrointestinal development: Current knowledge and future needs. *J Pediatr* **149**, S49–S61, doi:10.1016/j.jpeds.2006.06.052 (2006).
19. Zhao, C. *et al.* Noninvasive detection of candidate molecular biomarkers in subjects with a history of insulin resistance and colorectal adenomas. *Cancer Prev Res (Phila)* **2**, 590–597, doi:10.1158/1940-6207.CAPR-08-0233 (2009).
20. Hatzis, C. *et al.* Effects of tissue handling on RNA integrity and microarray measurements from resected breast cancers. *J Natl Cancer Inst* **103**, 1871–1883, doi:10.1093/jnci/djr438 (2011).
21. Brosseau, J. P. *et al.* High-throughput quantification of splicing isoforms. *RNA* **16**, 442–449, doi:10.1261/rna.1877010 (2010).
22. Wood, H. M. *et al.* Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. *Nucleic Acids Res* **38**, e151; doi:10.1093/nar/gkq510 (2010).
23. Kerick, M. *et al.* Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC Med Genomics* **4**, 68; doi:10.1186/1755-8794-4-68 (2011).
24. Nanthakumar, N. N., Fusunyan, R. D., Sanderson, I. & Walker, W. A. Inflammation in the developing human intestine: A possible pathophysiologic contribution to necrotizing enterocolitis. *Proc Natl Acad Sci U S A* **97**, 6043–6048 (2000).
25. Claud, E. C. *et al.* Developmentally regulated IkappaB expression in intestinal epithelium and susceptibility to flagellin-induced inflammation. *Proc Natl Acad Sci U S A* **101**, 7404–7408, doi:10.1073/pnas.0401710101 (2004).
26. Zeissig, S. *et al.* Butyrate induces intestinal sodium absorption via Sp3-mediated transcriptional up-regulation of epithelial sodium channels. *Gastroenterology* **132**, 236–248, doi:10.1053/j.gastro.2006.10.033 (2007).
27. Davidson, L. A., Lupton, J. R., Miskovsky, E., Fields, A. P. & Chapkin, R. S. Quantification of human intestinal gene expression profiles using exfoliated colonocytes: a pilot study. *Biomarkers* **8**, 51–61, doi:10.1080/1354750021000042268 (2003).
28. Illumina Inc., iGenomes. URL http://support.illumina.com/sequencing/sequencing_software/igenome.ilmn(Date of access:01/22/2014).
29. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590–596, doi:10.1093/nar/gks1219 (2013).
30. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, doi:10.1093/bioinformatics/btp324 (2009).
31. Zaharia, M. *et al.* Faster and more accurate sequence alignment with SNAP. *arXiv preprint arXiv:1111.5572* (2011).
32. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21, doi:10.1093/bioinformatics/bts635 (2013).
33. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111, doi:10.1093/bioinformatics/btp120 (2009).
34. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24–26, doi:10.1038/nbt.1754 (2011).
35. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511–515, doi:10.1038/nbt.1621 (2010).
36. Anders, S. HTSeq: Analysing high-throughput sequencing data with Python. (2010)*URL http://www-huber.embl.de/users/anders/HTSeq/doc/overview.htm*(Date of access:02/19/2014).
37. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet* **29**, 569–574 (2013).

Acknowledgments

This work was supported by Texas A&M AgriLife Research, National Institute of Health grants CA129444, U01CA162077, HD61929, R25TCA090301, P30ES023512, Hatch project ILLU-971-346 through the Division of Nutritional Sciences Vision 20/20 program and USDA–NIFA Grant Designing Foods for Health 2010-34402-20875.

Author contributions

J.M.K. performed bioinformatic and data analysis and assisted with manuscript writing, L.A.D. performed qPCR, contributed to RNA-Seq and writing of the manuscript, D.H. performed bioinformatic analysis, C.R.M. and S.M.D. obtained the fecal samples, and conceived the design with R.S.C. and I.V.I., J.S.G. performed RNA-Seq, and I.V.I. and R.S.C. supervised the work and assisted with manuscript writing. All authors discussed results and commented on the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Knight, J.M. *et al.* Non-invasive analysis of intestinal development in preterm and term infants using RNA-Sequencing. *Sci. Rep.* **4**, 5453; DOI:10.1038/srep05453 (2014).