Supplemental Digital Content (SDC) 1

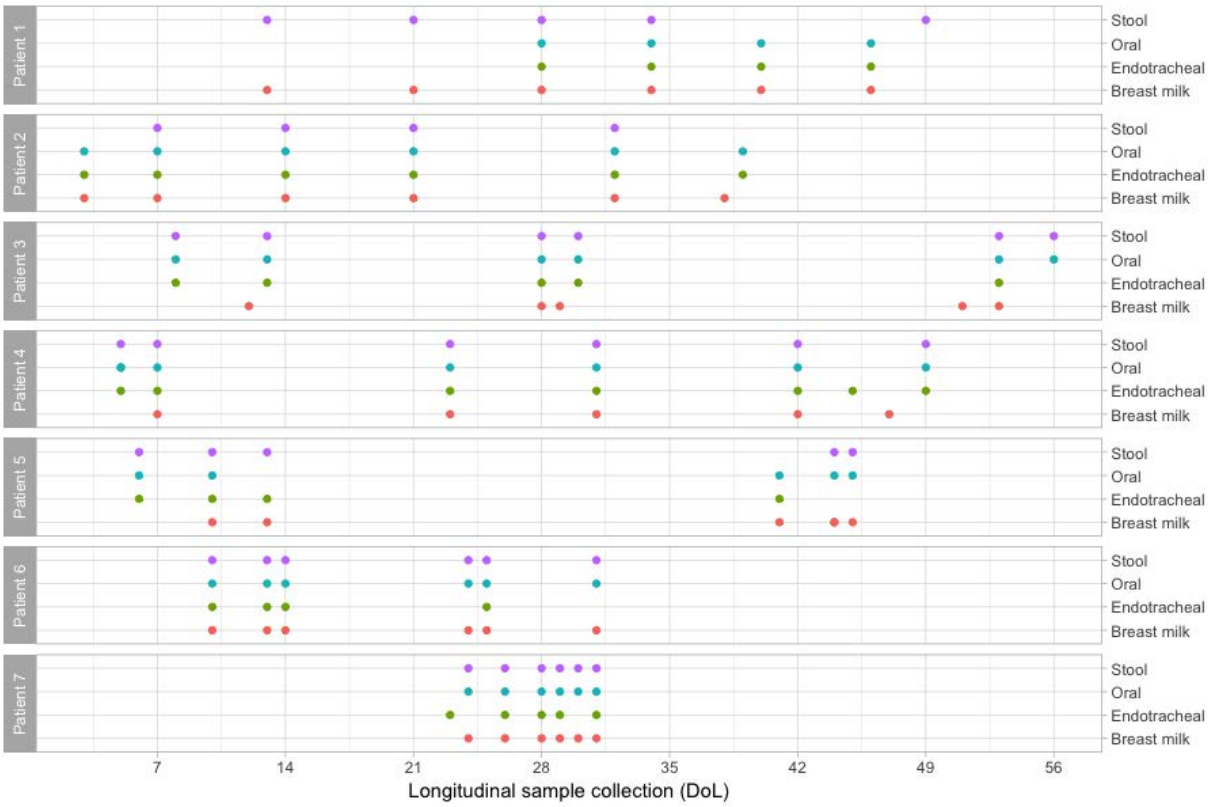
Participant Inclusion Criteria

Patients within the NICU for which parental consent was agreed were sampled as and when materials were available. No pre-defined sampling strategy was developed to minimise impact on clinical time and, as such, all sample collections were opportunistic.
Patients were identified as suitable for inclusion if they met the criteria outlined below (arranged by order of priority):

- Patients <26 weeks GA at birth
  (enabling characterisation of microbiota development in the most vulnerable infants in the NICU)
- All samples collected from timepoints spanning first 8 weeks of life
  (facilitate characterisation of initial microbiota acquisition, not just changes over time)
- Minimum of 5 of each sample type available per timepoint
  (enabling sufficient tracing of microbes across time)
- Maximum of 24h between all samples at a single timepoint
  (maximising longitudinal connectedness of microbiota timepoints)
- Patients inhabited single NICU from birth to discharge
  (minimising impacts of transfer from one NICU to another and differences arising due to variation in clinical care of patients)
- Sufficient volume for every sample to be split for both NVCE analysis
  (enabling both PMA- and non-PMA-treated fractions of each sample to be analysed to identify impact of relic DNA on results of previous studies)

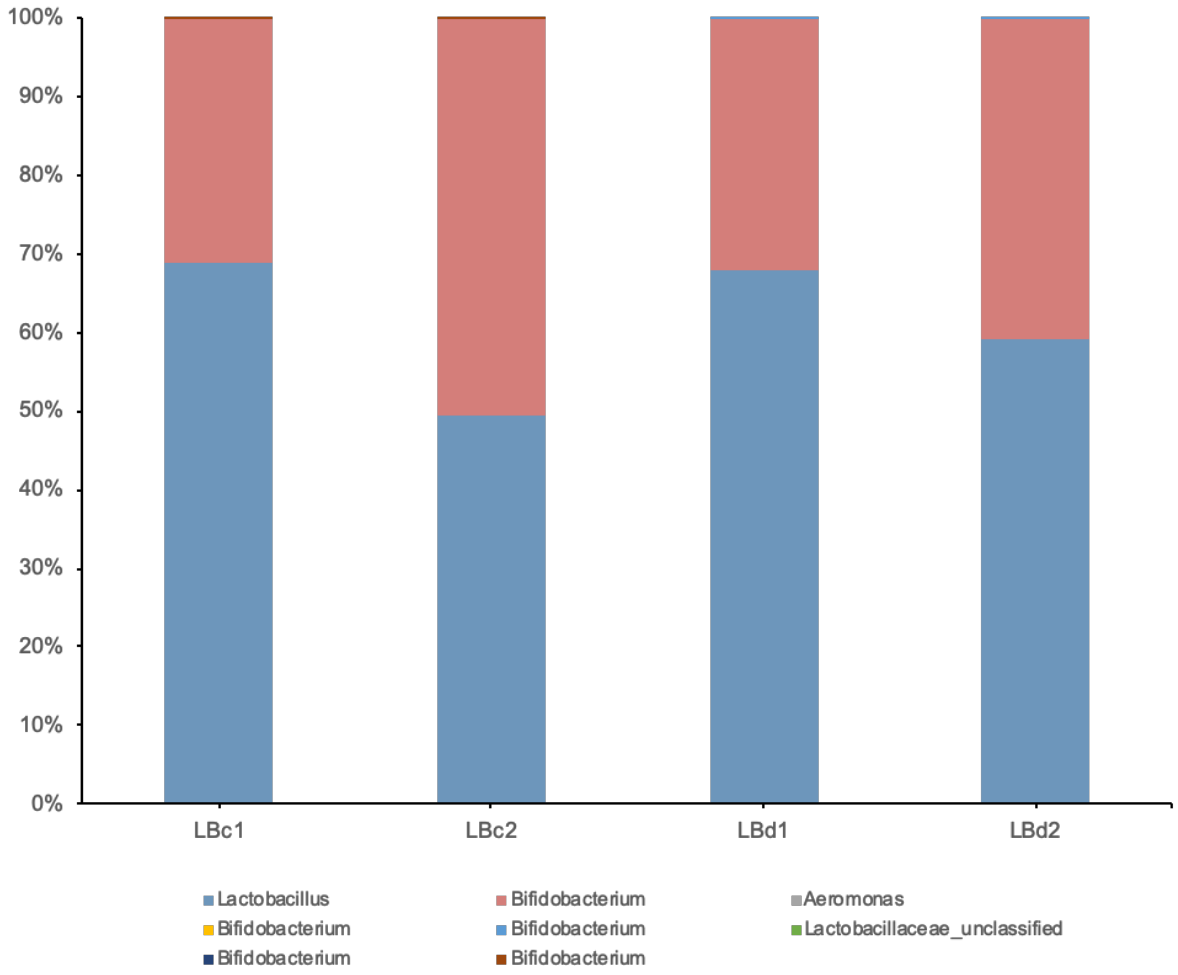
Due to low numbers of such extremely preterm infants providing suitable sample frequencies or volumes some patients provided samples across each of the first 8 weeks of life while others only provided samples across a single week (Figure, SDC2)

Figure SDC2



Longitudinal sampling achieved over the first 8 weeks of life for each patient, colour coded by sample type. Collection of waste breast milk (red), endotracheal (green), oral (blue) and stool (purple), samples was purely opportunistic in an attempt to reduce infant handling and intervention. Inclusion criteria and sampling strategy is described in Methods, SDC1. DoL = day of life

Figure SDC3



Composition of the Probiotic suspension administered to all preterm infants enrolled in this study. Stacked bars illustrate relative sequence abundance of all OTU level bacterial sequences contributing >5 total sequence reads. Each bar represents an individual sample with composition dominated by members of the *Bifidobacterium* & *Lactobacillus* genera. DNA was extracted as described in the methods from individual aliquots (LBd1/2), and swabs of the inner surface of the dropper cap (LBc1/2), in duplicate. Extracted DNA samples were prepared for targeted sequencing and processed for analysis as per the methods section. Included below is also the raw sequencing data used to generate the bar plots

	Raw reads				
Taxonomy	LBc1	LBc2	LBd1	LBd2	Total
Lactobacillus	16858	77549	12955	16692	124054
Bifidobacterium	7602	78989	6081	11591	104263
Aeromonas	0	118	0	0	118
Bifidobacterium	5	0	0	5	10
Bifidobacterium	0	3	4	2	9
Lactobacillaceae_unclassified	0	8	0	0	8
Bifidobacterium	2	5	0	0	7
Bifidobacterium	4	3	0	0	7
Stenotrophomonas	0	3	2	0	5
Haemophilus	0	4	0	1	5
Pasteurella	3	0	0	1	4
Streptococcus	4	0	0	0	4
Anaerococcus	3	0	0	0	3
Prevotella	3	0	0	0	3
Clostridiales_unclassified	3	0	0	0	3
Fusobacteriaceae_unclassified	3	0	0	0	3
Bacteroides	3	0	0	0	3
Xanthomonadaceae_unclassified	3	0	0	0	3
Bacteria_unclassified	0	0	3	0	3
Pseudomonas	0	0	0	2	2
Psychrobacter	0	0	2	0	2
Anaerostipes	2	0	0	0	2
Bacteria_unclassified	0	2	0	0	2
Lactobacillus	2	0	0	0	2
Schlegelella	0	2	0	0	2
Bifidobacterium	0	2	0	0	2
Bacteroides	2	0	0	0	2
Bacteroides	2	0	0	0	2
Moraxella	2	0	0	0	2
Acidobacteria_(Gp1_family_incertae_sedis)	0	0	0	2	2
Enterobacteriaceae_unclassified	2	0	0	0	2
Rhodoferax	2	0	0	0	2
TOTAL READS	24510	156688	19047	28296	

Figure SDC4

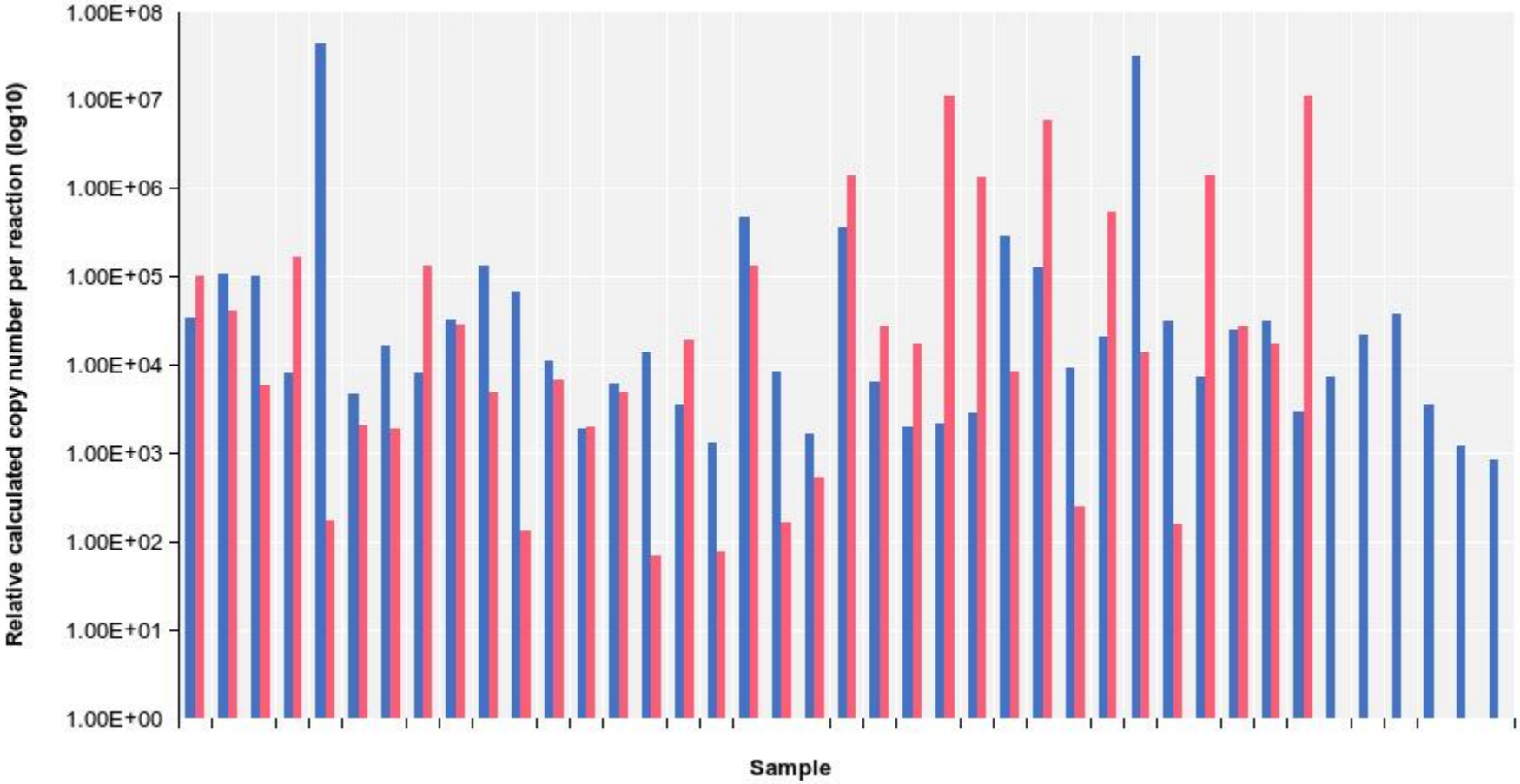


Figure SDC4 illustrates calculated copy number of whole 16S rRNA gene per 20uL reaction based on qPCR quantification as explained in methods and SDC5 Each bar represents a multiplexed sample of breast milk (blue; mean c/rxn = 1.90 x 10^6) and endotracheal suctions (red; mean c/rxn = 7.22 x 10^5). Bars are grouped by patient and time. Breast milk samples: n = 40. Endotracheal samples: n = 35. Calculated copy numbers per reaction, based upon the copy numbers of known standards used ranged from 69 to 4.36 x 10^7.

SDC5

Quantifying extracted 16S rRNA gene copy number

Due to high variability in extracted bacterial DNA yield pre-sequencing normalisation of sample DNA content was performed. Total 16S rRNA gene copy number per was quantified within all extracted DNA samples using *QIAGEN QuantiNova SYBR green qRT-PCR kits* and primers 1369F and 1492R as described by Suzuki and colleagues (Table, SDC5, Figure, SDC4). Following qPCR quantification, 16S rRNA gene copy number within samples was normalised to within 1 log unit of 16S rRNA copy number by dilution of extracted DNA samples with sterile PCR grade H_2O, prior to concentration in a *Christ 2-18 CDPlus Rotational vacuum concentrator* (Petershütte, DE). Concentration of normalised bacterial DNA was performed in 3 x 10 minute intervals at full speed and 60 °C.

Figure SDC5

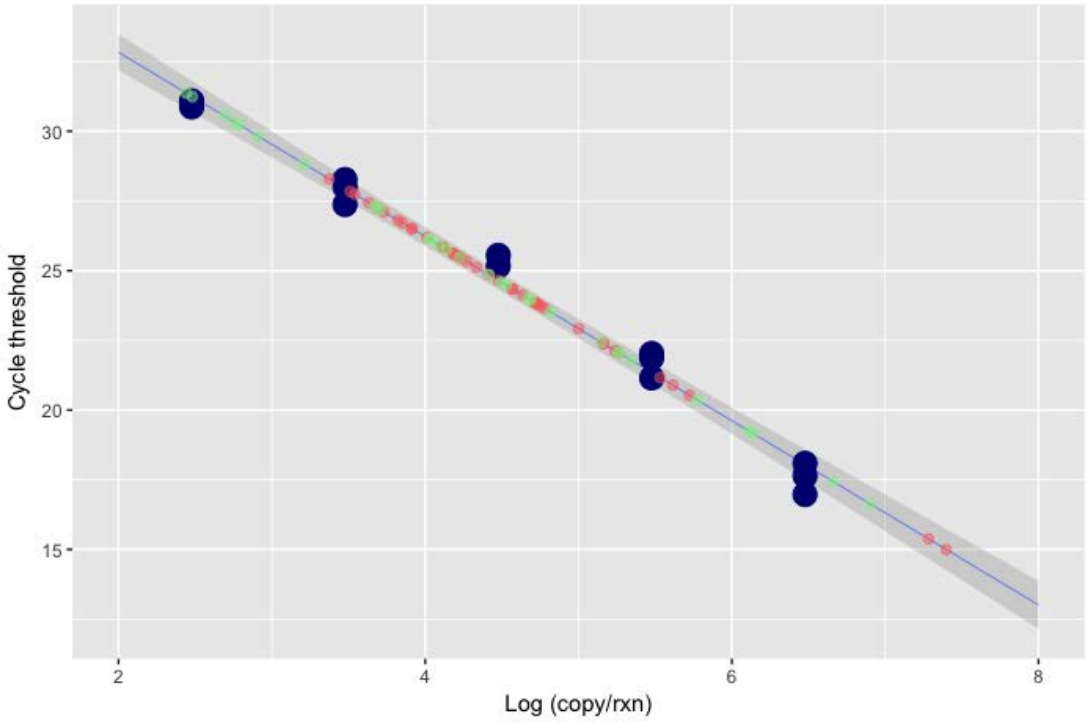


Figure SDC5 illustrates qPCR reaction efficiency and results. Copy number of 16S rRNA gene content per 20 µL reaction observed in breast milk (red) and endotracheal (green) samples, is plotted on a standard curve generated using samples of known DNA concentration (blue) (adjusted R^2 = 0.98, Efficiency = 99.51).

Suzuki MT, Taylor LT, DeLong EF. Quantitative Analysis of Small-Subunit rRNA Genes in Mixed Microbial Populations via 5′-Nuclease Assays. Appl Environ Microbiol 2000; 66:4605–14.

Table SDC6 details primers and cycling conditions used for 16S rRNA sequencing and amplification as performed in this study

Primer pair	Forward Sequence	Reverse Sequence	Cycling conditions
16S content qPCR Suzuki *et al.*, (2000)	1369F 5’ - CGG TGA ATA CGT TCY CGG - 3’	1492R 5’ - GGW TAC CTT GTT ACG ACT - 3’	95 °C 2 mins 95 °C 5 secs. \ _ 35 cycles 60 °C 10 secs. /
Universal 16S rRNA gene amplification Lane (1991) & Turner *et al.*, (1999)	27F 5’ - AGA GTT TGA TCM TGG CTC AG - 3’	1492R 5’ - TAC GGY TAC CTT GTT ACG ACT T - 3’	95 °C 5 mins 95 °C 30 secs \ 44.5 °C 30 secs | - 20 cycles 72 °C 60 secs / 72 °C 10 mins
V4 region 16S rRNA gene multiplexed sequencing by synthesis Caporaso *et al.*, (2011)	515F 5’ - GTG YCA GCM GCC GCG GTA A - 3’	806R 5’ - GGA CTA CNV GGG TWT CTA AT - 3’	95 °C 2 mins 95 °C 20 secs \ 55 °C 15 secs | - 30 cycles 72 °C 5 mins / 72 °C 10 mins

SDC7

Identification of potential contaminant taxa from targeted 16S rRNA sequencing analysis

Potential contaminant taxa were identified by processing kit extraction controls with each batch of samples analysed. Further controls included a sequencing negative, processed with each library prep and a positive control of known taxonomic content.

Once processed, each kit control was included in the library prep and sequenced in the same sequencing reaction as the corresponding samples. All samples and controls were processed as part of one batch in the same Mothur throughput.

Rules for determining potential contaminants were decided prior to processing and were as follows:

1. All total read counts per samples were converted to relative abundance
2. Mean relative abundance was calculated for each taxon in the sequencing negatives
3. Mean relative abundance was calculated for each taxon in all body-site samples
4. Any taxon for which mean relative abundance in the sequencing negative was greater than 20% of the mean relative abundance in the samples was classified as a 'sequencing contaminant' and removed from analysis
5. Mean relative abundance of remaining taxa was calculated for all kit controls
6. Any taxon for which mean relative abundance in the relevant kit control was greater than 50% of the mean relative abundance in the respective samples was classified as a 'kit contaminant' and removed from analysis

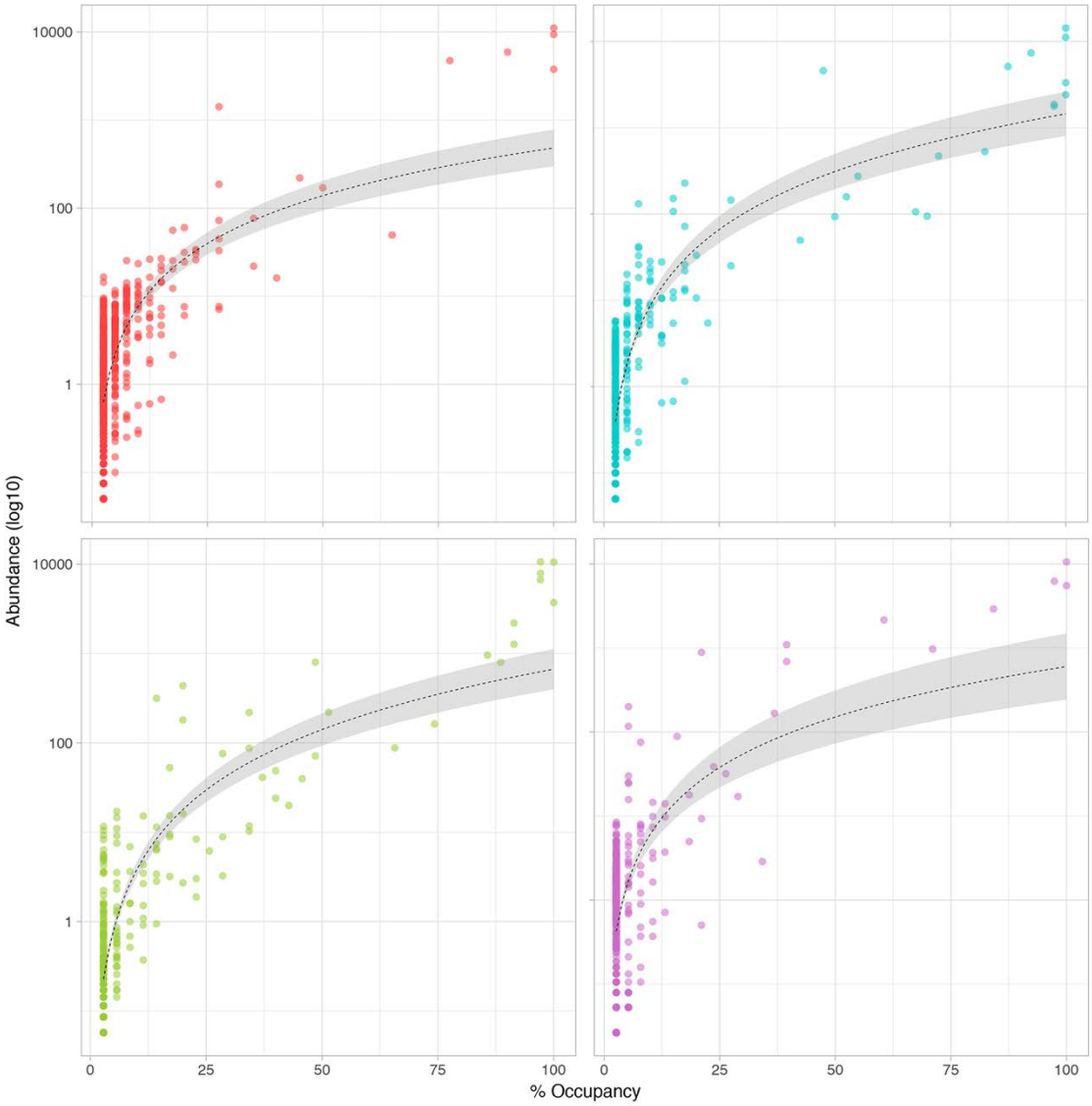
Employing these criteria we identified a total of 86 potential contaminant OTUs (table S1).

Identification of suspected Rhodococcus from breast milk samples

200 µL of remaining breast milk sample following extraction of DNA from all samples used in this study from patient 4 were spread on a large BHI agar plate. Initial culture conditions were 28 °C for 3 days on the large BHI agar spread plate. Subcultures of a single colony from the mixed culture were plated on to single BHI agar plates and incubated at 28 °C for 3 days prior to transfer to the fridge for storage over 2 days.

Single colonies were picked from this plate and DNA extracted using the QIAGEN PowerLyzer PowerSoil DNA extraction kit, as per manufacturers' instructions. Isolated DNA was amplified with universal bacterial PCR using primers 27F and 1492R under the following cycle conditions described in Table S2. PCR amplicons were cleaned using Sigma-Aldrich GenElute PCR cleanup kit (MI, USA) and sent to Eurofins genomics TubeSeq service (Ebersberg. DE) for 16S rRNA gene sanger sequencing using their in-house standard primers 27f and 1492r. Sequencing results were searched against nucleotide BLAST to identify the most likely species classification of the single colony.

Figure SDC8



Distribution abundance relationships for: breast milk (red) F=1032 R^2=0.58; oral (blue) F=358.8 R^2=0.46; endotracheal (green) F=278.5 R^2=0.47; and stool (purple) F=1193 R^2=0.69 microbiotas. All correlations are significant ($p \geq 0.01$). R^2 values given are adjusted, penalising any data-points which do not positively contribute to predictive capacity of the model. Distribution abundance relationships were calculated as being the correlation between relative abundance and sample distribution in all samples from all patients as described by van der Gast *et al.*, 2011.

van der Gast CJ, Walker AW, Stressmann FA, *et al*. Partitioning core and satellite taxa from within cystic fibrosis lung bacterial communities. ISME J, 2011; 5(5): 780–91. Available from: http://www.nature.com/articles/ismej2010175

Table SDC9
16S rRNA sequencing reads obtained for each body sampling site investigated.

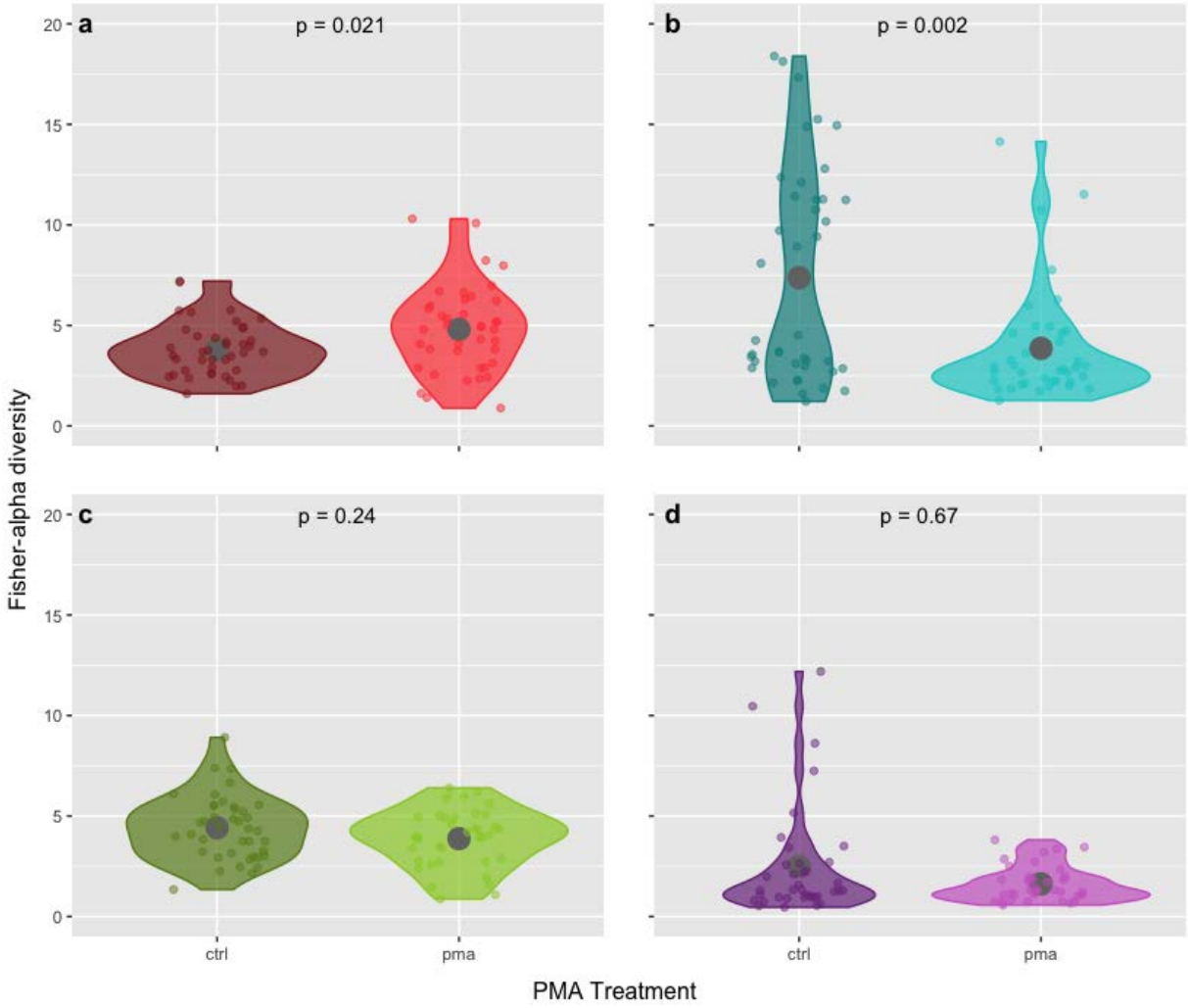
		Total	Mean per sample	Std. dev. per sample
Body sampling site	Breast milk (n = 40)	3.77×10^6	4.59×10^4	3.71×10^4
	Oral secretions (n = 40)	5.01×10^6	6.11×10^4	5.89×10^4
	Endotracheal secretions (n = 35)	3.78×10^6	4.76×10^4	3.93×10^4
	Stool (n = 38)	2.54×10^6	3.34×10^4	3.45×10^4

Table SDC10
Impact and significance of clinical factors on microbiota dissimilarity between samples as calculated by Adonis PERMANOVA with weighted Bray-Curtis dissimilarity across PMA-treated (PMA), and non-PMA-treated (CTRL) fractions of the same samples.

	ALL			PMA			CTRL		
	R^2	P	sig	R^2	P	sig	R^2	P	sig
Patient	0.11365	0.001	***	0.10974	0.002	**	0.129	0.001	***
DoL	0.16085	0.001	***	0.17952	0.027	*	0.18031	0.002	**
Site	0.1374	0.001	***	0.14249	0.001	***	0.14268	0.001	***
Condition	0.01261	0.001	***	0.01422	0.041	*	0.01558	0.001	***
Patient : DoL	0.03064	0.001	***	0.03648	0.04	*	0.03799	0.003	**
Patient : Site	0.15975	0.001	***	0.16572	0.011	*	0.18679	0.001	***
DoL : Site	0.22091	0.001	***	0.30112	0.049	*	0.25038	0.004	**
Patient : Condition	0.0045	0.001	***	N/A	N/A		N/A	N/A	
DoL : Condition	0.00988	0.001	***	N/A	N/A		0.00969	0.003	**
Site : Condition	0.00325	0.001	***	N/A	N/A		0.00502	0.005	**
Patient : DoL : Site	0.0349	0.001	***	0.04762	0.064	.	0.04126	0.01	**
Residuals	0.11165			0.00309			0.0013		
Total	0.99999			1			1		

Clinical factors identified as having a significant impact on microbiota community within each dataset are highlighted with an asterisk.
Nestedness of variables is illustrated by colons, where "Patient : DoL", explores the significance of any longitudinal impact on microbiota composition when controlled for by Patient.
PMA = Propidium monoazide-treated samples (those presented in manuscript); CTRL = non-PMA-treated fraction of the same samples (these results are not reported in manuscript); ALL includes all PMA and non-PMA-treated samples (these comparisons were stratified by PMA-treatment).
Corrected R^2 are reported and Bonferroni corrected P values are given to 3 decimal places. DoL = Day of life.

Figure SDC11



Impact of PMA-treatment on alpha diversity of samples from each sampling site. Fisher alpha diversity was calculated for each for breast milk (a); oral (b); endotracheal (c); and stool (d), sample and depicted by individual points. Value density is represented by cloud width and mean per group is depicted by grey circles. Results of Kruskal-wallis test are inset in each figure panel.

Table SDC12
Bonferroni-corrected P values for pairwise Mann-Whitney Wilcoxon test between Shannon alpha diversity of all body-sites sampled at WoL 7 & 8.

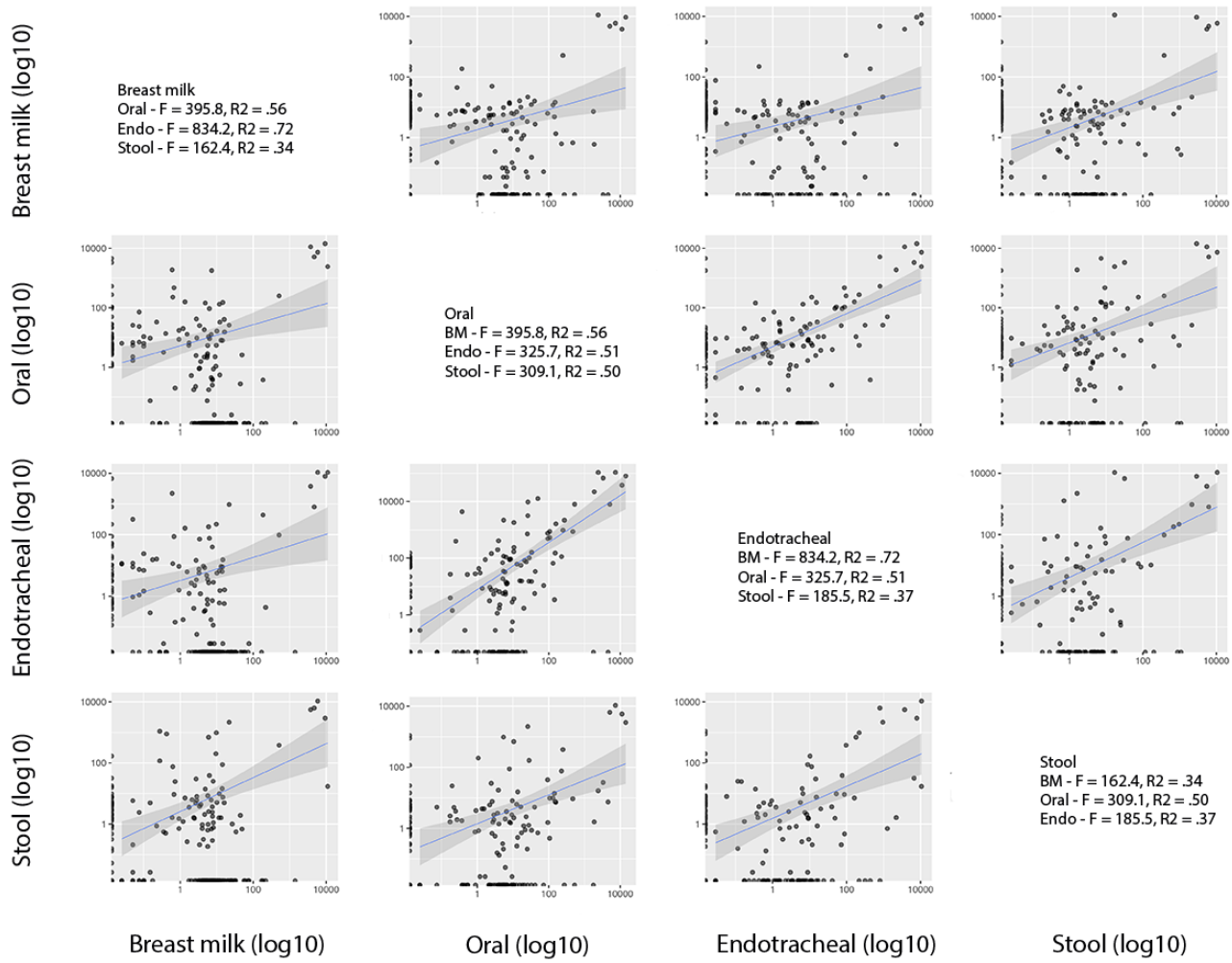
		SITE			
		Breast milk	Endotracheal	Oral	Stool
SITE	Breast milk	-	0.10	0.05*	0.001*
	Endotracheal	0.10	-	1	0.10
	Oral	0.05*	1	-	0.11
	Stool	0.001*	0.10	0.11	-

Cells highlighted with an asterisk contain values under the Bonferroni threshold for multiple testing (p = 0.05).

Table SDC13
Bonferroni-corrected P values and R^2 values for pairwise PERMANOVA analysis of weighted Bray-Curtis based beta-diversity values for all body-sites.

		SITE			
		Breast milk	Endotracheal	Oral	Stool
SITE	Breast milk	-	R^2 - 0.13 P.adj – 0.006	R^2 - 0.06 P.adj – 0.006	R^2 - 0.16 P.adj – 0.006
	Endotracheal	R^2 - 0.13 P.adj – 0.006	-	R^2 - 0.01 P.adj – 0.006	R^2 - 0.19 P.adj – 0.006
	Oral	R^2 - 0.06 P.adj – 0.006	R^2 - 0.01 P.adj – 0.006	-	R^2 – 0.10 P.adj – 0.006
	Stool	R^2 - 0.16 P.adj – 0.006	R^2 - 0.19 P.adj – 0.006	R^2 - 0.10 P.adj – 0.006	-

Figure SDC14



Strength of pairwise linear regression correlations between abundance of each individual taxa in all sampling sites, plotted on a log scale. Surprisingly, given the oral application of colostrum to babies via buccal mucosa in the first few days of life, only weak similarity was observed between breast milk and oral taxa abundances, or oral and endotracheal taxon abundances. Relative abundances of bacterial taxa in anatomically close sites are most similar.
All correlations proved significant ($p < .001$). F-statistic and adjusted R^2 values are stated for each correlation plotted. (BM = Breast Milk; Endo = Endotracheal). R^2 values are adjusted to penalise for data points which do not positively influence predictive capacity of the model.