Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens

Richard M. Leggett1,8*, Cristina Alcon-Giner2,8, Darren Heavens1, Shabhonam Caim2, Thomas C. Brook3, Magdalena Kujawska2, Samuel Martin1, Ned Peel1, Holly Acford-Palmer2, Lesley Hoyles4, Paul Clarke5,6, Lindsay J. Hall2,8* and Matthew D. Clark1,7*

The MinION sequencing platform offers near real-time analysis of DNA sequence; this makes the tool attractive for deployment in fieldwork or clinical settings. We used the MinION platform coupled to the NanoOK RT software package to perform shotgun metagenomic sequencing and profile mock communities and faecal samples from healthy and ill preterm infants. Using Nanopore data, we reliably classified a 20-species mock community and captured the diversity of the immature gut microbiota over time and in response to interventions such as probiotic supplementation, antibiotic treatment or episodes of suspected sepsis. We also performed rapid real-time runs to assess gut-associated microbial communities in critically ill and healthy infants, facilitated by NanoOK RT software package, which analysed sequences as they were generated. Our pipeline reliably identified pathogenic bacteria (that is, Klebsiella pneumoniae and Enterobacter cloacae) and their corresponding antimicrobial resistance gene profiles within as little as 1 h of sequencing. Results were confirmed using pathogen isolation, whole-genome sequencing and antibiotic susceptibility testing, as well as mock communities and clinical samples with known antimicrobial resistance genes. Our results demonstrate that MinION (including cost-effective Flongle flow cells) with NanoOK RT can process metagenomic samples to a rich dataset in < 5 h, which creates a platform for future studies aimed at developing these tools and approaches in clinical settings with a focus on providing tailored patient antimicrobial treatment options.

Next-generation sequencing (NGS) has revolutionized the profiling of environmental and clinical microbial communities. The culture-independent, sensitive, data-rich nature of metagenomic sequencing, combined with powerful bioinformatics tools, has allowed researchers to differentiate patient groups from healthy individuals based on their microbial profiles1–7, including those with increased risk of pathogen overgrowth. Metagenomics also allows the identification of functional traits, for example, antibiotic resistance genes, which are important in light of the antimicrobial resistance (AMR) threat8–10. Optimization of metagenomic methodologies and bioinformatics tools could allow the identification of at-risk individuals, profiling of infectious agents and tailoring of treatments11.

In contrast to many NGS platforms, which require large capital investments and numerous samples to be multiplexed, newer sequencing platforms such as the MinION by Oxford Nanopore Technologies (ONT) represent inexpensive portable sequencing devices capable of producing long reads12. The real-time nature of data generation could provide users with a rapid screening platform; however, this real-time functionality and a different error profile require development of methods and bioinformatics pipelines, particularly for the clinical arena. Despite technical challenges in metagenomic profiling and diagnostics13, MinIONs have been successfully used in medical research on low-complexity samples including: outbreak surveillance14; characterization of bacterial isolates15; and low microbial biomass samples16–17. Diagnostics in metagenomic samples is still challenging due to lower MinION yields and accuracy, but essential since many clinical samples are complex. To date no studies have explored MinION technology in clinical gut metagenomic samples. For such applications, it is important to confidently identify (1) species-level profiles, (2) species abundance within the microbiota and (3) AMR gene repertoires. The development of a software tool, NanoOK RT, allowed us to perform real-time analysis and benchmark MinION-based metagenomics using mock communities and clinical samples from healthy and ill preterm infants. These studies allowed us to determine longitudinal microbiota profiles, gut-associated pathogens linked with sepsis or necrotizing enterocolitis (NEC) and their AMR profiles.

Results

Accurate classification of a microbial mock community using MinION sequencing. We benchmarked MinION technology by profiling a bacterial mock community using R7.3 flow cells. Reads were analysed with NanoOK18 and produced alignments to the 20 microbial reference sequences with 82–89% identity19. Coverage ranged from almost 0 × (8 reads) of Actinomyces odontolyticus to 13 × (7,695 reads) of Streptococcus mutans, which is consistent with expected mock concentrations (Supplementary Table 1). Benchmarking to Illumina sequencing demonstrated high correlation with expected proportions (Fig. 1a, log-transformed Pearson’s r = 0.94 for MinION and 0.97 for Illumina), and with each other (log-transformed Pearson’s r = 0.98). Broadly similar abundance levels across both platforms were observed, with some differences in assignment to species versus genus/family (Fig. 1b). This is probable since the longer length Nanopore reads should provide...
better specificity; however, in some cases lower Nanopore per base accuracy may reduce the ability to discriminate between closely related species.

Monitoring microbial disturbances in the preterm gut microbiota using MinION. We next tested if MinION technology could be used for real metagenomic samples, profiling eight preterm infants (three healthy and five diagnosed with suspected sepsis or NEC; Supplementary Figs. 1 and 2 and Fig. 2a). These infants are born with underdeveloped gut physiology and immunity, and have an altered gut microbiota; this increases the risk of life-threatening infections. Principal coordinates analysis (PCoA) of faecal samples indicated three distinct clusters, driven by the presence of either beneficial *Bifidobacterium breve* or potentially pathogenic microbiota members *Enterobacter cloacae* or *Klebsiella pneumoniae* (Extended Data Fig. 1).

We carried out longitudinal profiling of a preterm infant patient (P10) at days 13, 28 and 64 after birth (Fig. 2a). Comparing MinION (R7.3) to Illumina shotgun metagenomics sequencing confirmed that MinION sequencing depth was sufficient to capture the complete species diversity of the samples (Extended Data Fig. 2 and Supplementary Fig. 3). Taxonomic assignments using...
MinION versus Illumina shotgun data were comparable (species-level, log-transformed Pearson’s $r = 0.95$, $r = 0.90$ and $r = 0.94$ for P10N, P10R and P10V, respectively; Fig. 2b,c), for example, *Klebsiella, Enterobacter, Enterococcus, Veillonella, Staphylococcus* and *Bifidobacterium*, which correlated to probiotic supplementation or suspected sepsis periods. These data highlight the potential for MinION shotgun metagenomics to confirm the impact of interventions (for example, probiotic supplementation) and profile potential pathogenic microbes.

Antibiotics can lead to disruption of the gut microbiota and create a selection pressure that may change the profile of AMR genes (the resistome)\(^2\). We determined AMR profiles by comparing MinION to Illumina results. To avoid overcalling numerous subtypes of resistance genes (due to the higher error rates of Nanopore sequencing), we grouped together those detected genes that shared sequence similarity (see Methods for details). Classifying AMR genes by mode of action indicated comparable detection efficiency of MinION and Illumina. However, since Illumina datasets were capped at 1 million reads, whereas MinION datasets ranged from 48,000 to 83,000, three low-abundance genes/groups with unique resistance mechanisms (*bacA, sat4* and group *mph2*) were only detected by the deeper Illumina sequencing (Extended Data Fig. 3).
Overall, four AMR classes—efflux pumps, β-lactamases, aminoglycosides and fluoroquinolones—were particularly prevalent (Supplementary Fig. 4), with MinION technology able to detect species-specific AMR genes, for example, _ileS_ encoding mupirocin resistance in _Bifidobacterium_ or _fosA2_ (ref. 25) encoding fosfomycin resistance in _E. cloacae_ (Supplementary Table 2).

Further sequencing and analysis (using newer R9.4 flow cells) also allowed accurate taxonomic profiling: the gut microbiota of healthy infants P106 and P116 were dominated with _B. breve_ and _Bifidobacterium bifidum_ (Extended Data Fig. 4a,b), with a correspondingly limited resistome, consistent with beneficial taxonomic profiles and short antibiotic treatments (Extended Data Fig. 4c,d). These data indicate that MinION technology can profile preterm gut metagenomic samples, including determination of known species (that is, _B. bifidum_) and AMR profiles.

Bioinformatics tools use MinION-specific features to rapidly characterize gut-associated bacteria and antibiotic resistance profiles. MinIONs provide near-real-time sequencing and longer reads than Illumina sequencing, but the available software must take advantage of these useful features. To improve speed and incorporate bespoke analyses, we added real-time functionality to the NanoOK software (v0.95)34 thereby creating NanoOK RT, which aligns reads to bacterial and AMR databases as they are generated.

To test these tools, we profiled samples from preterm infants who were clinically diagnosed with NEC26 (Supplementary Fig. 2a,b). Samples from infants P49 and P205 both contained high proportions of _E. cloacae_ (Fig. 3a,b), with the correlation plots of normalized reads assigned at 1 and 6 h being almost identical (log-transformed Pearson’s _r_ = 0.97 for P49 and _r_ = 0.98 for P205; Fig. 3c,d). Resistance analysis highlighted a substantial number of AMR genes and classes (that is, efflux pump and β-lactamases), which were detected within minutes of the start of sequencing (Fig. 3e,f). Although these infants had gut microbiota dominated by _E. cloacae_, they also harboured other potentially pathogenic bacteria, highlighting the importance of determining which bacteria are harbouring AMR genes if these approaches are to be developed for more clinically based analysis.

Since MinION reads are typically longer than Illumina reads, we reasoned that we could extract additional information by examining the flanking sequences either side of each AMR hit that were independent (defined as ≥50 bp). Using this walkout approach in the NanoOK RT tool, we determined that for infant P205 the vast majority of AMR genes mapped back to _E. cloacae_ (87%; Fig. 3h). Contrastingly, although infant P49 had similar levels of _E. cloacae_, only 54% of AMR hits were associated with _E. cloacae_ (and a further 15% to its order Enterobacterales), with (low-abundance) _Klebsiella_ containing a range of AMR genes, for example, _OXA-2_ (β-lactamases) and _patA_ (efflux pump), constituting 23% of total AMR genes present (Fig. 3g and Supplementary Table 4). These data highlight that MinION sequencing coupled with the NanoOK Reporter analysis software can potentially map AMR genes to specific bacteria.

Next we performed a real-time run to evaluate how rapidly MinION plus NanoOK RT could detect potential pathogens and their corresponding AMR profiles in preterm infant P8 (diagnosed with suspected NEC and treated with multiple antibiotics). Current rapid clinical microbiology tests, including determination of antibiotic susceptibility, take between 36 and 48 h. Our real-time run (from sample preparation to analysis) identified pathogens and resistances in approximately 5 h (Supplementary Fig. 2c and Fig. 4a).

Reads were analysed using NanoOK RT, with the first 500 reads indicating a dominance (332 reads) of _K. pneumoniae_ (a potential causative organism that has been associated with NEC)27. By 1 h after sequencing started (5 h total), the pipeline had analysed 20,000 reads with _K. pneumoniae_ accounting for approximately 70% of reads. Further analysis at 6 h showed no significant differences (Fig. 4b,d, log-transformed Pearson’s _r_ = 0.97) and was validated by Illumina sequencing (Extended Data Fig. 5a,b).

Our real-time run also indicated that we could rapidly (1 h after sequencing started) map AMR genes/groups (Fig. 4e) including fosfomycin, aminoglycoside and fluoroquinolone resistance, β-lactamases and efflux pumps. We detected _K. pneumoniae_-specific SHV variants28 as early as 38 min (at 13,000 reads, 4.38 min total time), whereas lower-abundance AMR genes in the sample, for example, those conferring tetracycline resistance, were not detected until 2 h post-sequencing (6 h total).

NanoOK Reporter AMR walkout analysis indicated that the majority of AMR genes within the sample were assigned to _K. pneumoniae_ (approximately 51%) or by the lowest common ancestor algorithm (Methods) to within its Enterobacteriaceae family (approximately 24%) (Fig. 4c), including efflux pumps _qoxB_ (group _qox-mex-amr1_), _mdtC_ (conferring multidrug resistance, group _mdt-mex-sm1_), _patA_ (resistance to fluoroquinolone) and _FosA5_ (resistance to fosfomycin) (Supplementary Table 4).

Whole-genome sequencing (WGS) analysis and phenotypic assays indicate the robustness of NanoOK RT walkout analysis. To validate the genotypes obtained from our real-time MinION run, 8 _K. pneumoniae_ isolates from P8 were obtained (the 16S rDNA gene sequence alignment indicated similarity levels ranging from 99.8 to 100%; Supplementary Table 5), with whole-genome shotgun sequencing and assembly on one _K. pneumoniae_ isolate performed using Illumina and Nanopore technologies. The longer Nanopore reads produced a single contig of 5.47 Mb and two further contigs of 0.37 Mb, while Illumina produced 69 contigs totalling 5.73 Mb. Many of the AMR genes/groups detected in the walkout analysis from the metagenomic sample P8 correlated with both the Illumina and MinION isolate data (Extended Data Fig. 6). A significant proportion (approximately 60%) of the resistance genes/groups in the metagenomics walkout and the WGS isolate correlated with the efflux pumps (for example, groups _mdt-mds-acr-mtr, mdt-mex-sne, mdt-acr_ and _qox-mex-amr_), while other hits correlated to known _K. pneumoniae_ AMR genes/groups including β-lactamases (for example, the _SHV-LEN-OKP_ group) or fosfomycin resistance (group _Fos3_).

To confirm these genomic AMR profiles, we carried out antibiotic phenotyping on three preterm bacterial isolates: two pathogenic (P8 _K. pneumoniae_ and P49 _E. cloacae_); and one beneficial (P103 _B. bifidum_). The _P8 K. pneumoniae_ isolate was tested against the seven most commonly used antibiotics in neonatal intensive care units (Supplementary Table 10), with the isolate found to have higher minimum inhibitory concentration (MIC) breakpoint values than those put forward by the European Committee on Antimicrobial Susceptibility Testing29 for previously prescribed antibiotics, that is, benzylpenicillin, amoxicillin and gentamicin. In contrast, the only MIC breakpoint value lower than the European Committee on Antimicrobial Susceptibility Testing was for cefotaxime, an antibiotic not prescribed. These data correlate with the AMR data generated by the NanoOK Reporter and walkout analysis (Extended Data Fig. 6). Phenotypic testing for _P49 E. cloacae_ indicated resistance to gentamicin and benzylpenicillin (Supplementary Table 10), correlating with prescribed antibiotics (Supplementary Fig. 2a), and the AMR genes detected by our ‘walkout’ analysis: the _ACT_ (resistance to benzylpenicillin) and _acrB_ genes (resistance to gentamicin) (Supplementary Table 4). P103 _B. bifidum_ showed resistance towards mupirocin (Supplementary Table 10), in agreement with the detection of the ileS gene (Supplementary Table 4) from the walkout analysis.
Fig. 3 | Rapid diagnostic using MinION technology for preterm infants clinically diagnosed with suspected NEC (P49 and P205). a, b, Taxonomic profiles comparing the results obtained at 1 h and 6 h for P49 (a) and P205 (b) after sequencing started. The pie chart legends comprise the eight most abundant taxa. Detailed counts can be found in Supplementary Table 3. c, d, Correlation plots representing normalized assigned reads at 1 h and 6 h for P49 (taxa n = 35, log-transformed Pearson’s r = 0.97) (c) and P205 (taxa n = 120, log-transformed Pearson’s r = 0.98) (d). The grey region either side of the fit line represents the 95% CIs. e, f, Heat maps displaying the number of CARD hits detected among the most common groups of antibiotic resistance genes found in preterm P49 (e) and P205 (f). Further information on all the AMR genes classified can be found in Supplementary Table 4. g, h, Walkout results for preterm infants P49 (g) and P205 (h) at 6 h, as reported by the NanoOK RT’s walkout option. Results shown are for independent bacterial hits (defined as ≥ 50 bp away from the AMR sequence), at 6 h of sequencing.
Further enhancements to the Nanopore sequencing technology. ONT now produce a rapid library kit that requires as little as 10 min preparation time. Profiling the gut microbiota of healthy infant P103 produced 1.2 million reads (read N50 of 1,957 bp), with a sample-to-analysis time around 60 min faster than our one-dimensional (1D) real-time run on infant P8. We confirmed dominance of commensal *Bifidobacterium* species, including *B. bifidum* (also probiotic species; Extended Data Fig. 7a,b), with NanoOK RT AMR profiling indicating a high proportion of mupirocin and tetracycline resistance (Extended Data Fig. 7c).

We performed a reference-guided assembly of the *B. bifidum* genome, which resulted in 3 contigs with an average identity of 98.86% (Supplementary Fig. 5a,b). A de novo assembly generated 24 contigs mapping to 1.7 Mb of the 2.2 Mb reference with an average identity of 98.64%, demonstrating the potential to resolve whole microbial genomes from metagenomic samples, although the error rate is currently high making SNP analysis, and therefore strain level profiling, challenging.

The Flongle flow cell adaptor (ONT) is another recent enhance-ment that facilitates the use of cheaper (approximately US$90) flow...
cells. Using Flongle flow cells on the MinION and GridION, we evaluated P129 (Supplementary Fig. 2d) and confirmed a dominance of potentially pathogenic Enterococcus faecalis (Extended Data Fig. 8a,b), as well as a diverse resistome conferring resistance to this infant's antibiotic treatment, that is, group AAC-APH genes (gentamicin resistance) and the PCI gene (benzylpenicillin resistance). Taxonomic and AMR profiles obtained for the MinION or GridION Flongle datasets were comparable (log-transformed Pearson's $r=0.92$ at the species level; Extended Data Fig. 8c,d).

**Benchmarking and validation of MinION and NanoOK RT using mock resistome samples.** The data presented so far indicates that we can detect AMR genes using MinION sequencing and bioinformatic tools. However, confirming the robustness and validity of these approaches is important for next-stage clinical studies. Thus, we analysed a mock AMR barcoded seven-strain community, spiking this with the P8 K. pneumoniae isolate. Analysis indicated that a significant proportion of AMR genes detected in the spiked mock community corresponded to the WGS isolate data, including group SHV-LEN-OKP (resistant to β-lactam antibiotics) and group mdt-mex-sme (efflux pumps) (Fig. 5a). Some genes in the isolate assemblies were not present in the mock community, probably a consequence of low sequence coverage. By barcoding the mock constituent species, we validated NanoOK RT's walkout decisions; 97 genes out of 107 were correctly assigned (Supplementary Table 9). Of those incorrectly assigned, five were assigned within the same genus and three appear to be due to barcodes that were wrongly identified by the ŌNT software (typically due to sequence error), thus independently of the walkout strategy. For the K. pneumoniae spike, 34 out of 35 genes were correctly assigned to species or higher taxa, the remaining gene suffering a misassigned barcode. We also spiked a metagenomic DNA sample (healthy preterm infant P103) with two different P8 K. pneumoniae sequenced isolate DNA concentrations (4 and 40%) to test sensitivity and specificity (Fig. 5b). The majority (22 out of 31) of AMR genes were detected at both concentrations, although the mdt-mds-acr-mtr group, mdrB, patA, acr-sme group, mdt-mex-sme group and ERM-7 group were only detected in the P103M 40% spike mock. Notably, reads assigned to P8 K. pneumoniae in P103M 40% were approximately 10× higher than the lower spiked mock, P103M 4% (Fig. 5c).

**Discussion**

With worldwide concerns about increasing AMR rates, there is a pressing need for optimized and rapid metagenomic sequencing platforms and bioinformatic tools that could be used to gather clinically relevant data. In this study, we used a combination of improved Nanopore sequencing chemistries and our own open source analysis packages to successfully profile mock and clinical metagenomes. MinION sequencing data were comparable in discriminatory power to Illumina sequencing data, allowing profiling and abundance of microbial species, community resistome profiling and species-specific antibiotic resistance profiles, which were benchmarked using mock communities and phenotypic testing.

Initial mock community profiling confirmed the MinION was a suitable tool (comparable to Illumina) for metagenome profiling 80, which we extended to preterm gut microbiota profiling, thereby identifying a supplemented probiotic species (that is, B. bifidum; Fig. 2c) and E. cloacae, a known sepsis pathogen 81. Furthermore, MinION and Illumina data indicated highly comparable AMR resistome profiles—low numbers of AMR groups within healthy Bifidobacterium-dominated preterm infants—whereas a larger AMR gene repertoire was present in the gut microbiota of infants dominated by Klebsiella and Enterobacter.

With the worldwide AMR threat, metagenomic profiling for resistance genes in a timely and accurate manner could be used in critical care settings. Notably, MinION- and Illumina-generated reads mapped to genes with similar antibiotic resistance mechanisms (Extended Data Fig. 3), including β-lactamase and aminoglycoside genes (conferring resistance to benzylpenicillin and gentamicin, respectively), and only 3 unique resistance mechanisms (bacA, sat4 and the mph-2 group) of all 70 AMR genes/groups were exclusively detected by Illumina sequencing. This result may be due to the lower MinION read count and might be mitigated by ongoing improvements in MinION technology. Because grouping of genes is based on sequence identity, this approach may not allow differentiation between grouped genes that in fact have different resistance mechanisms despite their sequence similarity. These caveats are important within a clinical context and further studies are required to understand these subtle differences in light of the potential limitations of Nanopore sequencing sensitivity.

Our NanoOK RT software allowed in-depth analysis of species abundance and antibiotic resistance genes in ill infants (P49 and P205). These preterm infants had high levels of E. cloacae and a significant resistome (AMR genes including ACT-27 mapping directly to E. cloacae; Supplementary Table 4), which may correlate with the clinical diagnosis of suspected NEC. Our software indicated specific taxa harbouring AMR genes, for example, gene ACT-27 mapping to E. cloacae (Supplementary Table 4). Notably, performing a walkout, rather than de novo metagenomic assembly, requires less computing time and therefore represents a faster method of characterizing potential multidrug-resistant pathogens. However, we also used MinION metagenomic data to assemble B. bifidum (P103) using a reference-guided approach and a more challenging de novo assembly, highlighting how more in-depth genomic follow-up studies can be performed from these data.

Next we sought to understand how rapidly we could determine microbial identification and corresponding AMR profiles by mimicking a more clinically relevant diagnostic approach by performing a real-time run using samples from an extremely ill preterm infant (P8) who had received multiple antibiotic courses since birth (46 d antibiotic treatment out of 63 d of life at sample collection). MinION sequencing generated high yields and revealed a K. pneumoniae-dominated profile after just 1 h of sequencing, which may link with the clinical NEC diagnosis since intestinal overgrowth of this pathogen can induce pathological inflammatory cascades 82. Profiling of additional and more complex samples from infants diagnosed with NEC (that is, P49 and P205) indicated distinct and differential microbiota profiles (when compared to P8) also 1 h after the start of sequencing (Fig. 3a,b). Real-time analysis of MinION data using NanoOK RT highlighted the presence of a significant resistome just 10 min after the start of sequencing, including β-lactamases, aminoglycoside resistance genes and multidrug efflux pumps, with greater sequencing depth correlating with higher numbers of AMR genes (Fig. 4e).

*Klebsiella* is of particular AMR concern due to the increasing emergence of multidrug-resistant isolates that cause severe infection and represent a real threat to patient outcomes 83. Benchmarking with WGS (Illumina and MinION) indicated broad agreement with AMR profiles from the MinION metagenomic run, although we noted a slightly expanded AMR profile at 6 h with the walkout analysis (Extended Data Fig. 5). These differences may correlate with intra-infant strain level variation; thus, single-isolate WGS analysis would not capture the wider AMR repertoire. However, further work is required to determine the utility of strain level analysis, including the development of a standardized framework determining the parameters for single-nucleotide polymorphism analysis, to compensate for the lower read accuracy observed in MinION data, and requiring substantial additional experimental validation. When subjecting strains to MIC testing (the current gold standard for profiling AMR), we observed phenotypic resistance to all main groups of antibiotics that had been prescribed to infant P8, with strong association between AMR gene detection and MIC testing.
for example, SHV and β-lactam antibiotics, and *oqxB* genes and gentamicin, thereby suggesting that MinION could be useful for rapid AMR profiling.

MIC phenotypic testing on preterm-associated *E. cloacae* and *B. bifidum* isolates agreed with our walkout analysis, with mock community experiments also providing the expected AMR profiles. However, if a potential pathogen is present at low levels within the total microbiome, ability to detect its AMR genes may be reduced. (This is potentially solvable using greater sequencing depth.) From a clinical standpoint, infection is typically associated with pathogen overgrowth; thus, these mock experiments provide strong indications that the MinION and NanoOK RT combination may provide robust antibiotic resistance data. Further (multicentre cohort) clinical studies are required to establish the accuracy of Nanopore/NanoOK methods before they could be considered as clinical diagnostic tools.

Rapid profiling and portability is crucial within clinical and fieldwork settings; however, current standard (large footprint) diagnostic tools.
NGS platforms (Illumina and PacBio) often take >10–40 h to run (excluding analysis). We obtained MinION bioinformatics results within 1 h of sequencing (5 h total time), with the recent rapid kit being even quicker and the Flongle representing a more cost-effective approach. However, the accuracy of Nanopore reads still lags behind short-read platforms, which necessitate the use of both lower BLAST thresholds and AMR gene groupings. As Nanopore read accuracy continues to reach that of short reads, this will no longer be necessary. The longer length of Nanopore reads results in longer (more significant) alignments, but further optimization and validation, including using standard clinical microbiology testing, is required for refinement and the development of clinical management of patients.

**Conclusion**

MinION technology in conjunction with NanoOK RT analysis represents a platform for rapid profiling of gut-associated bacterial species including potential pathogens and corresponding AMR profiles. The accuracy of this approach was confirmed by comparison to Illumina metagenomic sequencing, characterization of patient-derived bacterial isolates, including WGS and phenotypic (that is, MIC) testing, and using mock communities with known AMR profiles. Together these analyses and approaches may prove useful in healthcare settings, particularly with regard to resistome analysis and antibiotic stewardship interventions in the future.

**Methods**

**Mock community benchmarking.** DNA. We used genomic DNA from a microbial mock community used in the Human Microbiome Project (HM-277D; BEI Resources). This mock community contains a mixture of 20 bacterial strains containing staggered RNA operon counts. Details of the strains present in the community are indicated in Supplementary Table 1.

**Illumina sequencing of mock community.** Illumina-compatible, amplification-free, paired-end libraries were constructed with inserts spanning from 600 to >1,000 base pair (bp). A total of 600 ng of DNA was sheared in a 60 µl volume on a Covaris S2 for 1 cycle of 40 s with a duty cycle of 5%, cycles per burst of 200 and an intensity of 3. Fragmented DNA was then end-repaired using the NEBNext End Repair Module (New England Biolabs), size-selected with a 0.58 × Hi Prep bead clean-up (GE Biotech) and followed by A-tailing using the NEBNext dA-Tailing Module (New England Biolabs) and ligation of adaptors using the Blunt/TA Ligase Master Mix (New England Biolabs). Three 1x bead clean-ups were then undertaken to remove all traces of adaptor dimers. Library quality control was performed by running an Agilent Bioanalyzer High Sensitivity Chip and quantified using the KAPA Library Quantification Kit for Illumina Platforms (KAPA Biosystems). Based on the quantitative PCR quantification, libraries were loaded at 9 pmol on an Illumina MiSeq System and sequenced with 300 bp paired reads.

**MinION sequencing of mock community.** MinION two-dimensional (2D) libraries were constructed targeting inserts >8 kba-base pair (kbp). A total of 1 µg of DNA was fragmented in a 46 µl volume in a g-TUBE (Covaris) at 6,000 rpm in an Eppendorf 5417C centrifuge. Sheared DNA was then subjected to a repair step using the NEBNext FFPE Repair Mix (New England Biolabs) and purified with a 1x Hi Prep bead clean-up (GE Biotech). A DNA control was added to the repaired DNA and then end-repaired and A-tailed using the NEBNext Ultra II End Repair/ dA-Tailing Module (New England Biolabs), and purified with a 1x Hi Prep bead clean-up; then the AMX and HPA MinION adaptors were ligated using the Blunt/TA Ligase Master Mix (New England Biolabs). An HP tether was then added and incubated for 10 min at room temperature followed by a further 10 min room temperature incubation with an equal volume of pre-washed MyOne Streptavidin C1 beads (Thermo Fisher Scientific). The library-bound beads were washed twice with bead binding buffer (ONT) before the final library was eluted via a 10 min incubation at 37 °C in the presence of the MinION elution buffer. The final library was then mixed with running buffer, fuel mix and nuclease-free water and loaded onto an R7.3 flow cell according to the manufacturer’s instructions; sequencing data were collected for 48h.

**Mock community data analysis.** MinION reads were basecalled using the Metrichor service (https://metrichor.com/) and downloaded as FAST5 files. NanoOK v0.54 (ref. 34) was used to extract the FASTA files, align (via the LAST alignment (752)) against a reference database of the 20 genomes and generate an analysis report (Supplementary Note 1). Quality control of the Illumina data was carried out with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to ensure read quality was within the expected bounds. This demonstrated a mean quality control of 30 up to base 250. We subsampled a random set of 1,000,000 reads (subsample.pl script; https://github.com/richardmleggett/scripts) to represent the yield of a MiSeq nano flow cell and ran Trimmmorphic (v.0.30)34 to remove remaining adaptor content and apply a sliding window quality filter (size 4, mean quality ≥15). Illumina and MinION reads were then BLASTed (BLASTv+ v2.2.29, maximum e-value 10^-10) against the National Center for Biotechnology Information (NCBI) nucleotide database and the results were imported into MEGAN6 (ref. 35) for taxonomic analysis. In a separate analysis, the reads were mapped against references using minimap2 v2.17-1943 (ref. 36) and alignments processed using the bamstats.py script (https://github.com/guiojob/bamstats). Another script (parse_bamstats.pl, https://github.com/richardmleggett/bambl) then totalled the read mapping to each reference and these counts were imported into Microsoft Excel where they were normalized to relative abundances. Counts were log-transformed and the log-transformed Pearson’s coefficient was computed in Microsoft Excel using the PEARSON function. As Nanopore lags behind short-read platforms, which necessitate the use of both lower BLAST thresholds and AMR gene groupings. As Nanopore read accuracy continues to reach that of short reads, this will no longer be necessary. The longer length of Nanopore reads results in longer (more significant) alignments, but further optimization and validation, including using standard clinical microbiology testing, is required for refinement and the development of clinical management of patients.

DNA extraction from faeces samples (preterm infants). Bacterial DNA was extracted using the FastDNA Spin Kit for Soil (MP Biomedicals) according to the manufacturer’s instructions but extending the bead-beating step to 1 min and eluting the DNA with 55°C DNA Elution Solution. The starting material used to extract DNA was between 100 and 150 mg. DNA purity and concentration were assessed using a NanoDrop 2000c Spectrophotometer (Thermo Fisher Scientific) and Qubit 2.0 Fluorometer (Thermo Fisher Scientific). Samples with DNA concentrations higher than 25 ng/µl were considered acceptable.

MinION shotgun library preparation. MinION 2D libraries were constructed as outlined for the mock community (see earlier) except that, for the R9.4 flow cells, the final library was mixed with running buffer containing fuel mix, library loading beads and nuclease-free water and loaded onto the flow cell according to the manufacturer’s instructions. MinION 1D ligation libraries were constructed using 1 µg unfragmented DNA. This was subjected to a repair step using the NEBNext FFPE DNA Repair Mix (New England Biolabs) and purified with a 1x Hi Prep bead clean-up (GC Biotech). A DNA control was added to the repaired DNA and then end-repaired and A-tailed using the NEBNext Ultra II End Repair/dA-Tailing Module (New England Biolabs), and purified with a 1x Hi Prep bead clean-up. The ONT Adapter Mix MinION adaptors were then ligated using the Blunt/TA Ligase Master Mix (New England Biolabs). The library was eluted via a 10 min incubation at 37 °C in the presence of the MinION elution buffer. The final library was then mixed with running buffer, fuel mix and nuclease-free water and loaded onto a flow cell according to the manufacturer’s instructions. MinION 1D rapid libraries were prepared by incubating 200 ng of DNA with 2.5 µl FRM buffer for 1 min at 30 °C then 1 min at 75 °C, followed by adding 1 µl Rapid Adapters and incubating at room temperature for 5 min. The final library was then mixed with running buffer, fuel mix and nuclease-free water and loaded onto the flow cell according to the manufacturer’s instructions. Further details on the genomic sequencing kits and samples used in this study can be found in Extended Data Fig. 2.

**Illumina HiSeq 2500 shotgun library preparation.** Libraries for the samples (P10N, P10R and P10V) were prepared using the TrueSeq Nano DNA Library Prep Kit (Illumina) according to the manufacturer’s instructions and sequenced with the HiSeq 2500 System with 150 bp paired-end reads. The library was prepared as for the amplification-free library for the mock community (see earlier) and run at 9 pmol on an Illumina MiSeq System with a 2 × 250 bp read metric.

**Time series study for infant P10.** The Illumina and MinION sequencing data for samples P10N, P10V and P10R from infant P10 were studied. For the Illumina sample, the removed PCR data were downloaded from the raw sequencing data (https://github.com/richardmleggett/scripts), ran Trimmmorphic to remove the adaptors and applied a sliding window quality filter (size 4, mean quality ≥15) and then randomly subsampled 1 million reads (subsample.pl script; https://github/
NCBI nucleotide database (maximum e-value 10−10). We further validated this choice by sequencing a microbial mock community with a known AMR profile and investigating the effect of varying the minimum identity, but this coincides with the lowest true positive (117) and highest false negative (6) rates, while the false positive rate remained low (7). The lowest false negative rate (0) occurs at 100% identity, but this coincides with the lowest true positive (0) and highest false negative (123) rates.

**AMR gene grouping.** Because of the higher error rate of Nanopore sequencing compared to the Illumina platforms, we were not confident that genes with low coverage and high sequence similarity could be differentiated unambiguously. To address this, we wrote a Python script, AMR_gene_grouper.py, which groups similar genes in the CARD according to sequence similarity. This script can be found on the Github repository mentioned in the original text. To test the script, we used all pass reads (83,000 reads for P10N, 48,000 for P10V and 53,000 for P10V); for the barcoded mock community, we used all 90,000 (plot_card_heatmap.R; https://github.com/richardmleggett/bambi) takes this file and generates a text file giving per-read analysis. The walkout approach taken by MEGAN, we considered any hits with a bit score of at least 90% of the highest scoring hit. Results are displayed on a taxonomy tree, donut plot and generate a final file summarizing hits per group at each time point. An R script saves a text file at each time point (in this study, batches of 500 timestamped results are in, using their timestamps to indicate when a result is first obtained.

**Generation of AMR gene heat maps.** We opened the CARD results using NanoOK Reporter and used the option to save summary data as a plain text file. This saves a text file at each time point (in this study, batches of 500 timestamped reads) summarizing the counts of resistance genes identified up to that point (plot_card_heatmap.R) for each of these, writing them to a file in the folder ‘nanook_reporter_files’. CARD hits were only considered for this analysis if they possessed an e-value <1e-001, a sequence identity ≥80% and a length ≥200 bp. We took the latest time point file that the heat map was to show (for example, 6 h) and extracted a list of the antibiotic resistance ontology numbers from the 1D column. Each unique antibiotic resistance ontology number was manually assigned to its corresponding antibiotic resistance group according to the classification given by CARD. We wrote a script (gather_heatmap_data.pl; available at https://github.com/richardmleggett/bambi) to take the summary files, together with this mapping, and generate a final file summarizing hits per group at each time point. An R script (plot_card_heatmap.R: https://github.com/richardmleggett/bambi) takes this file and produces the heat map.

**Statistical analysis.** Read counts at different stages of the bioinformatics analysis are provided in Extended Data Fig. 2. For comparative analysis, MEGAN6 was set to subsample reads down to the read count of the sample with the lowest number of reads. For Pearson’s correlation comparisons of taxonomic profiles (for example,
at 1 and 6 h), the two samples to be compared were loaded into MEGAN6; its comparison function was used to display both on the same tree. MEGAN6 was set to display genus (Fig. 1) or species level (remaining figures); all nodes were selected and the assigned read counts were exported to a single CSV file. The CSV file was imported into Microsoft Excel, relative abundances were calculated and log-transformed, and the PEARSON function was used to calculate Pearson’s r from the log-transformed data. The Microsoft Excel data were exported to a tab-separated file and plots were produced using R (plot_correlation.R; https://github.com/richardmleggett/bambi).

Isolation and biochemical characterization of P8 K. pneumoniae strains. An aliquot (100 mg) of faecal sample was homogenized in 1 ml TBT buffer (100 mM Tris/HCl, pH 8.0, 100 mM NaCl, 10 mM MgCl₂) by pipetting and plate-mixing at 1,500 r.p.m. for 1 h. Homogenates were serially diluted to 10⁻⁴ in TBT buffer. Aliquots of 50 µl were spread over 385 K. pneumoniae agar plates (Oxoid) in triplicate and incubated aerobically at 37°C overnight.

Colonies were selectively screened for lactose-positive (that is, pink) colonies. One colony of each morphotype strain was restreaked on MacConkey agar three times to purify. Biochemical characterization was performed using API 20E tests (BioMérieux) according to the manufacturer’s instructions.

16S rRNA phylogenetic analysis of P8 K. pneumoniae isolates. Sequences of the 16S rRNA gene from nine K. pneumoniae isolates were prepared to perform the phylogenetic analysis. We extracted DNA using the FastDNA Spin Kit for Soil according to the manufacturer’s instructions (ref. 4); 150 ng of DNA was ligated with the Vent 96-Well Thermal Cycler (Applied Biosystems), master mix from Kapa2G Robust PCR reagents (Kapa Biosystems) and the following primers: D1 (forward, 5′-ACG GTT ACC TTG) and D2 (reverse, 5′-AGA GTT TCA TGG CTC AG-3′); D1 (forward, 5′-AGA GTT TCA TGG CTC AG-3′) and rP1 (reverse, 5′-AGG ACC TTG ACC TTA CGA CTT-3′) (ref. 4). PCR amplification conditions were: 1 cycle at 94°C for 5 min, followed by 35 cycles at 94°C for 1 min, 45°C for 1 min and 72°C for 2 min followed by a final strand extension at 72°C for 7 min. Amplicons were sequenced using an automated Sanger sequencing service (Eurofins Genomics). Partial 16S rRNA sequences (approximately 900 positions) of 9 isolates of K. pneumoniae, obtained using the automated Sanger sequencing service, were compared for similarity. Multiple sequence alignments were performed with the SILVA Incremental Aligner (v.1.211)4 and manually curated for quality. Nucleotides were coloured using BoxShade v.3.21 (http://www.ch.embnet.org/software/BOX_form.html). The similarity/identity matrix between sequences was calculated using MatGAT v.2.01 (Matrix Global Alignment Tool) using the BLOSUM 50 alignment matrix46.

Determination of MIC for P8 K. pneumoniae, P49 E. cloacae and P103 B. bifidum. Calculation of the antibiotic MIC was performed using the broth microdilution method47. Serial twofold dilution antibiotics (benzylpenicillin, gentamicin, vancomycin, metronidazole, meropenem, cefotaxime and mupirocin) were added to sterile broth microdilution plates. The MIC was calculated using MatGAT v.2.01 (Matrix Global Alignment Tool) using the BLOSUM 50 alignment matrix46.

DNA extraction from P8 K. pneumoniae isolate for WGS analysis. An overnight (10 ml) culture of the isolate was centrifuged at 4,000 r.p.m. for 10 min, resuspended in 30 ml of PBS (Sigma-Aldrich) and centrifuged again. The pellet was then resuspended in 2 ml of 25% sucrose (Thermo Fisher Scientific) in Tris-EDTA buffer (10 mM Tris (Thermo Fisher Scientific) and 1 mM EDTA at pH 8.0 (WVR Chemicals)); 50 µl of lysozyme (Roche Molecular Systems) at 100 mg ml⁻¹ in 0.25 M Tris, pH 8.0, was added. The mixture was incubated at 37°C for 1 h; 100 µl of Proteinase K at 20 mg ml⁻¹ (Roche Molecular Systems), 30 µl of RNase A at 10 mg ml⁻¹ (Roche Molecular Systems), 0.4 µl of 0.5 M EDTA, pH 8.0, and 250 µl of freshly prepared 10% Sarkosyl NL30 (Sigma-Aldrich) were added. The mixture was then mixed on ice for 2 h at 4°C, followed by a water bath at 50°C overnight. Next, E Buffer (10 mM Tris, pH 8.0) was added to the sample to a final volume of 5 ml, mixed with 5 ml phenol:chloroform:isoamyl alcohol and centrifuged for 10 min at 4,000 r.p.m. This procedure was repeated with a 5 min centrifugation time. Next, the aqueous phase was transferred into a new MaxAtract High Density Tube (QiAGEN) and centrifuged for 15 min at 4,000 r.p.m. The aqueous phase was transferred into a new MaxAtract High Density Tube, made up with E Buffer as necessary, mixed with 5 ml of phenol:chloroform:isoamyl alcohol and centrifuged for 10 min at 4,000 r.p.m. This procedure was repeated with a 5 min centrifugation time. Next, the aqueous phase was transferred into a MaxAtract High Density Tube made up to 5 ml with E Buffer as necessary, mixed with 5 ml of phenol:chloroform:isoamyl alcohol (2:4:1) (Sigma-Aldrich) and centrifuged for 5 min at 4,000 r.p.m. The chloroform:isoamyl alcohol step was repeated once more which the final aqueous phase was transferred into a sterile Corning 50 ml centrifuge tube and 2.5 volumes of ethanol (WVR Chemicals) were added. The sample was incubated for 15 min at −20°C, then centrifuged for 10 min at 4,000 r.p.m. and 4°C. Finally, the DNA pellet was washed with 10 ml of 70% ethanol and centrifuged at 4,000 r.p.m. for 10 min twice, dried overnight and resuspended in 300 µl of E buffer.

WGS library preparation and sequencing of P8 K. pneumoniae isolate. DNA samples containing 500 ng genomic DNA were analysed. DNA was sheared into fragments of 400–600 bp using a Covaris plate with glass well plates and Adaptive Focus Acoustic Fibres. Sheared DNA was purified by size, which was used as the input for the Nanopore RR analysis. Walkout analysis was performed by clicking on the ‘Walk’ icon in NanoKo Reporter; the resultant walkout_results.txt file was processed with a custom Perl script (parse_walkout_barcode.pl; https://github.com/richardmleggett/bambi), which looks up each read ID in the walkout to discover which barcode is associated with it (Supplementary Table 1). Annotated assemblies

Assembly of WGS isolate (P8 K. pneumoniae). Presence or absence of AMR genes was performed on one K. pneumoniae isolate from sample P8, benchmarking two different sequencing platforms: MinION and Illumina HiSeq 2500. Sequencing data from the MinION run was assembled with Canu v.1.5 (ref. 3) corrected with Reblock v.1.0 with nanopolish v.2.0. Sequencing data from the Illumina HiSeq 2500 run was assembled using Velvet (v.1.1)48. Gene presence/absence diagrams were generated as described earlier.

Preparation of in vitro mock resistome and clinical mock. DNA. We extracted DNA from 2 ml overnight cultures of 7 National Collection of Type Cultures bacterial samples and P8 K. pneumoniae isolate. The DNA extraction protocol followed the manufacturer’s instructions (MagAttract HMW DNA Kit; QIAGEN). Details of the strains present in this community are shown in Supplementary Table 1.

MinION sequencing of mock resistome. Nanopore 1D native barcoded libraries were constructed targeting inserts >8 kb using the ONT SQK-LSK109 and EXP-NBD104 kits based on the DNA concentration of the native barcode adaptor-ligated molecules. The mock resistome consisted of 12.5% of each of the 8 strains (for example, an even mock). A total of 1 µg of each DNA was fragmented in a 46 µl volume in a q-TUBE at 6,000 r.p.m. in an Eppendorf 5417 centrifuge. Sheared DNA was then subjected to a combined repair and A-tailing step using the FFPE DNA Repair Mix and NEBNext Ultra II End Repair/dA-Tailing Module and purified with a 1x KAPA bead (Roche Sequencing) clean-up. Repaired and A-tailed DNA had native barcode adaptors ligated using the Blunt/TA Ligase Master Mix further followed by a purification step with a 1x KAPA bead clean-up. To create an even abundance mock, 87.5 ng of each native barcode adapter-ligated molecules were pooled. AMXXI (ONT) adaptors were ligated to the two pooled mock samples using the Quuck T4 DNA ligase. The samples were then separated by size, washed twice with ONT’s long fragment buffer and then eluted in MinION elution buffer by incubating for 10 min at room temperature. The final library was mixed with the sequencing buffer and loading beads, and then loaded onto a FLO-MIN108D Flow Cell (R9.4.1) flow cell according to the manufacturer’s instructions; sequencing data were collected for 48 h.

Analysis of barcoded mock data. Barcoded reads from the mock data were basecalling with ONT’s Guppy v.2.3.1. From the pass reads, we took all reads ≤3,000 bp in length (to reflect the reduced read lengths probably found in real samples) and randomly sampled approximately 11,000 reads from each of the barcode files to make a singe random 10% of each file, which was used as the input for the NanokO RT analysis. Walkout analysis was performed by clicking on the ‘Walk’ icon in NanoKo Reporter; the resultant walkout_results.txt file was processed with a custom Perl script (parse_walkout_barcode.pl; https://github.com/richardmleggett/bambi), which looks up each read ID in the walkout to discover which barcode is associated with it (Supplementary Table 1). Annotated assemblies
of the mock constituents are available from the Public Health England reference collections at the Wellcome Sanger Institute (https://www.sanger.ac.uk/resources/downloads/bacteria/ntc/). We BLASTed these and the P8 K. pneumoniae isolate assembled against the CARD (as used previously), filtering for a maximum e-value of 0.001, minimum identity of 80% and minimum length of 200 bp, to determine the expected AMR profile. The amr analyser Java program described earlier was then used to create the presence/absence maps, comparing the profiles of the metagenomic sample with the genome assemblies.

**MiniON sequencing of clinical mock data.** We spiked a sample from healthy infant P103 with the DNA from the P8 K. pneumoniae isolate. DNA from P103 was run on a TapeStation 2100 (Agilent Technologies) to determine average molecule length. DNA from the P8 K. pneumoniae isolate was then fragmented to a similar length using a gTUBE. The DNA from P8 K. pneumoniae was spiked into the P103 sample targeting 10% and 50% of total DNA, based on concentration, and MiniON 1D libraries constructed using the ONT SQK-LSK109 Kit. Libraries were constructed as outlined in the manufacturer’s protocol, loaded onto a FLO-MIN106D Flow Cell (R9.4.1) flow cells and sequence data were collected for 48 h. Reads were basecalled with ONT’s Guppy and analysed after sequencing with NanoOK RT.

**Analysis of Flange flow cells.** For the GridION run, reads were basecalled live, then passed to NanoOK RT for analysis, as per the real-time diagnostic study (see earlier). For the MiniON run, reads were basecalled post-sequencing with Guppy, then passed to NanoOK RT for analysis.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

Sequence data (Illumina and MiniON) that support the findings of this study have been deposited with the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession no. PRJEB22207.

**Code availability**

NanoOK RT is available as an extension to NanoOK, selectable as a runtime option, from https://github.com/richardmleggett/NanoOK. NanoOKReporter is available from https://github.com/richardmleggett/NanoOKReporter.

**Received:** 29 July 2019; **Accepted:** 1 November 2019; **Published online:** 16 December 2019

**References**

1. Blander, J. M., Longman, R. S., Iliev, I. D., Sonnenberg, G. F. & Artis, D. Regulation of inflammation by microbiota interactions with the host. *Nat. Immunol.* 18, 851–860 (2017).

2. Lewis, B. B. & Pamer, E. G. Microbiota-based therapies for *Clostridium difficile* and antibiotic-resistant enteric infections. *Annu. Rev. Microbiol.* 71, 157–178 (2017).

3. Pedersen, H. K. et al. Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* 535, 376–381 (2016).

4. Lane, E. R., Zisman, T. L. & Suskind, D. L. The microbiota in inflammatory bowel disease: current and therapeutic insights. *J. Inflamm.* 10, 63–73 (2017).

5. Lipterr, K. et al. Gut microbiota dysbiosis associated with glucose metabolism disorders and the metabolic syndrome in older adults. *Benef. Microbes* 8, 545–556 (2017).

6. Wekerle, H. Brain autoimmunity and intestinal microbiota: 100 trillion game changers. *Trends Immunol.* 38, 483–497 (2017).

7. Kheireddine, A. & Sadowsky, M. J. Understanding the mechanisms of fecal microbiota transplantation. *Nat. Rev. Gastroenterol. Hepatol.* 13, 508–516 (2016).

8. Mayor, S. First WHO antimicrobial surveillance data reveal high levels of resistance globally. *BMJ* 360, k462 (2018).

9. Hoffman, S. J. et al. Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J. Antimicrob. Chemother.* 72, 104–114 (2017).

10. Greninger, A. L. et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med.* 7, 59 (2015).

11. Leppert, R. M., Heavey, D. Cacccomo, M., Clark, M. D. & Davey, R. P. NanoOK: multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. *Bioinformatics* 32, 142–144 (2016).

12. Ip, C. L. C. et al. MiniON Analysis and Reference Consortium: phase 1 data release and analysis. *F1000Res.* 4, 1075 (2015).

13. Melville, J. M. & Moss, T. J. The immune consequences of preterm birth. *Front. Neurosci.* 7, 89 (2013).

14. Alcon-Giner, C. et al. Optimisation of 16S rRNA gut microbiota profiling of extremely low birth weight infants. *BMC Genomics* 18, 841 (2017).

15. Chen, H. N., Lee, M. L., Yu, W. K., Lin, Y. W. & Tiao, L. Y. Late-onset *Enterobacter cloacae* sepisin in very-low-birth-weight neonates: experience in a medical center. *Pediatr. Neonatol.* 50, 3–7 (2009).

16. Hu, Y., Gao, G. F. & Zhu, B. The antibiotic resistance gene: flow game in environments, animal and human beings. *Front. Med.* 11, 161–168 (2017).

17. Serafini, F. et al. Insights into physiological and genetic mupirocin susceptibility in bifidobacteria. *Appl. Environ. Microbiol.* 77, 3141–3146 (2011).

18. Xu, H., Miao, V., Kwong, W., Xia, R. & Davies, J. Identification of a novel fosfomycin resistance gene (fosA2) in *Enterobacter cloacae* from the Salmon River, Canada. *Lett. Appl. Microbiol.* 52, 427–429 (2011).

19. Brooks, B. et al. The developing premature infant gut microbiome is a major factor shaping the microbiome of neonatal intensive care unit rooms. *Pediatr. Microbiol.* 6, 112 (2018).

20. Sim, K. et al. Dysbiosis anticipating necrotizing enterocolitis in very premature infants. *Clin. Infect. Dis.* 60, 389–397 (2015).

21. Liakopoulos, A., Mevis, D. J. & Ceccarelli, D. A review of SHV extended-spectrum β-lactamases: neglected yet ubiquitous. *Front. Microbiol.* 7, 1374 (2016).

22. *Antimicrobial wild type distributions of microorganisms* Version 5.26 (European Committee on Antimicrobial Susceptibility Testing, 2018); https://mic.eucast.org/Eucast2/

23. Brown, B. L., Watson, M., Minot, S. S., Rivera, M. C. & Franklin, R. B. MiniON™ nanopore sequencing of environmental metagenomes: a synthetic approach. *Gigascience* 6, 1–10 (2017).

24. Anand, R. J., Leaphart, C. L., Mollen, K. P. & Hackham, D. J. The role of the intestinal barrier in the pathogenesis of necrotizing enterocolitis. *Shock* 27, 124–133 (2007).

25. Hodzic, Z., Bolock, A. M. & Good, M. The role of mucosal immunity in the pathogenesis of necrotizing enterocolitis. *Front. Pediatr.* 5, 40 (2017).

26. Vargas, L. C. L. & Holh, K. E. Klebsiella pneumoniae population genomics and antimicrobial-resistant clones. *Trends Microbiol.* 24, 944–956 (2016).

27. Kielbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21, 487–493 (2011).

28. Bolger, A. M., Lobse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2141–2142 (2014).

29. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410 (1990).

30. Huson, D. H. et al. MEGAN Community Edition: interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput. Biol.* 12, e1004957 (2016).

31. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100 (2018).

32. Nicholls, S., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* 8, giz043 (2019).

33. McNab, R. A. G. et al. *The comprehensive antibiotic resistance database. Antimicrob. Agents Chemother.* 57, 3348–3357 (2013).

34. van der Helm, E. et al. Rapid resistome mapping using nanopore sequencing. *Nucleic Acids Res.* 45, e61 (2017).

35. Xia, Y. et al. MiniON nanopore sequencing enables correlation between resistance phenotype and genotype of coliform bacteria in municipal sewage. *Nucleic Acids Res.* 48, 2105 (2018).

36. Arango-Argoty, G. A. et al. NanoARG: a web service for detecting and contextualizing antimicrobial resistance genes from nanopore-derived metagenomes. *Microbiome* 7, 88 (2019).

37. Weisburg, W. G., Barns, S. M., Pelletier, D. A. & Lane, D. J. 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.* 171, 697–703 (1989).

38. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28, 1823–1829 (2012).
Acknowledgements

R.M.L., D.H. and M.D.C’s MinION work is supported by Biotechnology and Biological Sciences Research Council (BBSRC) Tools and Resources Development Fund award no. BB/N023196/1, BBSRC National Capability grant nos. BB/J010375/1 and BB/CCG17201/1, BBSRC Institute Strategic Programme grant no. BB/J004669/1 and BBSRC Core Strategic Programme grant no. BB/CSP17270/1. This work was funded by a Wellcome Trust Investigator Award (no. 100/974/C/13/Z); a BBSRC Norwich Research Park Bioscience Doctoral Training grant no. BB/M011216/1 (supervisor L.J.H., students C.A.G. and M.K.); an Institute Strategic Programme Gut Microbes and Health grant no. BB/R012490/1 and its constituent projects BBS/E/F/000PR10353 and BBS/E/F/000PR10356; and an Institute Strategic Programme Gut Health and Food Safety grant no. BB/J004529/1 to L.J.H. Isolation work was funded by a Microbiology Society Research Visit grant no. RVG16/03 to T.C.B. L.J.H. is in receipt of a Medical Research Council Intermediate Research Fellowship in Data Science (UK MED-BIO, grant no. MR/L01632X/1). We are grateful for the assistance of the Genomics Pipelines team at Earlham Institute, as well as the NBI Computing Infrastructure for Science team. We are also grateful to research nurse K. Few for obtaining consent from parents and collecting the samples. We thank C. Bennett and S. Stanbridge of the Earlham Institute Communications team for producing the accompanying video. The following reagent was obtained through BEI Resources, National Institute of Allergy and Infectious Diseases, National Institutes of Health as part of the Human Microbiome Project: Genomic DNA from Microbial Mock Community B (Staggered, High Concentration), v:5.2H, for Whole Genome Shotgun Sequencing, HM-277D.

Author contributions

L.J.H., M.D.C. and R.M.L. designed the research. L.J.H. (clinical), M.D.C. (technical) and R.M.L. (bioinformatics) managed the research. C.A.G. led on clinical sample selection, DNA extraction, phenotypic testing studies and AMR data analysis. D.H. and N.P. performed the library preparation and MinION sequencing. D.H. prepared the mock communities and carried out the sequencing. M.K. performed the 16S rRNA alignments. R.M.L. performed the majority of the bioinformatics analysis and contributed the new software toolsNanoOK RT and NanoOKReporter. H.A.P performed the antimicrobial susceptibility testing. S.C. provided bioinformatics support and analysis (Illumina data). T.C.B. isolated Klebsiella from patient samples. L.H. contributed to Klebsiella data interpretation and manuscript drafting. S.M. carried out the nanopore assemblies and wrote the AMR gene grouping scripts. P.C. provided access to clinical samples and clinical context for research. L.J.H. is the chief investigator on the preterm microbiota study (BAMBI) providing the samples used in this study. C.A.G., L.J.H., M.D.C. and R.M.L. helped with data interpretation/analysis and wrote the paper.

Competing interests

The authors have not received direct financial contributions from ONT; however, R.M.L. and M.D.C. have received a small number of free flow cells as part of the MAP and MARC programmes. R.M.L. is in receipt of travel and accommodation expenses to speak at two ONT-organized conferences and is on a PhD student advisory team with a member of ONT staff.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-019-0626-z.

Supplementary information is available for this paper at https://doi.org/10.1038/s41564-019-0626-z.

Correspondence and requests for materials should be addressed to R.M.L., L.J.H. or M.D.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s), 2019
Extended Data Fig. 1 | Principal Coordinate Analysis (PCoA) based on taxonomic assignments of MinION preterm faecal samples. PCoA was performed in MEGAN using the species-level taxonomic assignments of the preterm infant samples (n=9 biologically independent samples). Green arrows indicate the biplot loading vectors determined by MEGAN (the taxa with the largest influence in the plot). Samples were drawn from two groups: (i) healthy preterms (P106, P103 and P116) and (ii) preterms diagnosed with NEC or sepsis (P8, P49, P205 and P10). The PCoA plot indicates that distribution of samples from healthy preterms was distinct from samples belonging to infants diagnosed with suspected sepsis or NEC.
Extended Data Fig. 2 | Nanopore flow cell versions and yields. Yield and length metrics for Nanopore runs. Flow cell for sample P8 (2D) was used for two experiments: (a) Initial Metrichor basecalled run which was abandoned due to network lag and (b) Local basecalled run which was used for subsequent analysis. Runs 1-8 were performed using an earlier version of MinKNOW (ONT’s control software) which classified reads into “pass” and “fail” according to quality. Runs 9-14 used a later version of MinKNOW which removes the quality distinction, however some reads can still fail basecalling. The low yield for run 10 (P49A) is due to the flow cell having a low number of active pores at QC.

| Run | Sample   | Flow cell | Seq’ing kit | Library type | Total no. of reads | No. of pass reads | Mean length of pass reads (bp) | Pass read N50 (bp) | Longest pass read (bp) |
|-----|----------|-----------|-------------|--------------|--------------------|-------------------|---------------------------------|-------------------|-----------------------|
| 1   | Mock     | R7.3 (MAP103) | MAP006     | 2D           | 148,441            | 71,675            | 3,047                           | 5,497             | 40,561                |
| 2   | P10N     | R7.3 (MAP103) | MAP006     | 2D           | 145,342            | 82,734            | 2,926                           | 3,910             | 17,979                |
| 3   | P10R (1) | R7.3 (MAP103) | MAP006     | 2D           | 103,705            | 35,560            | 2,958                           | 3,967             | 18,069                |
| 4   | P10R (2) | R7.3 (MAP103) | MAP006     | 2D           | 45,486             | 12,118            | 3,832                           | 4,832             | 23,423                |
| 5   | P10V (1) | R7.3 (MAP103) | MAP006     | 2D           | 165,026            | 53,437            | 925                             | 1,087             | 11,300                |
| 6   | P10V (2) | R7.3 (MAP103) | MAP006     | 2D           | 69,427             | 21,780            | 2,027                           | 2,318             | 15,206                |
| 7   | P8       | R9.4 (MIN106) | RAD002     | 1D Rapid     | 69,442             | 44                | 1,480                           | 3,089             | 7,848                 |
| 8   | P8       | R9.5 (MIN107) | LSK108     | 1D Ligation  | 1,369,544          |                   | basecalling stopped at          |                   |                       |
|     |          |           |            |              |                    |                   | 633,226                         |                   |                       |
| 9   | P49A     | R9.5 (MIN107) | LSK108     | 1D Ligation  | 84,527             | 72,814            | 1,046                           | 1,338             | 34,975                |
| 10  | P205G    | R9.5 (MIN107) | LSK108     | 1D Ligation  | 2,745,619          | 2,415,324         | 966                             | 1,102             | 11,619                |
| 11  | P106i    | R9.5 (MIN107) | LSK108     | 1D Ligation  | 593,183            | 443,299           | 536                             | 533               | 3,655                 |
| 12  | P116i    | R9.5 (MIN107) | LSK108     | 1D Ligation  | 422,155            | 392,783           | 464                             | 466               | 20,356                |
| 13  | P103M    | R9.5 (MIN107) | RAD004     | 1D Rapid     | 1,355,250          | 1,221,699         | 1,484                           | 1,957             | 153,729               |
| 14  | P8 isolate | R9.4 (MIN106) | LSK108     | 1D Ligation  | 513,022            | 513,022           | 5,891                           | 7,192             | 96,613                |
| 15  | 8-species Mock with spike-in | R9.4.1 (MIN106D) | LSK109     | 1D Ligation  | 3,025,193          | 2,456,427         | 6,743                           | 8,379             | 128,257               |
| 16  | P103M with 4% spike-in | R9.4.1 (MIN106D) | LSK109     | 1D Ligation  | 256,296            | 229,585           | 6,353                           | 8,166             | 125,506               |
| 17  | P103M with 40% spike-in | R9.4.1 (MIN106D) | LSK109     | 1D Ligation  | 176,189            | 153,055           | 8,076                           | 11,052            | 201,346               |
| 18  | P129B    | R9.4.1 (FLG001) | LSK109     | 1D Ligation  | 430,314            | 165,175           | 1,646                           | 3,339             | 26,367                |
| 19  | P129B    | R9.4.1 (FLG001) | LSK109     | 1D Ligation  | 268,981            | 154,774           | 1,682                           | 3,296             | 26,592                |
| 20  | Zymo mock | R9.4.1 (MIN106D) | LSK109     | 1D Ligation  | 6,337,027          | 5,517,313         | 2,761                           | 3,539             | 30,413                |
Extended Data Fig. 3 | Presence/absence heat maps of AMR hits found in P10 using MinION and Illumina. Three samples of preterm P10 were analysed and sequenced using MinION (P10N-N, P10R-N and P10V-N), and Illumina technology (P10N-I, P10R-I and P10V-I). (a) Heat map representing efflux pumps inhibitors or regulators genes found in samples P10N, P10R and P10V. (b) Heat map highlighting β-lactamases. (c) Heat map showing (1) aminocoumarin resistance genes (2) aminoglycosides resistance (3) bacitracin resistance (4) colistin resistance (5) erythromycin resistance (6) fluoroquinolone resistance (7) fosfomycin resistance (8) mupirocin resistance (9) quinolone resistance (10) streptothricin resistance (11) sulphonamide resistance (12) tetracycline resistance (13) trimethoprim resistance. AMR genes were grouped according to sequence similarity. Blue indicates presence, grey indicates absence. Further information on genes detected can be found in Supplementary Table 2.
Extended Data Fig. 4 | Rapid diagnostic of healthy preterms P106 and P116 receiving probiotic supplementation. (a), (b) Taxonomic profiles obtained using MiniON technology for preterms P106 and P116, respectively. Figure legend comprises the 8 most abundant taxa represented. Further information on number of reads classified can be found in Supplementary Table 3. (c), (d) Heat maps displaying number of CARD database hits detected among the most common groups of antibiotic resistance genes found in preterms P106 and P116. Top and lower panel indicate the hours since sequencing started and the number of reads analyzed, respectively within this timeframe.
Extended Data Fig. 5 | Comparison of taxonomic profiles using Illumina and MinION technology for preterm P8. (a) Taxonomic profiles obtained using Illumina HiSeq 2500 and MinION technology for preterm P8. Figure legend comprises the 15 most abundant taxa classified. (b) Species-level correlation plot between normalised read count using Illumina (x-axis) and MinION (y-axis) sequencing. Taxa n=121, with log transformed Pearson’s r = 0.91. The grey region either side of fit line represents 95% confidence intervals.
Extended Data Fig. 6 | AMR genes associated to *Klebsiella pneumoniae* from a metagenomic sample P8 compared to those found in P8 *Klebsiella pneumoniae* isolate. Heat maps displaying AMR genes associated to *K. pneumoniae* from faecal sample P8 and AMR genes detected from a *K. pneumoniae* isolate from the same sample. Sequencing of the metagenomic sample was performed for only 6h using MinION and NanoOK RT tool (highlighted as “P8”). Sequencing and assembly of the *K. pneumoniae* isolate was performed using both MinION (‘Isolate MinION’) and Illumina HiSeq (‘Isolate Illumina’). AMR genes were grouped according to sequence similarity. Blue indicates presence, grey indicates absence. The “Walkout taxa” column shows the taxa that NanoOK RT’s walkout function classified the nanopore AMR hits to. An “n/a” indicates the gene is not found in the metagenomic nanopore sample. A single read can only be classified to a single taxa, but multiple taxa are shown if different reads classify to different taxa. Further details of specific gene names can be found in Supplementary Table 8.
Extended Data Fig. 7 | Rapid diagnostic of healthy preterm P103 using SQK-RAD004. (a) Taxonomic profiles detected for preterm P103. Figure legend comprises the 8 most abundant taxa obtained. Further information on specific taxa read counts can be found in Supplementary Table 3. (b) Correlation plot of normalised assigned reads from P103 at 1h of sequencing (x-axis) and at 6h of sequencing (y-axis). Taxa n=131, with log transformed Pearson’s $r = 0.99$. The grey region either side of fit line represents 95% confidence intervals. (c) Heat map displaying number of CARD database hits detected among the most common groups of antibiotic resistance genes found in preterm P103. Top and lower panel indicate the hours since sequencing started and the number of reads analyzed within this timeframe. Further information on the specific AMR genes classified can be found in Supplementary Table 4. (d) Walkout results reported by NanoOK RT software. Results shown are from independent hits (bacterial host hit $\geq 50$ bp from the AMR hit) at 6 hours of sequencing.
Extended Data Fig. 8 | Use of flongle flowcells to profile preterm infant P129 suffering from NEC. (a) Taxonomic profiles obtained using flongle flowcell on GridION and MinION for preterm P129. Figure legend comprises the 12 most abundant taxa. (b) MEGAN taxonomy tree representing species assigned for GridION (blue) and MinION (brown) flongle flowcells. The height of the bars corresponds to the number of reads assigned, with normalised counts given next to the species name. (c) Species-level correlation plot between normalised read counts from GridION and MinION. Taxa n=60, with log transformed Pearson’s $r = 0.92$. The grey region either side of fit line represents 95% confidence intervals. (d) Heat map indicating AMR genes detected for preterm P129 using NanoOK RT tool and Flongle flowcells run on a GridION and MinION.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

☑️ ○ n/a Confirmed

☐☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐☒ The statistical test(s) used AND whether they are one- or two-sided
  Only common tests should be described solely by name; describe more complex techniques in the Methods section.
☐☒ ☐ A description of all covariates tested
☐☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐☒ ☐ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐☒ ☐ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  Give P values as exact values whenever suitable.
☐☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐☒ ☐ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated
☐☒ ☐ Clearly defined error bars
  State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

No software was used that is not described in detail in the methods. Collection software (MinKnow version not given) is available from Oxford Nanopore Technologies (the MinION sequencer provider), basecalling was via Guppy (version 2.3.1), all other software is available through an open source license.

Data analysis

No software was used that is not described in detail in the methods and available through an open source license. NCBI BLAST (version 2.2.29), MEGAN (version 6), NanoOK (version 0.54), NanoOK RT (version 1), MiniMap2 (version 2.17-rd43), Canu (version 1.5), Nanopolish (version 0.9.0), Falcon (version 1.3.1), Velvet (version 1.2.10), Fyce (version 2.4), minimus2, mummer3.

For manuscripts utilizing custom algorithms or software that are not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. Github). See the Nature Research guidelines for submitting code & software for further information.
Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The Illumina and MinION read data supporting the conclusions of this article are available in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under study accession PRJEB22207.
No figures have associated raw data.
Data is freely available.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences  Behavioural & social sciences  Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flap.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | We did not conduct statistical analysis. |
|-------------|------------------------------------------|
| Data exclusions | Some sequencing runs were excluded if the QC metrics showed that the data was poor quality (greater than 50% were "pass reads" in MinKNOW), or if there was insufficient sequence data (less than 800 pores were available and flowcells were returned to ONT). We believe much of this to be due to batch variation from the sequencing provider, which is common with new start ups simultaneous developing their technology and their production capacity. In these cases the samples were reprocessed and this is referred to in the manuscript. |
| Replication | For patient sample P8 we repeated the sample twice on the MinION (both runs gave equivalent microbiota profiles see also summary in extended data 2). In an earlier version of the study we sequenced P8 another time with a 2D run (an older chemistry) which gave equivalent microbiota profiles see preprint Fig8. We also verified the P10 samples using Illumina sequencing technology which gave equivalent microbiota profiles see Fig2. |
| Randomization | We selected patients from a clinical study (BAMBi study, PI Hall) samples were selected by criteria of the amount of fecal sample remaining from the BAMBi study being sufficient for at least one sequencing run (ONT recommended 1ug of DNA at the start of this project, later decreased to 400ng). |
| Blinding | We were aware of the status of the patients, when selecting samples (to ensure some healthy and some NEC patients were selected). In this study we were not trying to demonstrate an ability to discriminate between healthy and NEC patients, only that MinION technology could capture the microbiota and AMR profiles. |

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Study description | Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study). |
| Research sample | State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source. |
| Sampling strategy | Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to determine sample size. If no sample size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed. |
| Data collection | Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection. |
| Timing | Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort. |
Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

**Study description**
Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.

**Research sample**
Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.

**Sampling strategy**
Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.

**Data collection**
Describe the data collection procedure, including who recorded the data and how.

**Timing and spatial scale**
Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken.

**Data exclusions**
If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

**Reproducibility**
Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.

**Randomization**
Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.

**Blinding**
Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

**Field work, collection and transport**

**Field conditions**
Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).

**Location**
State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).

**Access and import/export**
Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).

**Disturbance**
Describe any disturbance caused by the study and how it was minimized.

**Reporting for specific materials, systems and methods**
### Materials & experimental systems

| Method | n/a | Involved in the study |
|--------|-----|-----------------------|
|        |     | Unique biological materials |
|        |     | Antibodies |
|        |     | Eukaryotic cell lines |
|        |     | Palaeontology |
|        |     | Animals and other organisms |
|        |     | Human research participants |

### Methods

| Method | n/a | Involved in the study |
|--------|-----|-----------------------|
|        |     | ChiP-seq |
|        |     | Flow cytometry |
|        |     | MRI-based neuroimaging |

### Unique biological materials

**Policy information about availability of materials**

| Activity | Description |
|----------|-------------|
| Obtaining unique materials | Describe any restrictions on the availability of unique materials OR confirm that all unique materials used are readily available from the authors or from standard commercial sources (and specify these sources). |

### Antibodies

**Policy information about antibodies**

| Activity | Description |
|----------|-------------|
| Antibodies used | Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number. |
| Validation | Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer’s website, relevant citations, antibody profiles in online databases, or data provided in the manuscript. |

### Eukaryotic cell lines

**Policy information about cell lines**

| Activity | Description |
|----------|-------------|
| Cell line source(s) | State the source of each cell line used. |
| Authentication | Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated. |
| Mycoplasma contamination | Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination. |
| Commonly misidentified lines | None any commonly misidentified cell lines used in the study and provide a rationale for their use. (See ITAG register) |

### Palaeontology

**Policy information about palaeontology**

| Activity | Description |
|----------|-------------|
| Specimen provenance | Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). |
| Specimen deposition | Indicate where the specimens have been deposited to permit free access by other researchers. |
| Dating methods | If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided. |

- Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

### Animals and other organisms

**Policy information about studies involving animals**

| Activity | Description |
|----------|-------------|
| Laboratory animals | For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals. |
| Wild animals | Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals. |
| Field-collected samples | For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field. |
Human research participants

Policy information about studies involving human research participants.

Population characteristics
The BAMBI study (PI Hall) from which samples were taken includes patients of less than 37 weeks gestation, or that had entered the Neonatal Intensive Care Unit (NICU). Our subset included samples from patients ranging in ages 22-33 weeks of gestation, and were sampled in the NICU aged 1-9 weeks.

Recruitment
Samples were selected based from a wider study of pre-term babies (BAMBI) and we wanted to capture a range of health outcomes including healthy, NEC, sepsis and probiotic supplementation. From within those four classes we sampled randomly based on availability of material.

ChIP-seq

Data deposition
☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.
☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
For "initial submission" or "revised version" documents, provide reviewer access links. For your "final submission" document, provide a link to the deposited data.

Files in database submission
Provide a list of all files available in the database submission.

Genome browser session (e.g. UCSC)
Provide a link to an anonymized genome browser session for "initial submission" and "revised version" documents only, to enable peer review. Write "no longer applicable" for "final submission" documents.

Methodology

Replicates
Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth
Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies
Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters
Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality
Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5 fold enrichment.

Software
Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots
☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
☐ All plots are contour plots with outliers or pseudocolor plots.
☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation
Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

Instrument
Identify the instrument used for data collection, specifying make and model number.

Software
Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.

Cell population abundance
Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.
Magnetic resonance imaging

Experimental design

| Design type | Indicate task or resting state, event-related or block design. |
|-------------|------------------------------------------------------------|
| Design specifications | Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials. |
| Behavioral performance measures | State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects). |

Acquisition

| Imaging type(s) | Specify: functional, structural, diffusion, perfusion. |
| Field strength | Specify in Tesla |
| Sequence & imaging parameters | Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle. |
| Area of acquisition | State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined. |

Preprocessing

| Preprocessing software | Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.). |
| Normalization | If data were normalized/standardized, describe the approach(es); specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization. |
| Normalization template | Describe the template used for normalization/registration, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized. |
| Noise and artifact removal | Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration). |
| Volume censoring | Define your software and/or method and criteria for volume censoring, and state the extent of such censoring. |

Statistical modeling & inference

| Model type and settings | Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or autocorrelation). |
| Effect(s) tested | Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used. |
| Specify type of analysis: | Whole brain | ROI-based | Both |
| Statistical type for inference | Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods. |
| Correction | Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo). |

Models & analysis

| n/a | Involved in the study |
|-----|-----------------------|
|      | Functional and/or effective connectivity |
|      | Graph analysis |
|      | Multivariate modeling or predictive analysis |
| Section                              | Guidance                                                                                          |
|-------------------------------------|--------------------------------------------------------------------------------------------------|
| Functional and/or effective connectivity | Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information). |
| Graph analysis                      | Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.). |
| Multivariate modeling and predictive analysis | Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics. |