Premature neonatal gut microbial community patterns supporting an epithelial TLR-mediated pathway for necrotizing enterocolitis

Alexander G Shaw (a.shaw@imperial.ac.uk)
Imperial College London https://orcid.org/0000-0002-3166-872X
Kathleen Sim
Imperial College London
Graham Rose
Imperial College London
David J. Wooldridge
Public Health England
Ming-Shi Li
Imperial College London
Raju V. Misra
Natural History Museum
Saheer Gharbia
Public Health England
J. Simon Kroll
Imperial College London

Research article

Keywords: Metagenome, Necrotising enterocolitis, premature infant, microbiome, TLR4, TLR9

Posted Date: March 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-289345/v1

Version of Record: A version of this preprint was published at BMC Microbiology on August 6th, 2021.
See the published version at https://doi.org/10.1186/s12866-021-02285-0.

Abstract

Background

Necrotising enterocolitis (NEC) is a devastating bowel disease, primarily affecting premature infants, with a poorly understood aetiology. Prior studies have found associations in different cases with an overabundance of particular elements of the faecal microbiota (in particular Enterobacteriaceae or *Clostridium perfringens*), but there has been no explanation for the different results found in different cohorts. Immunological studies have indicated that stimulation of the TLR4 receptor is involved in development of NEC, with TLR4 signalling being antagonised by the activated TLR9 receptor. We speculated that differential stimulation of these two components of the signalling pathway by different microbiota might explain the dichotomous findings of microbiota-centered NEC studies. Here we used shotgun metagenomic sequencing and qPCR to characterise the faecal microbiota community of infants prior to NEC onset and in a set of matched controls. Bayesian regression was used to segregate cases from control samples using both microbial and clinical data.

Results

We found that the infants suffering from NEC fell into two groups based on their microbiota; one with low levels of CpG DNA in bacterial genomes and the other with high abundances of organisms expressing LPS. The identification of these characteristic communities was reproduced using an external metagenomic validation dataset. We suggest that these two patterns represent the stimulation of a common pathway at extremes; the LPS leading to overstimulation of TLR4, whilst low levels of CpG fail to sufficiently stimulate TLR9.

Conclusions

The identified microbial community patterns support the concept of NEC resulting from TLR-mediated pathways. Identification of these signals suggests characteristics of the gastrointestinal microbial community to be avoided to prevent NEC. Potential pre- or pro-biotic treatments may be designed to optimise TLR signalling.

Background

Necrotising enterocolitis

Necrotising enterocolitis (NEC) is a devastating bowel disease primarily affecting premature infants, the aetiology of which is poorly understood. The increasing incidence of the disease which has come with the advent of improved neonatal intensive care, and our ability to keep more and more premature infants

alive, combined with the high mortality rate, limited treatment options, rapid onset, and lack of a screening tool, has made research into the disease a priority.

Bacteria are considered to play a key role in the pathogenesis of NEC – the condition does not occur in sterile settings (e.g. *in utero* or in germ-free animal models). Antibiotics are used to treat the disease, but paradoxically, antibiotic treatment early in life as empirical therapy for sepsis been shown to increase the risk of developing the disease [1].

We have previously described our findings of two different microbial signatures anticipating the development of NEC in a cohort of premature infants [2]. Using 16S rRNA characterisation of faecal samples collected from the infants from birth until disease onset, we showed that there was either a 'bloom' of *C. perfringens* immediately prior to diagnosis, or, unusually high levels of Enterobacteriaceae from birth which persisted until NEC developed. Other investigators have reported similar findings [3–5], posing a question as to how such very different organisms can provoke the same comparable pathogenic process.

The role of TLR4 and 9 in necrotising enterocolitis

In a mouse model of NEC, the pattern recognition receptor TLR4, a key component of the innate immune system, has been found to play a central role, its deletion being protective [6]. TLR4 expression in the fetal/neonatal murine gut is temporally regulated, levels in the gut epithelium and endothelium rising as the foetus matures and then rapidly falling in the perinatal period [7]. On premature (over)exposure of neonatal mouse epithelial TLR4 to its ligand (lipopolysaccharide (LPS): a major component of the outer membrane of Gram-negative bacteria), binding leads to activation of the NF-KB pathway and destruction of the gut epithelium that is the hallmark of NEC [7].

A similar pattern of TLR4 expression has been reported in human infants, with a peak at 16 weeks gestation and greatly increased expression at 21 weeks gestation compared to four months after birth [8]. An infant born prematurely could therefore be expected to have a comparatively high expression of TLR4 in the gut epithelium, and premature exposure to a microbiota by mischance dominated by Gram-negative organisms, leading to widespread activation of TLR4, to carry the risk of catastrophic epithelial damage.

Of course, not all infants colonised with high levels of Gram-negative organisms develop NEC. TLR4 signalling is down-regulated by TLR9, for which the activating ligand is the CpG motif in DNA [9]. CpG motifs are more prevalent in bacterial than in eukaryotic DNA, although the frequency/kb varies greatly across the bacterial kingdom [10]. Activation of TLR9 acts to inhibit TLR4 mediated signalling via IL-1R-asscoaited kinase M [7]. TLR9 therefore acts to prevent a constant inflammatory response to the commensal bacteria that are resident in the lumen of the gut [11].

We hypothesised that an imbalance in the stimulation of TLR4 and TLR9 in the faecal gut, potentially leading to the development of NEC, could be quantified through calculation of the abundance of Gram-negative bacteria and bacterial DNA CpG content which would serve as proxies of immuno-stimulation.

To address this, we undertook both a metagenomic analysis and a bacterial quantification by qPCR of the microbial content of faecal samples closest to the time of diagnosis in infants prior to the development of NEC, and in a set of paired controls. From this we calculated the number of CpG motifs present per gram of faecal matter. Quantitative PCR data and taxonomic identifications were also used to estimate the number of Gram-negative bacteria present. We report here that faecal samples collected from infants taken just before NEC development have either a lower CpG motif frequency in bacterial DNA or a higher amount of Gram-negative bacteria per gram of faeces compared to samples from control infants matched for postnatal age. These findings were validated using an external dataset which also included both a metagenomic analysis and bacterial quantification for faecal samples collected from premature infants.

Results

Sequencing results

Twenty-six samples entered the shotgun metagenomic sequencing pipeline. These comprised 12 NEC samples (N1-N12) collected as close to diagnosis as possible, 12 matched control samples (C1-C12), a DNA extraction negative control and a technical repeat of sample C12. The negative control and two samples, N7 and C9, failed library amplification (requiring more than eight PCR cycles). Canonical correlation analysis (CCA) was used to compare the similarity of samples by microbial content, confirming that the technical repeat clustered tightly to its pair (see Fig. 1), indicating the robustness of the sequencing pipeline.

Microbial communities of sequenced samples

After removal of failed samples, controls and repeats, eleven NEC samples and eleven control samples remained, of which twenty were matched pairs. Samples were clustered according to the similarity of their microbial communities (see Fig. 2), displaying some loose grouping between sample types. NEC samples tended towards domination by either Enterobacteriaceae or *Enterococcus*, *Staphylococcus*, *Anaerococcus* or *Clostridium* species. Some control samples fell into these categories, whilst others featured high abundances of Bifidobactericeae.

Quantification of the Bacterial communities

qPCR data for each faecal sample was corrected for the 16S rRNA gene copy number of its bacterial community. This provided an estimate of the number of bacteria per gram of faeces (see Additional file 1). No significant difference in bacterial abundance was found between control and NEC samples.

TLR9 stimulatory potential of typical premature infant gut colonisers

We sought to establish the stimulatory properties of the bacteria present in the guts of premature infants with regard to TLR9 through calculation of the frequency of CpG DNA per megabase of their sequenced DNA. Taxonomic groups relating to *Bifidobacterium* had the highest frequency of CpG motifs, whilst *Clostridium*, *Anaerococcus* and *Staphylococcus* taxonomic groups had a CpG motif frequency between 4 % and 25 % of this (see Fig. 3). These results are consistent with previous findings [10].

Stratification of samples by immunostimulatory proxies

Following prior studies in neonatal mice,[6, 7] we hypothesised that infants at high risk of NEC would either display over-stimulation of TL4 or under stimulation of TLR9. To investigate whether the segregated into the two categories we first sought to identify the infants most likely to fall into either category.

To quantify the overall immunostimulatory potential of the faecal microbiota for TLR 9, we estimated the abundance of CpG DNA per gram of faeces and the total number of bacteria per gram of faeces. The expected relationship between the two would be linear; an increase in the amount of bacteria would lead to a proportional increase in CpG motifs. The majority of samples indeed followed this relationship (see Fig. 4A). However, four NEC cases were found to have a particularly low ratio of CpG motifs to bacteria.

We then sought to identify the NEC samples most likely to feature high immunostimulation of TLR 4 due to an overabundance of its ligand, LPS. As a proxy for the abundance of LPS, the number of Gram-negative bacteria per gram of faeces was calculated by classifying the previously quantified bacterial taxonomic groups according to Gram staining (see Fig. 4B). Four NEC cases displayed an over-abundance of Gram-negative bacteria compared to the control samples.

The remaining three NEC cases were not assigned an initial classification.

Stratification of samples by immunostimulatory proxies and clinical factors

Two Bayesian linear regression models were used to segregate the two groups of classified NEC cases from the control infants, with each model focusing on one of the immunostimulatory proxies. In anticipation that the abundance of these proxies and the bacteria per gram of faeces would not remain static over time, day of life and gestation at birth were included as additional factors along with days of antibiotics prior to sampling. Reclassification of the NEC cases was permitted in response to the inclusion of these additional factors, with the overall aim of achieving the most accurate segregation of cases and controls.

The first model, aiming to identify the risk of NEC associated with low stimulation of TLR9 (“CpG-associated NEC”), sought to initially separate the cases highlighted in Fig. 4A) from control infants. Predicted probabilities of NEC risk were calculated for each NEC and control sample. Seven cases had predicted probabilities of NEC greater than the 90 % quantile of control infant risk probabilities (see

Fig. 5). Increased risk was associated with lower CpG motifs per bacteria within the sample, a lower total abundance of bacteria, fewer days with antibiotics, increased days of life and decreased gestation at birth.

The second model, aiming to identify the risk of NEC associated with over-stimulation of TLR4 ("Gram-negative-associated NEC") sought to initially segregate the cases highlighted in Fig. 4B) from control infants. Five cases had predicted probabilities of NEC greater than the 90 % quantile of control infant risk probabilities (see Fig. 5). Increased risk for Gram-negative-associated NEC was associated with increased proportion of the bacterial community being Gram-negative, increased overall bacterial abundance, increased days of antibiotics, increased day of life and increased gestation at birth. Gram-negative bacteria comprised 93 % of this cohort's total bacteria, compared to only 51 % in the control infants. The coefficients of the two models are shown in additional file 4. NEC severity was not associated with predicted probabilities.

The two categories of cases were clearly separated by antibiotics use; of the three cases predicted high risk exclusively by the Gram-negative-associated NEC model all had antibiotics courses prior to NEC (a mean of nine days). The six cases with predicted high risk exclusively of CpG-associated NEC had received a mean of only one day of antibiotics, with three cases having received none.

To test whether these immunostimulatory signatures were reproducible, an external metagenomic dataset with quantification by digital droplet PCR was used for validation [12]. Samples from five control infants and three infants that developed NEC were classified using the models. One of the cases developed NEC twice, and each incident was included. The cases each had a sample within a five-day window prior to NEC diagnosis, and a total of nine earlier samples. 23 samples were tested from the five control infants. Sequencing data and quantification of CpG and bacteria were processed using the same pipeline as our own dataset. These data were entered into the two models with the coefficients derived from the testing on our own cohort to give predictive probabilities (See Fig. 5).

The samples closest to NEC diagnosis from one case and the sample prior to the first incidence of NEC from the recurrent case had high predicted probabilities of CpG-associated NEC along with two samples taken two and six days prior to the sample closest to NEC. The remaining NEC case and the second incidence of the recurrent case of NEC (following a ten-day treatment with antibiotics) had a high predicted probability of Gram-negative associated NEC, along with two samples taken one and two days prior. Seven validation control samples had high risk of Gram-negative associated NEC and all were within the first two weeks of life, whilst samples from NEC cases attributed high risk were all from week three to week eight of life.

Discussion

In this study we have used characterisation of the infant faecal microbiota by shotgun metagenomics and qPCR to demonstrate how infants who go on to develop NEC display characteristic communities of faecal microbiota prior to diagnosis. These communities feature high proportions of LPS-expressing

bacteria and/or a low frequency of CpG motifs within the bacterial DNA. These findings have been reproduced in an external cohort and fall in line with a recent theory concerning the development of NEC, with high levels of LPS stimulating the TLR4 receptor leading to inflammation, whilst low CpG frequencies lead to reduced TLR9 signalling and reduced IRAK-M dependent inhibition of TLR4 [7].

The mostly dichotomous nature of our findings suggests why prior microbial studies have confusingly implicated a range of organisms, with some studies highlighting association with an excess of Enterobacteriaceae [2, 3, 5, 13], whilst others with a range of organisms including *Clostridium* and *Staphylococcus* species [2, 4, 13]. We suggest that where NEC is associated with an Enterobacteriaceae-dominated microbiota the pathological basis for the epithelial necrosis is overstimulation of TLR4 by over-abundant LPS, while where the association is with the bloom of a pathogen like *C. perfringens*, it is the (strikingly) low CpG frequency in the resulting microbiota that leads to a failure of counter-regulation of TLR4 through TLR9.

Both of our models include terms representing temporal as well as bacterial associations, as the occurrence of the microbial patterns must be considered in relation to clinical time-courses. In each model, increase in risk is associated with increased day of life, in line with the recognised paucity of NEC cases prior to 8–10 days post-partum [14]. For our CpG-associated NEC model this association may represent the requirement that the gut becomes anaerobic prior to the flourishing of *Clostridium* species - the organisms with the lowest CpG motifs observed within our cohort. For the Gram-negative-associated NEC model increased risk associated with increased day of life may reflect the uncharacteristic persistence of Gram-negative bacteria within infant's gut microbiota. A shift away from this pattern is typically seen from week three of life as the gut becomes more anaerobic and obligate anaerobes begin to dominate [15]. Both in our study cohort and in the validation dataset there are a small number of early life control and pre-NEC samples - predominantly taken in the first two weeks of life - that have high levels of Gram-negative bacteria. Control samples collected after this time period tend not display this, with communities following the typical pattern of succession. However, in infants that develop NEC, the dominance of Gram-negative bacteria persists. Reduced gestation is also associated with increased risk of Gram-negative associated NEC, and this may relate to expression of TLR4 in the infant gut with gestational age, which is observed in mice to peak prior to birth [7]. These two factors combined could lead to high risk infants experiencing the confluence of peaking TLR4 expression and prolonged exposure to high abundances of LPS.

LPS from different organisms causes varying degrees of TLR4 stimulation, with *Veillonella parvula* at one extreme causing minimal stimulation. This may be a characteristic of the Negativicutes class as a whole, given their evolutionary distance from other Gram-negative organisms, and may explain why Negativicutes have been negatively associated with NEC [5]. Further characterisation of the immune-stimulatory properties of individual Gram-negative bacterial species will be important to fully parametrise our Gram-negative-associated NEC model.

Our CpG-associated-NEC model identified an association between fewer days of antibiotic treatment and the development of NEC. While previous studies have shown that increased antibiotic usage is linked with NEC [1], in the case of these specific infants in our cohort, the reduced antibiotic duration may have facilitated the succession of the gut microbiota towards organisms such as Clostridia. In term and pre-term infants, antibiotic treatment has been demonstrated to lead to a higher proportion of Proteobacteria in the post-treatment gut microbiota [16, 17], as seen by the positive association between days of antibiotics and increased Gram-negative associated NEC risk. Interestingly, the validation case with two incidents of NEC was initially found at high risk of CpG-related NEC and then progressed to high risk of Gram-negative-associated NEC after an additional ten days of antibiotic treatment. The proportion of the bacterial population that was Gram-negative was only 24 % prior to the first incidence of NEC and had risen to 74 % before the second incident.

Whilst our two models classify all eleven NEC cases in our cohort and the three validation cases as being at high risk, the credible intervals for the coefficients that were determined are wide and require further data to provide greater certainty. Both day of life and gestation were retained in the models given the highly time dependent nature of the remaining factors. The resulting coefficients were relatively small however, possibly due to the correlation with the other factors, for example the negative association between Gram negative bacteria in the faecal microbiota and day of life. Additional data points for the first two weeks of life or longitudinal data could better inform this relationship to avoid high risk classifications for healthy infants who lack long-term faecal colonisation by Gram-negative bacteria. We also acknowledge that whilst faeces is a convenient sampling methods, it may not contain a true representation of the gut microbiota at the site of the small intestine where NEC commonly occurs [18].

Although the predicted risk scores for the two models are plotted against each other in Fig. 5, the mathematical interaction between these two pathways is unknown. One control sample has relatively high risk by both models yet did not develop NEC. Mathematical characterisation of the interactions between the two pathways would be essential to combine the two models into a single unified predictor of NEC risk, and would likely also require the interplay of other immunological meditators that could potentially be involved such as those discussed by Cho *et al* [19].

Conclusions

The microbial community patterns presented here support the possibility of NEC occurring due to TLR-mediated pathways. Our findings suggest that clinical management of very premature infants to avoid NEC may benefit from careful steering of the faecal microbiota community to avoid the development of either of the two potentially destructive patterns. The prophylactic use of antibiotics in the NICU has been observed to lead to increased community domination by Enterobacteriaceae [16], leading to microbiota conforming to our high Gram-negative pattern. Whilst a suitably anaerobic gut environment may harbour low-CpG organisms such as *Clostridium perfringens*, characteristic of our second destructive pattern, appropriate nutrition, with consideration of the use of pre- or probiotics, may promote community development towards a more beneficial pattern, dominated by bacteria such as *Bifidobacterium*.

Interventions aimed at rehabilitating the gut microbiota should be designed so as to avoid avoiding both negative community patterns. We believe our findings provide an explanation for the different, yet consistent, microbiota associated with NEC that have been reported by researchers around the world over the course of decades, improving understanding of the interplay between the host and the resident gastrointestinal microbiota which may indicate new strategies for the prevention of NEC.

Methods

Study population

Faecal samples analysed were from infants enrolled on the “Defining the Intestinal Microbiota in Premature Infants” study between January 2011 and December 2012 at an Imperial College Healthcare National Health Service Trust neonatal intensive care unit (NICU) (St Mary's Hospital, Queen Charlotte's and Chelsea Hospital). Recruited infants were born before 32 completed weeks of gestation, with 369 of the 388 eligible infants being recruited over the two-year study period.

The study was approved by West London Research Ethics Committee Two, United Kingdom (reference number 10/H0711/39).
Parents gave written informed consent for their infant to participate in the study.

Sample collection

Faecal samples were collected by nursing staff from diapers using a sterile spatula, placed in a sterile DNase-, RNase-free Eppendorf tube, stored at − 20°C within two hours of collection and stored at − 80°C within five days. Almost every faecal sample produced by each enrolled infant between recruitment and discharge was collected.

Case Definition, Control Selection, and Clinical Management

NEC cases were defined using the Vermont Oxford Network criteria and staged using the Bell modified staging criteria [20, 21]. The closest available sample prior to the day of diagnosis was selected for twelve infants with NEC. For each of these samples, a sample from a control infant was matched by day of life the samples were taken, and the delivery mode and antibiotics use of each of the infants. Infant details are shown in Additional File 2. Investigators were not involved in clinical care.

DNA extraction and shotgun library preparation

DNA extractions and shotgun library preparation were performed as described previously [22], with the inclusion of a negative extraction control and a technical replicate (C12R). Briefly, DNA extractions were performed with 200 mg of faeces and the process included selective lysis of eukaryotic cells and degradation of eukaryotic DNA, followed by bead-beating to extract bacterial DNA prior to purification. The DNA was fragmented using the NEBNext dsDNA fragmentase kit (NEB), and shotgun libraries prepared using the KAPA HyperPrep kit (KAPA Biosystems). Ligated libraries were PCR amplified with the number of cycles being dependant on starting material (between two and eight cycles). The negative

extraction control and two faceal samples (one NEC, one control) required > 8 PCR cycles, hence the samples were excluded from downstream analysis. The library was normalised, pooled and diluted to 1.6pM prior to loading on an Illumina NextSeq 500 system. The fragment size of the libraries sequenced ranged from 244 bp to 288 bp, with a mean of 261 bp.

Shotgun metagenomic sequencing

Paired end sequencing was performed using a v2 300 cycle high output reagent kit (Illumina), generating over 300 million PE reads and yielding 90.4 Gbp of sequence data. Within the 22 infant samples, and excluding replicates, this translates to a mean 10.2 million PE reads or 2.6 Gbp sequence per sample (see Additional file 6).

Sequencing data availability

Metagenomic sequencing data is available from the EBI European Nucleotide Archive under the study accession PRJEB24015.

Processing of metagenomic sequences

Sequence quality was calculated using FastQC (v0.11.3) [23]. Read filtering was performed using Trimmomatic (v0.36)[24], with removal of the leading and trailing basepairs with phred qualities < 20 and reads with a mean base phred score quality < 20 over a 4 bp sliding window (parameters of: LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20). Sequences with less than 40 bp remaining were discarded (MINLEN:40).

Surviving sequences were screened for human host and vector contamination using FastQ Screen (v0.8.0) and the short read aligner Bowtie2 (v2.2.6) [25, 26]. For validation samples, this screening process was performed on a subset of approximately 100,000 reads per sample. Reads were mapped against the human genome (GRCh38) and the UniVec (build 9.0) vector database, with unmapped reads continuing to the downstream analysis.

Taxonomic identification and read characterisation

Taxonomic binning was performed using DIAMOND (v0.8.36) and MEGAN (v6.6.7) [27, 28]. A Lowest Common Ancestor algorithm with default MEGAN settings was used to assign an NCBI taxonomy, with relative abundances and binned bacterial reads being extracted. Taxonomic assignments were used to classify whether the reads belonged to Gram-negative organisms (with sub-classification of the Negativicutes). The average CpG content per megabase for each bacterial taxonomic group was calculated based on the reads assigned by the taxonomic binning and counting occurrences of “CG” using grep.

Taxonomic groups were summarised to species level where strain level identifications were provided.

Bacterial quantification

Quantitative PCR (qPCR) was performed on the same extracted DNA samples used for the shotgun metagenomic sequencing. qPCR standards were derived from a *Pseudomonas aeruginosa* PAO1 full-length 16S rRNA gene cloned into TOPO TA vector and primers were BAC338F (500 nM) and BAC805R (200 nM) with a BAC516P reporter ((FAM)-TGCCAGCAGCCGCGGTAATAC-(BHQ-1)). The qPCR protocol involved an initial incubation at 95˚C for 10 min, followed by 40 cycles of (95˚C, 15 sec; 60˚C, 1 min) with data collected at 60˚C. The reaction was performed using a TaqMan Universal PCR Master Mix kit on a StepOnePlus Real-Time PCR System (Applied Biosystems). Results are shown in Additional file 2.

qPCR results were corrected for the 16S rRNA gene copy numbers of constituent bacteria. This was performed using the community structures provided by the shotgun metagenomic data, corrected for genome length, and 16S rRNA gene copy numbers for each taxonomic group (see Additional file 2). Where available copy numbers were taken from existing literature [29, 30], and were otherwise calculated from complete NCBI genomes [31]. If no genome was available, the copy number of a higher ranking taxonomic group was used as an estimate.

Calculation of immunostimulatory proxies

To estimate the activation of TLR9, the amount of CpG DNA per gram of faeces was calculated using the following steps; qPCR provided an absolute 16S rRNA copy number per gram of faeces, and metagenomic sequencing estimated the proportions of different bacteria in the community (with the proportions of binned DNA being corrected for genome length). 16S rRNA copy numbers were sourced for each of these bacterial groups (see Additional file 2) and used to weight the corrected bacterial community proportions (given that bacteria with higher 16S rRNA copy numbers appear to be more numerous by 16S rRNA quantification). The qPCR derived 16S rRNA copies per gram was divided according to these proportions giving the number of 16S rRNA copies for each taxonomic group per gram of faeces. Division of these values by the number of 16S rRNA copies for each bacterial group gave the absolute number of bacteria. These could then be multiplied by the average genome length of the taxonomic group to give the total amount of DNA for each taxonomic group per gram of faeces. Multiplication by the average CpG copies per megabase of DNA for each bacterial group (as previously calculated) then gave an estimate of the total CpG motif content per gram.

The same process was also performed on the DNA motif “GTCGTT”, which has been identified as the optimal motif for stimulation of TLR9 [32]. GTCGTT occurrences per gram were found to be linearly related to CpG per gram and given that data is unavailable for the stimulatory potential of other CpG topologies, CpG per gram was taken forward for this analysis.

To estimate the activation of TLR4, the amount of Gram-negative bacteria in each sample was calculated using the previously derived abundances per taxonomic group. We excluded the order Negativicutes from the total of Gram negative bacteria given the reduced TLR4 stimulation exhibited by *Veillonella parvula* [33], which we extrapolate to be a feature of the family resulting from their genetic distance from other Gram negative organisms.

Statistics

Clustering of samples by Euclidean distance was performed in MATLAB. Statistical analyses were performed in R studio version 1.2.5042 running R version 4.0.0 and using the Vegan package [34–36]. The CCA used the most abundant taxonomic groups that constitute 95% of the sequencing reads of the dataset. Bacterial counts per gram of faeces were compared using the Wilcoxon Signed-Rank test due to small group sizes. Linear regressions were used to explore the relationship between total bacterial counts and both CpG counts and Gram negative bacteria counts for control infants with the spread of the data being expressed as IQR around the regression line due to sample sizes. Bayesian generalized linear models with binomial distributions, implemented with the arm package [37], were used segregate cases and control groups, with initial models comparing the two sets of cases highlighted in Fig. 4 versus all non-validation control infants. Cases that fell within either NEC group were added or removed iteratively as indicated by the models. The coefficients of factors retained in the two final models were used to calculate predicted probabilities for the validation dataset.

Abbreviations

CCA Canonical correlation analysis

IRAK-M Interleukin-1 receptor associated kinase

LPS Lipopolysaccharide

NEC Necrotising enterocolitis

NICU Neonatal intensive care unit

NIHR National Institute for Health Research

qPCR Quantitative polymerase chain reaction

TLR Toll-like receptor

Declarations

Ethics approval and consent to participate

The study “Defining the Intestinal Microbiota in Premature Infants” (ClinicalTrials.gov identifier NCT01102738) was approved by West London Research Ethics Committee Two (National Health Service Health Research Authority), United Kingdom (reference number 10/H0711/39). Parents gave written informed consent for their infant to participate in the study.

Consent for publication

Not applicable.

Availability of data and material

Metagenomic sequencing data is available from the EBI European Nucleotide Archive under the study accessions PRJEB15257 and PRJEB19677.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by funding from the Winnicott Foundation; Micropathology Ltd; and the National Institute for Health Research (NIHR) Biomedical Research Centre based at Imperial Healthcare NHS Trust and Imperial College London. KS was funded during this work by an NIHR Doctoral Research Fellowship [NIHR-DRF-2011-04-128]. This article presents independent research funded by the NIHR. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health or other funders. Funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

AGS contributed to the manuscript and analysed the data. KS conducted the sample collection study, extracted DNA from samples and contributed to the manuscript. GR processed the metagenomic sequencing data and contributed to the manuscript. DJW performed the metagenomic sequencing. MSL contributed to experimental design and performed the qPCR experiments. RVJ contributed to experimental design and analysis of data. SG and JSK contributed to experimental design and directed analysis. All authors read and approved the final manuscript.

Acknowledgements

We thank the participants and their families for their contribution to this study.

References

1. Cotten CM, Taylor S, Stoll B, Goldberg RN, Hansen NI, Sanchez PJ, Ambalavanan N, Benjamin DK, Jr.: Prolonged duration of initial empirical antibiotic treatment is associated with increased rates of necrotizing enterocolitis and death for extremely low birth weight infants. *Pediatrics* 2009, 123:58-66.
2. Sim K, Shaw AG, Randell P, Cox MJ, McClure ZE, Li MS, Haddad M, Langford PR, Cookson WO, Moffatt MF, Kroll JS: Dysbiosis anticipating necrotizing enterocolitis in very premature infants. *Clin Infect Dis* 2015, 60:389-397.
3. Mai V, Young CM, Ukhanova M, Wang X, Sun Y, Casella G, Theriaque D, Li N, Sharma R, Hudak M, Neu J: Fecal microbiota in premature infants prior to necrotizing enterocolitis. *PLoSOne* 2011, 6:e20647.

4. Heida FH, van Zoonen AG, Hulscher JB, te Kiefte BJ, Wessels R, Kooi EM, Bos AF, Harmsen HJ, de Goffau MC: A Necrotizing Enterocolitis-Associated Gut Microbiota Is Present in the Meconium: Results of a Prospective Study. *Clin Infect Dis* 2016, 62:863-870.
5. Warner BB, Deych E, Zhou Y, Hall-Moore C, Weinstock GM, Sodergren E, Shaikh N, Hoffmann JA, Linneman LA, Hamvas A, et al: Gut bacteria dysbiosis and necrotising enterocolitis in very low birthweight infants: a prospective case-control study. *Lancet* 2016, 387:1928-1936.
6. Sodhi CP, Neal MD, Siggers R, Sho S, Ma C, Branca MF, Prindle T, Jr., Russo AM, Afrazi A, Good M, et al: Intestinal epithelial Toll-like receptor 4 regulates goblet cell development and is required for necrotizing enterocolitis in mice. *Gastroenterology* 2012, 143:708-718 e701-705.
7. Gribar SC, Sodhi CP, Richardson WM, Anand RJ, Gittes GK, Branca MF, Jakub A, Shi XH, Shah S, Ozolek JA, Hackam DJ: Reciprocal expression and signaling of TLR4 and TLR9 in the pathogenesis and treatment of necrotizing enterocolitis. *J Immunol* 2009, 182:636-646.
8. Meng D, Zhu W, Shi HN, Lu L, Wijendran V, Xu W, Walker WA: Toll-like receptor-4 in human and mouse colonic epithelium is developmentally regulated: a possible role in necrotizing enterocolitis. *Pediatr Res* 2015, 77:416-424.
9. Bauer S, Kirschning CJ, Hacker H, Redecke V, Hausmann S, Akira S, Wagner H, Lipford GB: Human TLR9 confers responsiveness to bacterial DNA via species-specific CpG motif recognition. *Proc Natl Acad Sci U S A* 2001, 98:9237-9242.
10. Kant R, de Vos WM, Palva A, Satokari R: Immunostimulatory CpG motifs in the genomes of gut bacteria and their role in human health and disease. *J Med Microbiol* 2014, 63:293-308.
11. Ghadimi D, Vrese M, Heller KJ, Schrezenmeir J: Effect of natural commensal-origin DNA on toll-like receptor 9 (TLR9) signaling cascade, chemokine IL-8 expression, and barrier integritiy of polarized intestinal epithelial cells. *Inflamm Bowel Dis* 2010, 16:410-427.
12. Raveh-Sadka T, Thomas BC, Singh A, Firek B, Brooks B, Castelle CJ, Sharon I, Baker R, Good M, Morowitz MJ, Banfield JF: Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. *Elife* 2015, 4.
13. Stewart CJ, Embleton ND, Marrs EC, Smith DP, Nelson A, Abdulkadir B, Skeath T, Petrosino JF, Perry JD, Berrington JE, Cummings SP: Temporal bacterial and metabolic development of the preterm gut reveals specific signatures in health and disease. *Microbiome* 2016, 4:67.
14. Neu J, Walker WA: Necrotizing enterocolitis. *N Engl J Med* 2011, 364:255-264.
15. La Rosa PS, Warner BB, Zhou Y, Weinstock GM, Sodergren E, Hall-Moore CM, Stevens HJ, Bennett WE, Jr., Shaikh N, Linneman LA, et al: Patterned progression of bacterial populations in the premature infant gut. *Proc Natl Acad Sci U S A* 2014, 111:12522-12527.
16. Fouhy F, Guinane CM, Hussey S, Wall R, Ryan CA, Dempsey EM, Murphy B, Ross RP, Fitzgerald GF, Stanton C, Cotter PD: High-throughput sequencing reveals the incomplete, short-term recovery of infant gut microbiota following parenteral antibiotic treatment with ampicillin and gentamicin. *Antimicrob Agents Chemother* 2012, 56:5811-5820.

17. Arboleya S, Sanchez B, Milani C, Duranti S, Solis G, Fernandez N, de los Reyes-Gavilan CG, Ventura M, Margolles A, Gueimonde M: Intestinal microbiota development in preterm neonates and effect of perinatal antibiotics. *J Pediatr* 2015, 166:538-544.
18. Zmora N, Zilberman-Schapira G, Suez J, Mor U, Dori-Bachash M, Bashiardes S, Kotler E, Zur M, Regev-Lehavi D, Brik RB, et al: Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features. *Cell* 2018, 174:1388-1405 e1321.
19. Cho SX, Berger PJ, Nold-Petry CA, Nold MF: The immunological landscape in necrotising enterocolitis. *Expert Rev Mol Med* 2016, 18:e12.
20. Kliegman RM: Neonatal necrotizing enterocolitis: bridging the basic science with the clinical disease. *JPediatr* 1990, 117:833-835.
21. Vermont Oxford Network Database. Manual of Operations. Part 2: Data Definitions and Data Forms For Infants Born in 2013 [http://www.vtoxford.org/tools/ManualofOperationsPart2.pdf]
22. Rose G, Shaw AG, Sim K, Wooldridge DJ, Li MS, Gharbia S, Misra R, Kroll JS: Antibiotic resistance potential of the healthy preterm infant gut microbiome. *PeerJ* 2017, 5:e2928.
23. FastQC: a quality control tool for high throughput sequence data [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/]
24. Bolger AM, Lohse M, Usadel B: Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014, 30:2114-2120.
25. FastQ Screen [http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/]
26. Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012, 9:357-359.
27. Buchfink B, Xie C, Huson DH: Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015, 12:59-60.
28. Huson DH, Auch AF, Qi J, Schuster SC: MEGAN analysis of metagenomic data. *Genome Res* 2007, 17:377-386.
29. Vetrovsky T, Baldrian P: The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One* 2013, 8:e57923.
30. Sun DL, Jiang X, Wu QL, Zhou NY: Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. *Appl Environ Microbiol* 2013, 79:5962-5969.
31. National Center for Biotechnology Information [http://www.ncbi.nlm.nih.gov/]
32. Wang X, Bao M, Wan M, Wei H, Wang L, Yu H, Zhang X, Yu Y, Wang L: A CpG oligodeoxynucleotide acts as a potent adjuvant for inactivated rabies virus vaccine. *Vaccine* 2008, 26:1893-1901.
33. Matera G, Muto V, Vinci M, Zicca E, Abdollahi-Roodsaz S, van de Veerdonk FL, Kullberg BJ, Liberto MC, van der Meer JW, Foca A, et al: Receptor recognition of and immune intracellular pathways for Veillonella parvula lipopolysaccharide. *Clin Vaccine Immunol* 2009, 16:1804-1809.
34. Team RStudio (2016) RStudio: Integrated Development for R. RStudio, Inc., Boston, MA, URL http://www.rstudio.com/

35. R Core Team (2018), R: A Language and Environment for Statistical Computing, URL https://www.R-project.org.
36. vegan: Community Ecology Package [https://CRAN.R-project.org/package=vegan]
37. arm: Data Analysis Using Regression and Multilevel/Hierarchical Models [https://CRAN.R-project.org/package=arm]

Figures

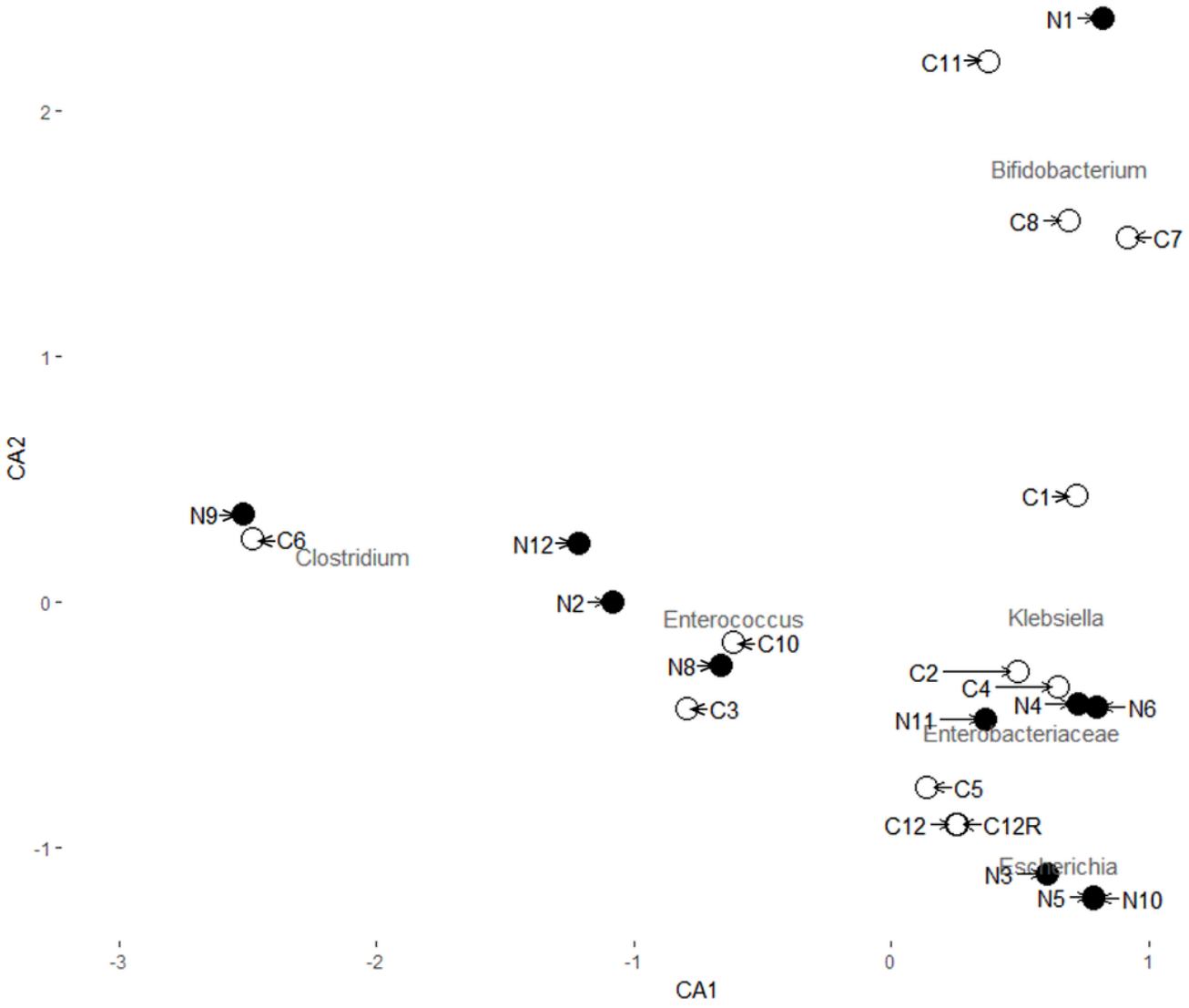

Figure 1

Canonical correlation analysis plot showing the similarity of the processed samples. Analysis was performed using data quantifying the most abundant taxonomic groups comprising 95% of the sequencing reads (n=24; 11 NEC samples (black), 11 control samples (white), 1 technical repeat (white, C12R)). Taxonomic names indicate the major bacterial groups associated with nearby samples which drive the separation.

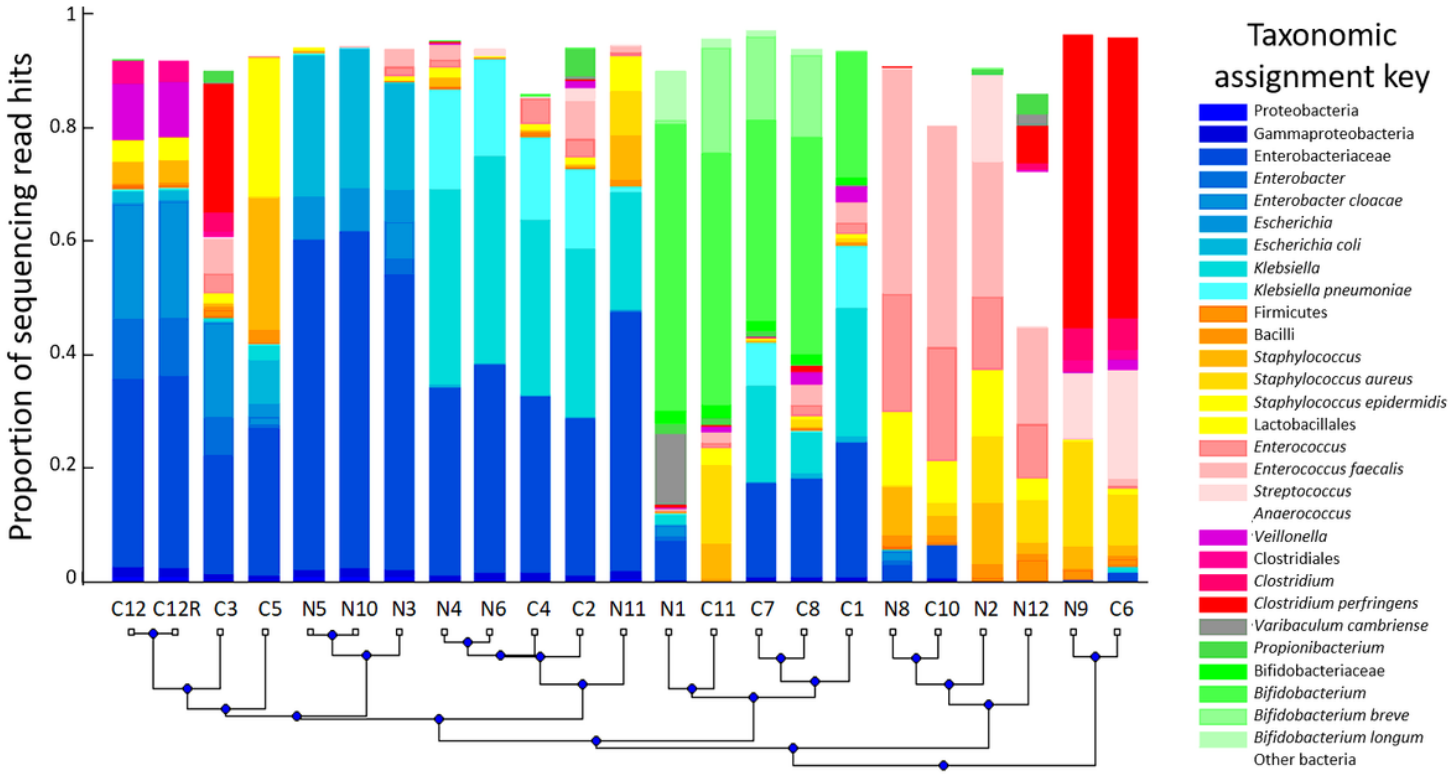

Figure 2

Samples clustered by taxonomic assignment of sequencing reads. Y axis indicates proportion of reads assigned to each taxonomic category as coded in the colour key. X axis indicates the samples, clustered by similarity (n=22; 11 NEC samples, 11 control samples).

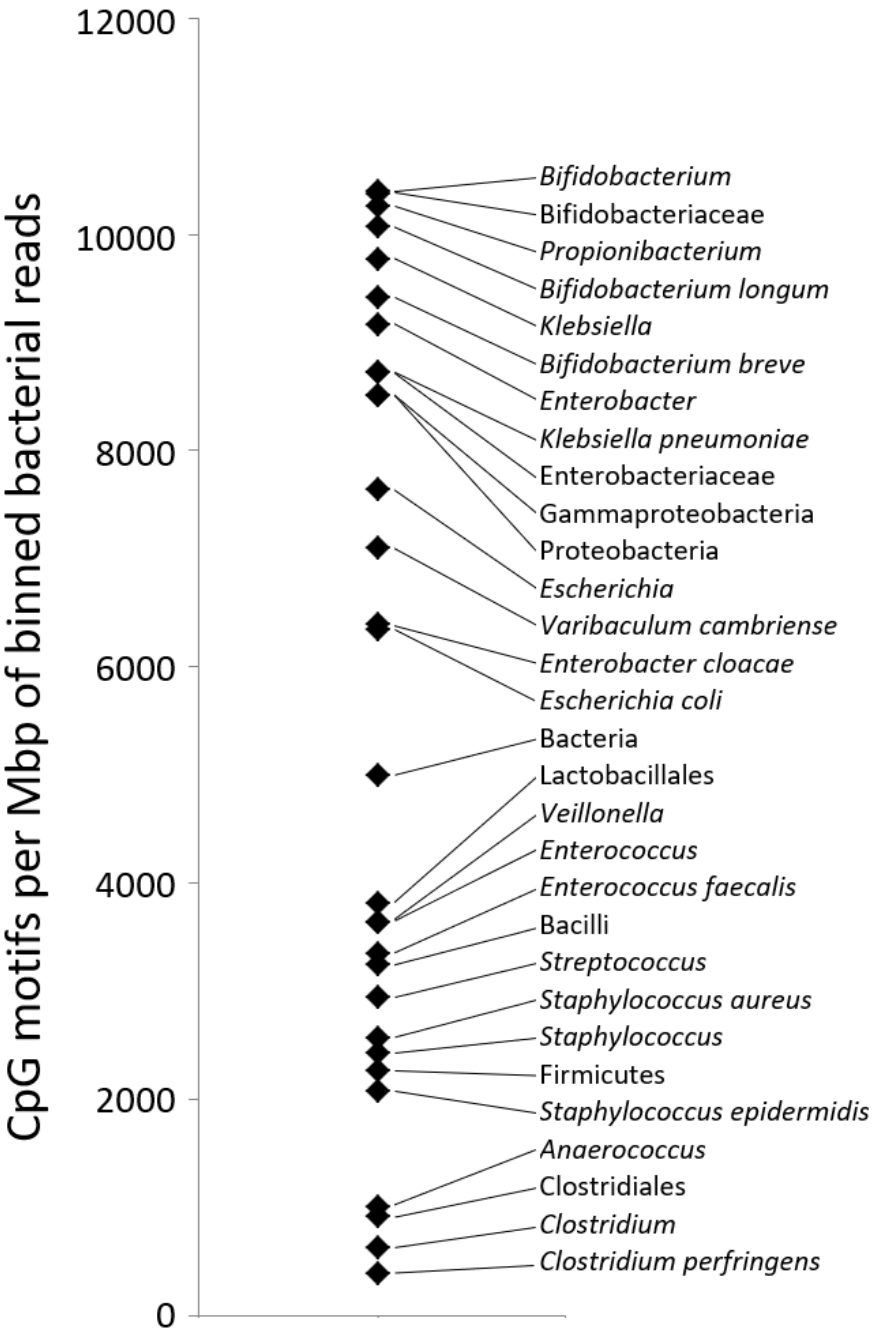

Figure 3

CpG motif frequency in premature infant gut colonisers. Binned bacterial reads for the taxonomic groups making up the top 95% of classifications are stratified according to the number of CpG motifs per megabase of DNA.

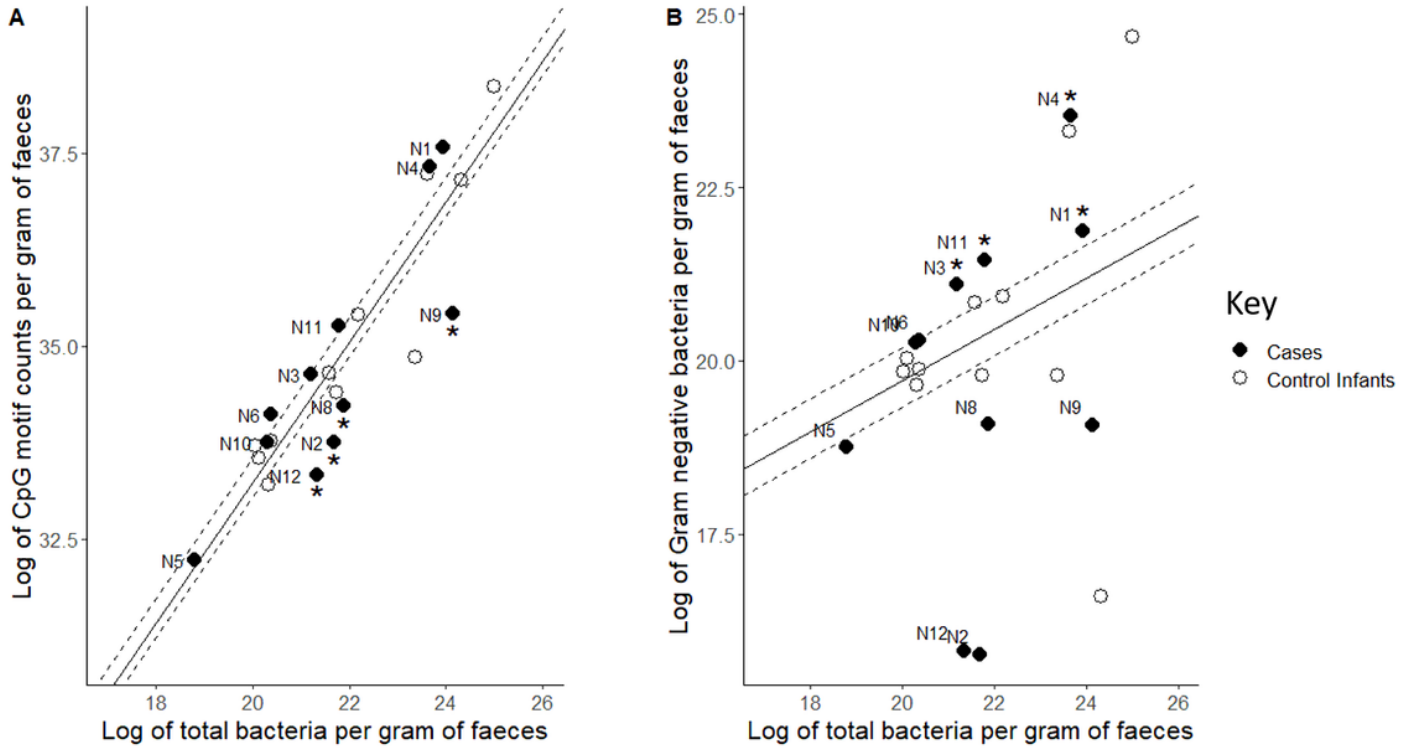

Figure 4

Stratification of NEC and Control samples by bacterial load and A) CpG DNA content or B) Gram-negative bacteria. Control and NEC samples are stratified by bacteria per gram of faeces (x axis) and A) the occurrences of CpG DNA per gram of faeces (y axis) and B) number of Gram-negative bacteria per gram of faeces (y axis). The solid diagonal lines indicate the relationship established between each pair of factors for control infants using linear regression. The interquartile range (IQR) of the deviation of control samples from the regression line is indicated by dashed lines. Stars indicate NEC cases identified as either having particularly low CpG DNA per gram of faeces or particularly high abundances of Gram-negative bacteria compared to control infants, as defined by being outside the IQR of the control samples.

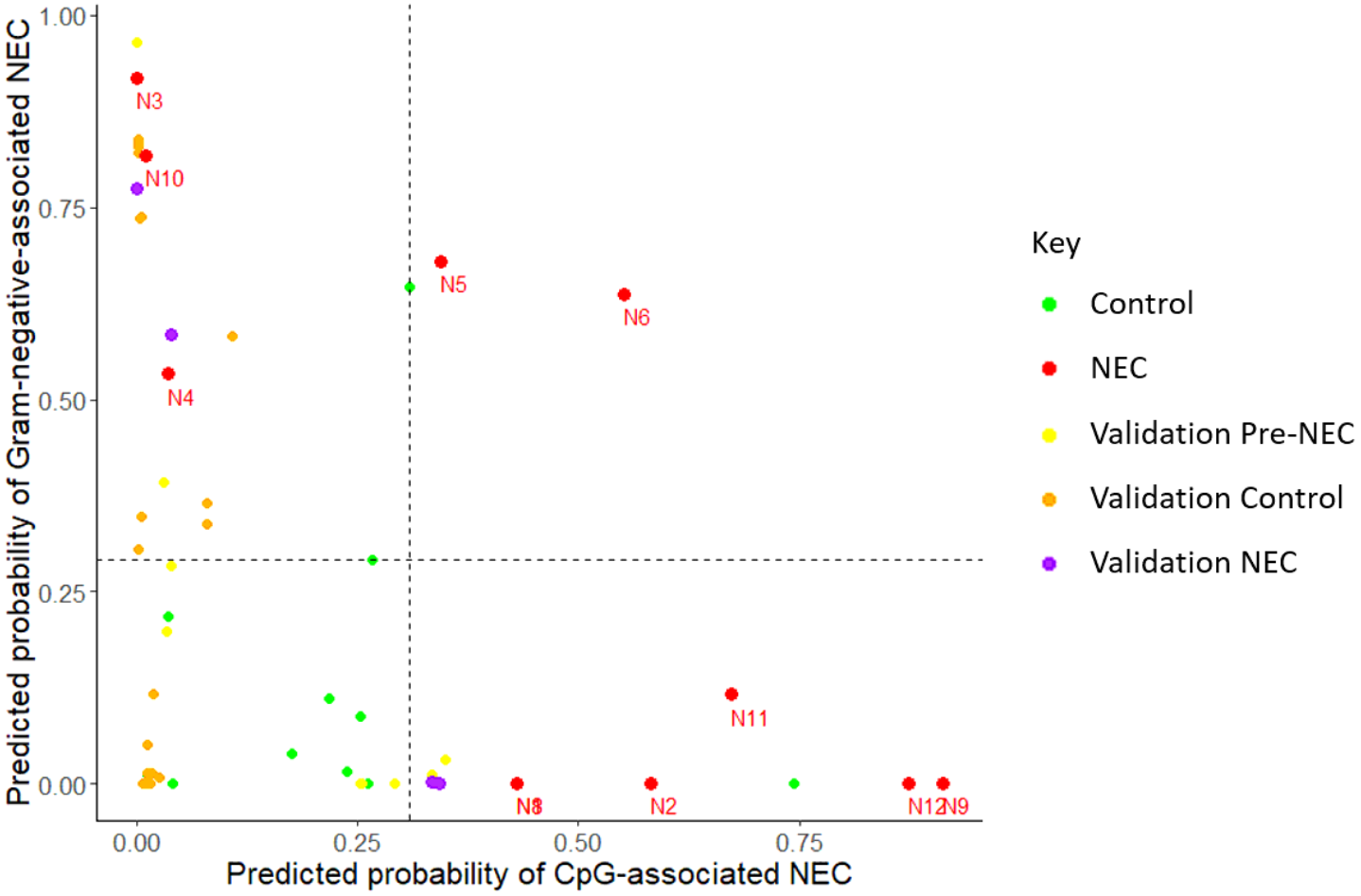

Figure 5

Predicted probabilities from the Bayesian regression models for training and validation datasets. Predicted probabilities for the CpG-related NEC and Gram-negative-related NEC models are shown for 11 control infants, 11 NEC cases and for the validation dataset. Dashed lines indicate the 90% quantile of predicted probability for the eleven control samples along each axis. “NEC” samples are the closest available samples to an infant’s NEC diagnosis for each dataset and “Pre-NEC” refers to any prior samples from these infants. The points for samples “N1” and “N8” are overlaid.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.tif
- Additionalfile2.xlsx
- Additionalfile3.xlsx
- Additionalfile4.xlsx