Preterm infants harbour diverse *Klebsiella* populations, including atypical species that encode and produce an array of antimicrobial resistance- and virulence-associated factors

Yuhao Chen[1]*, Thomas C. Brook[2]*, Cho Zin Soe[3], Ian O'Neill[3], Cristina Alcon-Giner[3], Onnicha Leelastwattanagul[4], Sarah Phillips[3], Shabhonam Caim[3], Paul Clarke[5,6], Lindsay J. Hall[3]†, Lesley Hoyles[1,7]†

[1]Department of Surgery & Cancer, Faculty of Medicine, Imperial College London, London, United Kingdom

[2]Department of Biomedical Sciences, Faculty of Science and Technology, University of Westminster, London, United Kingdom

[3]Gut Microbes & Health, Quadram Institute Bioscience, Norwich Research Park, Norwich, United Kingdom

[4]Bioinformatics & Systems Biology Program, School of Bioresources and Technology, King Mongkut's University of Technology Thonburi (Bang Khun Thian Campus), Bangkok, Thailand

[5]Neonatal Intensive Care Unit, Norfolk and Norwich University Hospitals NHS Foundation Trust, Norwich, United Kingdom

[6]Norwich Medical School, University of East Anglia, Norwich, United Kingdom

[7]Department of Biosciences, School of Science and Technology, Nottingham Trent University, Nottingham, United Kingdom

*These authors contributed equally to this work.

†**Correspondence:** Lindsay J. Hall, Lindsay.Hall@quadram.ac.uk; Lesley Hoyles, lesley.hoyles@ntu.ac.uk

ABSTRACT

Klebsiella spp. are frequently enriched in the gut microbiota of preterm neonates, and overgrowth is associated with necrotizing enterocolitis (NEC), nosocomial infections and late-onset sepsis. Little is known about the genomic and phenotypic characteristics of preterm-associated *Klebsiella* as previous studies have focussed on recovery of antimicrobial-resistant isolates or culture-independent molecular analyses. The aim of this study was to better characterize preterm-associated *Klebsiella* populations using phenotypic and genotypic approaches. Faecal samples from a UK cohort of healthy and sick preterm neonates (*n*=109) were screened on MacConkey agar to isolate lactose-positive *Enterobacteriaceae*. Whole-genome sequences were generated for *Klebsiella* spp., and virulence and antimicrobial resistance genes identified. Antibiotic susceptibility profiling, and *in vitro* macrophage and iron assays were undertaken for the *Klebsiella* strains. Metapangenome analyses with a manually curated genome dataset were undertaken to examine diversity of *Klebsiella oxytoca* and related bacteria in a publicly available shotgun metagenome dataset. Approximately one-tenth of faecal samples harboured *Klebsiella* spp. (*Klebsiella pneumoniae*, 7.3 %; *Klebsiella quasipneumoniae*, 0.9 %; *Klebsiella grimontii*, 2.8 %; *Klebsiella michiganensis*, 1.8 %). Isolates recovered from NEC- and sepsis-affected infants and those showing no signs of clinical infection (i.e. 'healthy') encoded multiple β-lactamases. No difference was observed between isolates recovered from 'healthy' and sick infants with respect to *in vitro* siderophore production (all encoded enterobactin in their genomes). All *K. pneumoniae*, *K. quasipneumoniae*, *K. grimontii* and *K. michiganensis* faecal isolates tested were able to reside and persist in macrophages, indicating their immune evasion abilities. Metapangenome analyses of published metagenomic data confirmed our findings regarding the presence of *K. michiganensis* in the preterm gut. There is little difference in the phenotypic and genomic characteristics of *Klebsiella* isolates recovered from 'healthy' and sick infants. Identification of β-lactamases in all isolates may prove problematic when defining treatment regimens for NEC or sepsis, and suggests 'healthy' preterm infants contribute to the resistome. Refined analyses with curated sequence databases are required when studying closely related species present in metagenomic data.

Keywords: *Klebsiella oxytoca*, shotgun metagenomics, taxonomy, microbiome.

Author notes: All supporting data have been provided within the article or through supplementary data files, available from figshare.

Abbreviations: AMR, antimicrobial resistance; ANI, average nucleotide identity; CARD, Comprehensive Antibiotic Resistance Database; GI, gastrointestinal; LOS, late-onset sepsis; LPE, lactose-positive *Enterobacteriaceae*; LPS, lipopolysaccharide; MAG, metagenome-assembled genome; NEC, necrotizing enterocolitis; NICU, neonatal intensive care unit; NNUH, Norfolk and Norwich University Hospital; OTU, operational taxonomic unit; VFDB, Virulence Factors of Pathogenic Bacteria Database.

Impact statement

Polyphasic characterization of isolates recovered from the faeces of preterm infants has demonstrated that *Klebsiella* spp. recovered from these patients are genomically more diverse than previously recognized. All *K. pneumoniae*, *K. quasipneumoniae*, *K. grimontii* and *K. michiganensis* faecal isolates studied were able to reside and persist in macrophages, indicating their immune evasion abilities and potential for causing infections in at-risk infants. The identification of *K. michiganensis* in samples, and the abundance of *K. michiganensis* genomes in public repositories, adds to the growing body of evidence indicating that *K. michiganensis* is likely to be more clinically relevant than *K. oxytoca* in human-associated infections. Metapangenome analyses of publicly available shotgun metagenomic data confirmed the prevalence of *K. michiganensis* in the faeces of preterm infants, and highlighted the need for refined taxonomic analyses when splitting closely related species from one another in metagenomic studies.

Data summary

16S rRNA gene sequence data associated with this article have been deposited at DDBJ/ENA/GenBank under BioProject accession PRJEB34372. The Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under BioProject accession PRJNA471164. The metagenome-assembled genomes are available from figshare. Published sequence data of Ward *et al.* (1), used to generate the metagenome-assembled genomes, are available under BioProject accession number 63661.

INTRODUCTION

The gut microbiota encompasses bacteria, archaea, lower eukaryotes and viruses, with these microbes contributing to host gastrointestinal (GI) and systemic health. Host–microbiome interactions within the intestine are particularly important in neonates, contributing to development of the immune response, establishment of the gut microbiome and protection from infections (2,3). Term infants (i.e. gestation 37 weeks) are rapidly colonised after exposure to the mother's microbiota and the environment, with streptococci and *Enterobacteriaceae* dominating in the initial phases (2), and *Bifidobacterium* spp. becoming prominent as the infant grows (2).

In contrast, colonization of preterm infants (i.e. <37 weeks' gestation) occurs in neonatal intensive care units (NICUs) and is shaped by the significant number of antibiotics ('covering' (i.e. to cover possible early onset infection from birth) and treatment) these infants receive in the first days and weeks post birth. The microbiota in preterm infants is enriched for bacteria such as *Enterobacteriaceae*, *Enterococcus* and *Staphylococcus* (4,5).

Critically, colonization of these at-risk infants with potentially pathogenic taxa, in concert with an unstable microbiome, and immaturity of their GI tract and immune system, is thought to contribute to nosocomial infections such as late-onset sepsis (LOS) or necrotizing enterocolitis (NEC) (6–12).

The family *Enterobacteriaceae* comprises more than 25 genera of catalase-positive, oxidase-negative Gram-negative bacteria and encompasses many pathogens [e.g. *Escherichia* (*Esc.*) *coli, Klebsiella pneumoniae, Shigella* (*Shi.*) *dysenteriae, Enterobacter* (*Ent.*) *cloacae, Serratia* (*Ser.*) *marcescens* and *Citrobacter* spp.] (13). While coagulase-negative staphylococci are the most common cause of LOS in preterm infants, *Enterobacteriaceae* that translocate from the preterm gut to the bloodstream also cause this condition (8,9). In addition, *Enterobacteriaceae* are associated with higher morbidity than the staphylococci, and blooms in *Proteobacteria* – thought to be linked to impaired mucosal barrier integrity – have been reported immediately prior to the diagnosis of LOS (8,9,14). Predictions made from shotgun metagenomic data show replication rates of all bacteria – and especially the *Enterobacteriaceae* and *Klebsiella* – are significantly increased immediately prior to NEC diagnosis (15). This altered gut microbiome influences intestinal homeostasis and contributes to NEC (16), in tandem with the immature preterm immune system contributing to intestinal pathology in response to blooms of *Proteobacteria*.

Associations between *Klebsiella*-related operational taxonomic units (OTUs) and the development of NEC have been noted, suggesting members of this genus contribute to the aetiology of NEC in a subset of patients (17,18). Although Sim *et al.* (17) found one of their two distinct groups of NEC infants had an overabundance of a *Klebsiella* OTU, these researchers failed to identify a single predominant species of *Klebsiella*, recovering representatives of several genera (*K. pneumoniae*, *Klebsiella oxytoca*, *Klebsiella aerogenes*, *Ent. cloacae*, *Esc. coli* and *Ser. marcescens*) from samples. *Klebsiella* spp. and their fimbriae-encoding genes were significantly enriched in faeces collected immediately prior to the onset of NEC in a US infant cohort. These fimbriae may contribute to the overexpression of TLR4 receptors observed in preterm infants (15). Confirming the role of these bacteria in NEC will require reproducing certain aspects of the disease in model systems, using well-characterized bacteria recovered from preterm infants (14,17).

To date, there is limited information on the genomic and phenotypic features of preterm-associated *Klebsiella* spp. Thus, to characterise these important opportunistic pathogens, and to build a collection of preterm-associated *Klebsiella* strains for use in future mechanistic studies relevant to preterm-infant health, we isolated and characterized (phenotypically and genomically) bacteria from a cohort of preterm neonates enrolled in a study at the Norfolk and Norwich University Hospital (NNUH), Norwich, United Kingdom. Recovered *Klebsiella* isolates were subject to additional phenotypic tests that complemented genomic data. In addition, for the increasingly important species *K. oxytoca*, in which our laboratories have a specific interest, metapangenome analyses were undertaken to better understand the prevalence and potential virulence of this organism and related species in the context of the preterm neonate gut microbiota.

METHODS

Collection of faecal samples. Faeces were collected from premature neonates (<37 weeks' gestation) (**Supplementary Table 1**). The Ethics Committee of the Faculty of Medical and Health Sciences of the University of East Anglia (Norwich, UK) approved this study. The protocol for faeces collection was laid out by the Norwich Research Park (NRP) Biorepository (Norwich, UK) and was in accordance with the terms of the Human Tissue Act 2004 (HTA), and approved with licence number 11208 by the Human Tissue Authority. Infants admitted to the NICU of the NNUH were recruited by doctors or nurses with informed and written consent obtained from parents. Collection of faecal samples was carried out by clinical researchers and/or research nurses, with samples stored at -80 °C prior to DNA extraction.

16S rRNA gene sequencing and analyses. DNA was extracted from samples using the FastDNA SPIN Kit for Soil (MP Biomedicals) and processed for sequencing and analyses as described previously (19). This 16S rRNA gene sequence data associated with this project have been deposited at DDBJ/ENA/GenBank under BioProject accession PRJEB34372.

Isolation of bacteria and biochemical characterization. For isolation work, a single faecal sample from each baby (*n*=109; **Supplementary Table 1**) was thawed and 0.1 g homogenised in 1 mL TBT buffer (100 mM Tris/HCl, pH 8.0; 100 mM NaCl; 10 mM $MgCl_2•6H_2O$). Homogenates were serially diluted 10^{-1} to 10^{-4} in TBT buffer. Aliquots (50 µL) of homogenate were spread on MacConkey agar no. 3 (Oxoid Ltd) plates in triplicate and incubated aerobically at 37 °C overnight.

Differential counts (based on colony morphology) of all lactose-positive (i.e. pink) colonies were made in triplicate to calculate colony-forming units (CFUs) per gramme wet-weight faeces. One of each colony type per plate was selected and re-streaked on MacConkey agar three times to purify, incubating aerobically at 37 °C overnight each time. A single colony from each pure culture was re-suspended in 5 mL of sterile distilled water; the API 20E kit (bioMérieux) was used according to the manufacturer’s instructions to give preliminary identities for each of the isolates recovered.

DNA extraction, whole-genome sequencing and assembly. DNA was extracted using a phenol–chloroform method fully described previously (20) from overnight cultures of strains, and sequenced using the 96-plex Illumina HiSeq 2500 platform to generate 125 bp paired-end reads (21). Raw data provided by the sequencing centre were checked using fastqc v0.11.4 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/); no adapter trimming was required, and reads had an average Phred score >25. MetaPhlAn2.6 (22) was used to identify which species genome sequences represented. According to the results given by MetaPhlAn2.6, appropriate reference genomes were retrieved from Ensembl Genome (http://bacteria.ensembl.org/index.html) to guide reference-based assembly using BugBuilder v1.0.3b1 (default settings for Illumina data) (23). Summary statistics for the *Klebsiella* genome sequences generated in this study, including accession numbers, can be found in **Supplementary Table 2**. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under BioProject accession PRJNA471164.

Genome analyses. Average nucleotide identity (ANI) between genome sequences of isolates and reference strains (*Klebsiella grimontii* 06D021^{T}, GCA_900200035; *K. oxytoca* 2880STDY5682490, GCA_900083895.1; *Klebsiella michiganensis* DSM 25444^{T}, GCA_002925905) was determined using FastANI (default settings) (24).

K. oxytoca, *K. michiganensis* and *K. grimontii* genomes were uploaded to the *Klebsiella oxytoca*/*michiganensis* MLST website (https://pubmlst.org/koxytoca/) sited at the University of Oxford (25) on 28 July 2019 to determine allele number against previously defined house-keeping genes (*rpoB*, *gapA*, *mdh*, *pgi*, *phoE*, *infB* and *tonB*). *K. pneumoniae* genomes were analysed using the Institut Pasteur MLST database (https://bigsdb.pasteur.fr/klebsiella/klebsiella.html). Kleborate (26,27) and Kaptive (28) were used to identify capsular type and O antigen type.

Virulence genes were identified by BLASTP of genome amino acid sequences against the Virulence Factors of Pathogenic Bacteria Database (VFDB; ‘core dataset’ downloaded 27 July 2019) (29); results are reported for >70 % identity and 90 % query coverage. Antimicrobial resistance (AMR) genes were identified by BLASTP against the Comprehensive Antibiotic Resistance Database (CARD) download (27 July 2019; protein homolog dataset) (30); only strict and perfect matches with respect to CARD database coverage and bit-score cut-off recommendations are reported.

Genomic traits were visualized using anvi’o-5.5 (31) according to the pangenomic workflow. Briefly, for each figure presented herein, genomes were used to create an anvi’o contigs database, which contained ORFs predicted using Prodigal v2.6.3 (32). A multiple-sequence alignment was created using BLASTP. Markov CL algorithm (33) was used to identify gene clusters (--mcl-inflation

10; high sensitivity for identifying gene clusters of closely related species or strain level). Gene clusters and genomes were organized using Euclidean distance and Ward linkage, with results visualized using GoogleChrome.

Phenotypic characterization of *Klebsiella* isolates

Iron assays. Pre-cultures of *Klebsiella* isolates (5 mL) were grown overnight in LB broth (37 °C, 160 rpm). Aliquots (500 µL) were harvested (4000 rpm, 20 min) and the cell pellets washed twice with PBS. The cell suspensions (50 µL) were used to inoculate 5 mL cultures containing M9 minimal medium (Na_2HPO_4, 6.9 g/L; KH_2PO_4, 3 g/L; NaCl, 0.5 g/L; NH_4Cl, 1 g/L; $CaCl_2$, 0.1 mM; $MgSO_4$, 2 mM; 0.2 % glucose) at 37 °C. At 20 h, bacterial growth and siderophore production were measured using the CAS assay (34). An aliquot (100 µL) of the cell culture supernatant was mixed with CAS dye (100 µL), followed by the shuttle solution (4 µL) and siderophore production monitored at 620 nm at 4 h using a BioRad Benchmark Plus microplate spectrophotometer. A decrease in the blue colour of the CAS dye was measured using uninoculated medium as control. The estimated amount of total siderophore produced by *Klebsiella* isolates was calculated using the CAS standard curve based upon a desferrioxamine B standard (1:1).

Macrophage assays. All strains were grown on LB broth + 1.5 % agar and incubated overnight at 37 °C. THP-1 monocytes were obtained from ATCC (TIB-202) and were maintained in RPMI (Gibco: 72400021) plus 10 % heat-inactivated foetal bovine serum (FBS; Gibco: 10500064) in a humidified incubator at 37 °C with 5 % CO_2. THP-1 monocytes were differentiated into macrophages in RPMI + 10 mM HEPES + 10 % FBS + 10 ng/mL phorbol 12-myristate 13-acetate (PMA; Sigma) and seeded at 1×10^5 cells per well of a 96-well tissue culture dish and incubated for 15 h. Overnight cultures of bacteria were diluted 1:100 into fresh LB broth and grown until mid-exponential phase. Bacteria were then washed twice with PBS and diluted to 1×10^7 cfu/mL in RPMI + 10 mM HEPES + 10 % FBS and 100 µL of bacteria was added to each well. Plates were then centrifuged at 300 *g* for 5 min to synchronize infections. Bacteria–macrophage co-culture was incubated at 37 °C/5 % CO_2 for 30 min to allow for phagocytosis. Cells were then washed three times in PBS and medium was replaced with above culture medium supplemented with 300 μg/mL gentamicin and 100 units/mL polymyxin B to eliminate extracellular bacteria. Cell were again incubated at 37 °C/5 % CO_2 for 1.5 h. Cells were then washed three times with PBS and medium for cells for later time points was replaced with culture medium supplemented with 300 μg/mL gentamicin and incubated for a further 4.5 h. Intracellular bacterial load was enumerated by lysing macrophages in PBS + 1% Triton X-100 for 10 min at room temperature, serially diluting cultures and plating on LB agar. Plates were incubated overnight at 37 °C and colonies counted the following day.

Calculation of antibiotic minimal inhibitory concentration (MIC) for the *Klebsiella* isolates.

Broth microdilution method was used to calculate the MIC of the *Klebsiella* isolates. Serial two-fold dilutions of benzylpenicillin, gentamicin, and meropenem were added to sterile nutrient broth. The antibiotics used in this assay were supplied by the NICU of NNUH. Inoculum for each of the isolates was prepared using 10 mL from a fresh overnight culture. Microplates were incubated for 24 h at 37 °C under aerobic conditions. Optical density was monitored using a plate reader (BMG Labtech, UK) at 595 nm. MICs were determined as the lowest concentration of antibiotic inhibiting any bacterial growth. All experiments were repeated in triplicate. For the aminoglycoside gentamicin and the carbapenem meropenem, *Klebsiella* (*Enterobacteriaceae*) breakpoints were determined according to European Committee on Antimicrobial Susceptibility Testing (EUCAST) guidelines (version 8.1, published 16 May 2018, http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Breakpoint_tables/v_8.1_Breakpoint_Tables.pdf). No EUCAST data were available for benzylpenicillin (EUCAST states this aminopenicillin has no clinically useful activity against *Enterobacteriaceae*).

Estimation of abundance of *K. oxytoca* in shotgun metagenomic data. We chose to analyse a published preterm gut metagenome dataset (1) in this study, as it had been previously used to identify associations between uropathogenic *Esc. coli* and NEC. Trimmed, human-filtered, paired-end read data deposited in the Sequence Read Archive by Ward *et al.* (1) are available under BioProject accession number 63661. Information on Ward samples included in this study can be found in

Supplementary Table 3. Ward *et al.* (1) used MetaPhlAn to determine abundance of bacteria in samples. However, the marker genes used to enumerate *K. oxytoca* in the MetaPhlAn2.6 database are derived from 11 genomes, five of which are not *K. oxytoca* (*Raoultella ornithinolytica* 10-5246, GCF_000247895; *K. michiganensis* E718, GCF_000276705; *K. michiganensis* Kleb_oxyt_10-5250_V1, GCF_000247915; *K. michiganensis* KCTC 1686, GCF_000240325; *K. michiganensis* Kleb_oxyt_10-5242_V1, GCF_000247835). Therefore, relative abundance of bacteria was instead determined using Centrifuge (35). While MetaPhlAn2.6 relies on a pre-compiled database of unique marker genes for determining taxonomic abundance, the Centrifuge database can be updated at will using genomes downloaded from NCBI. A bacteria- and archaea-specific complete genome database was generated for use with Centrifuge via NCBI on 1 July 2018. Species-level abundances, based on read-level data, for *K. oxytoca* and *K. michiganensis* in the study of Ward *et al.* (1) were determined. (NB: *K. grimontii* genomes were not included in the Centrifuge database, nor are they included in MetaPhlAn2.6 or the most-recent version of Kraken2.)

Metapangenome analyses of *K. oxytoca, K. michiganensis* and *K. grimontii*. A total of 162 *K. oxytoca*-related whole-genome sequences were retrieved from GenBank on 31 May 2018 (**Supplementary Table 4**). On the basis of *bla*$_{\mathrm{OXY}}$, phylogenetic and ANI analyses (36), these had been confirmed to belong to *K. oxytoca* (*n*=64), *K. grimontii* (*n*=24) and *K. michiganensis* (*n*=74) (**Supplementary Table 5**). Prokka v1.13.3 (37) was used to annotate the 162 downloaded and five infant genomes. The resulting .gff files were subject to pangenome analyses using Roary v3.12.0 (default settings) (38). Genes present in 165–167 strains were defined as the core cluster, while those present in 25–164 strains were defined as the accessory cluster. Remaining genes that only existed in single strains were classified into strain-specific clusters. FastTree v2.1.10 (39) was used to generate a phylogenetic tree from the core gene alignment, with the tree visualized using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

PanPhlAn (panphlan_pangenome_generation.py v1.2.3.6; panphlan_profile.py v1.2.2.3; panphlan_map.py v1.2.2.5; (40)) was used to profile strains within metagenomes using the Roary-generated pangenome dataset. Gene-family clusters across all 167 available genomes and centroid sequence files outputted from Roary were uploaded to PanPhlAn to build a Bowtie2-indexed pangenome database, against which raw reads (concatenated read pair files) were mapped using Bowtie2 v2.3.0. The coverages of all gene positions were detected and extracted using samtools v1.4.1, and then integrated to a gene-family coverage profile for each sample. Pangenome-related strains were predicted to exist if a consistent and similar coverage depth across a set of 5774 gene families was detected under the non-default parameters of PanPhlAn (--min_coverage 1, --left_max 1.70, --right_min 0.30; panphlan_profile.py). Principal component analysis (PCA) was performed on 500 accessory genes randomly selected from the pangenome using the R package FactoMineR (41), allowing us to distinguish different species at the gene level.

Recovery of metagenome-assembled genomes (MAGs) from metagenomes. For metagenome samples in which *K. oxytoca*-related strains were identified using PanPhlAn, attempts were made to recover them as MAGs. All reads in samples were mapped against the pangenome database using Bowtie2 and mapped paired-end reads were extracted by using FastQ Screen v0.11.3 as new fastq files (with parameter --tag, --filter 3). The extracted paired-end reads were assembled using SPAdes v3.12.0 (42). These assemblies were known as original MAGs. A genome size of 5.5 Mb was set as a strict threshold: any assembly whose genome size was lower than this threshold was not considered in downstream analyses. FastANI was applied to calculate the ANI (cut-off 95 %) between MAGs and the three species reference genomes to double-check the predominant species in corresponding samples. The quality of each MAG was assessed using CheckM v1.0.18 (43).

A small number of the original MAGs were of high quality (44), but some contained a large number of contaminant contigs. All original MAGs were decontaminated as follows. Coding sequences of species-specific genomes in the pangenome were predicted using Prodigal v2.6.3 (Hyatt et al., 2010) with default settings and resulting multi-FASTA files containing protein sequences were concatenated to single files, which were used to build *K. oxytoca*-, *K. michiganensis*- and *K. grimontii*-specific databases in Diamond (45) format. Contigs of the original MAGs were aligned against the corresponding database under different minimum identity (%) cutoffs to report sequence alignments (with parameter --id 95, 96, 97, 98, 99 and 100). All unmapped contigs and contigs <500 nt in length were discarded from the original MAGs and the quality of new MAGs acquired was evaluated again using CheckM, to identify a Diamond BLAST identity threshold at which decontamination was effective while maintaining high genome completeness.

RESULTS AND DISCUSSION

Composition of the microbiota of preterm neonates

First faecal samples available after birth were collected from 109 hospitalized preterm infants (*n* = 50 female; *n* = 59 male) in the NICU of the NNUH (**Supplementary Table 1**; **Figure 1a**). On the basis of 16S rRNA gene sequencing data, *Enterobacteriaceae* were detected in the faeces of 42 (38.5 %) of the infants (**Figure 1b**).

All faecal samples were screened for *Enterobacteriaceae* using MacConkey agar no. 3 (**Figure 1c**). Forty-six (42.2 %) samples were positive for *Enterobacteriaceae* (*n* = 9 lactose-negative, carriage rate 8.3 %; *n* = 37 lactose-positive, carriage rate 33.9 %). Lactose-negative isolates were not characterized further. API 20E was used to provide tentative identification of LPE from 36 neonates (isolate could not be resuscitated for neonate P054) (**Figure 1d**). Of the 36 infants from whose faeces isolates were recovered, 23 were healthy, three had or were subsequently diagnosed with NEC, eight had suspected sepsis, one had an operation for gastroschisis and one was diagnosed with an eye infection after the faecal sample was taken (**Supplementary Table 1**).

Whole-genome sequencing of neonatal faecal LPE

Whole-genome sequences were obtained for 56 LPE. MetaPhlAn2.6 was used to assign identities to genomes (not shown). Among the 56 isolates sequenced, 20 were identified as *K. pneumoniae* (carriage rate 8.3 %; **Supplementary Table 2**), 14 were *Ent. cloacae* complex (carriage rate 11.9 %), 13 were *Esc. coli* (carriage rate 11.9 %), five were *K. oxytoca* (**Supplementary Table 2**), two were *Citrobacter freundii* (carriage rate 1.8 %; (46)), one was *Citrobacter murliniae* (carriage rate 0.9 %; (46)) and one was *R. ornithinolytica* (carriage rate 0.9 %; (47)). *Esc. coli* and *Ent. cloacae* complex isolates will be discussed in detail elsewhere. MetaPhlAn2.6-generated identities matched those given by API 20E (**Supplementary Table 2**).

Reference-based assembly of genomes was performed using BugBuilder (23) (**Supplementary Table 2**). To determine whether preterm neonates may harbour more than one strain of a species in their faecal microbiota, nine isolates (#64–#73) were collected from neonate P008. These had all been identified as *K. pneumoniae* by API 20E and genome data. ANI across the nine isolates was >99.99 %. To determine whether the isolates were identical, gene content analysis was performed using Roary (38). The average number of CDSs among these isolates was 5,385 (6.25) (**Supplementary Figure 1a–d**). Anvi'o showed the isolates were highly similar (**Supplementary Figure 1e**). Isolates of the same species from other neonates were also found to be identical to one another (#102 and #103 from P080; #118 and #119 from P124). For sets of identical isolates, only one was taken forward for further analyses. This left 14 distinct *Klebsiella* strains (9 *K. pneumoniae*; 5 *K. oxytoca*) for further analyses.

Genome analyses of *K. pneumoniae* strains

K. pneumoniae is a commensal of the human gut microbiota and can cause nosocomial infections, NEC and LOS in premature neonates (9,17,48–50),(51). The genetic backgrounds of the neonate isolates were explored, to determine virulence and AMR genes encoded within the strains' genomes.

Each isolate was genetically different: i.e. no two infants harboured the same strain of *K. pneumoniae* (**Supplementary Figure 2**). ANI analyses with representative strains of the seven phylogenetic groups of *K. pneumoniae* (52) showed eight of the neonatal isolates were *K. pneumoniae* (98.83–98.98 % ANI with *K. pneumoniae* ATCC 13883^T (GCA_000742135)) and one (#91) was *K. quasipneumoniae* (98.5 % ANI with *Klebsiella quasipneumoniae* subsp. *quasipneumoniae* 01A030^T (GCA_000751755)). MLST identified six STs within the *K. pneumoniae* strains (**Figure 2a**). *K. quasipneumoniae* #91 had a novel *mdh* allele, so no ST could be specified for this strain. None of the STs belonged to clonal complex (CC) 258, responsible for hospital outbreaks due to its frequent carriage of KPC and other acquired AMR genes (53). The capsule of *K. pneumoniae* and related species is considered one of its major virulence factors. K1, K2 and K5 capsular types and hypervirulent types have strong associations with human infectious diseases (54,55). None of our neonatal isolates had a capsular type commonly associated with infections or hypervirulent *K. pneumoniae*, though K7, K10, K11, K16 and K38 isolates have previously been recovered from clinical samples in Taiwan (56). Although the capsular type of strain #74 was identified as K62 with 99.33 % confidence and 100 % coverage, there was one gene (*KL62-12*, according to Kaptive) missing from the locus, leaving it a non-perfect match. Further analyses showed the genes associated with K62 to be disrupted in strain #74 and not encoded on a contiguous stretch of DNA (**Supplementary Figure 3**). Of the nine *K. pneumoniae* strains analysed using Kleborate (26,27), O1v1 and O1v2 were represented equally among the O-antigen types (n =4 for both). These can be distinguished using genomic data but are serologically cross-reactive (27). *K. quasipneumoniae* #91 was O3/O3a.

The vast majority (e.g. ~90 % in the NNUH NICU) of preterm infants receive antibiotics during their NICU stay, often started routinely from admission (i.e. 'covering') if they are born very premature and/or very low birth weight. Administration of antibiotics can lead to disruption of early colonization by microbes, potentially encouraging growth of opportunistic pathogens such as LPE, creating a selection pressure that may promote development of AMR. All strains encoded homologues of *acrB*, *acrD*, *marA*, *emrR*, *CRP*, *mdtB*, *mdtC*, *baeR*, *Escherichia mdfA*, PmrF, *msbA*, OmpK37, KpnE, KpnF, KpnG, KpnH and *Escherichia ampH* β-lactamase, associated with antibiotic efflux and its regulation (*acrB*, *acrD*, *marA*, *emrR*, *CRP*, *mdtB*, *mdtC*, *baeR*) (57–59) and resistance to: aminoglycosides; cationic antimicrobial peptides and antibiotics such as polymyxin (PmrF (60)); chloramphenicol (*mdfA* (61)), cefotaxime and cefoxitin (OmpK37); cefepime, ceftriaxone, colistin, erythromycin, rifampin, tetracycline, streptomycin as well as enhanced sensitivity toward sodium dodecyl sulfate, deoxycholate, dyes, benzalkonium chloride, chlorohexidine, and triclosan (KpnE, KpnF (62)); azithromycin, ceftazidime, ciprofloxacin, ertapenem, erythromycin, gentamicin, imipenem, ticarcillin, norfloxacin, polymyxin-B, piperacillin, spectinomycin, tobramycin, and streptomycin (KpnG, KpnH (63)); β-lactams and penicillin (*Escherichia ampH* β-lactamase (64)). As expected, core genes *bla*$_{\mathrm{SHV}}$ and *bla*$_{\mathrm{OKP}}$, respectively, were found in *K. pneumoniae* and *K. quasipneumoniae* genomes (53). *K. quasipneumoniae* #91 also encoded homologues of the acquired

AMR gene *emrB* (a translocase that recognizes substrates including carbonyl cyanide *m*-chlorophenylhydrazone, nalidixic acid and thioloactomycin) and *bacA* (confers resistance to bacitracin) (**Figure 2a**). Plasmid-encoded *bla*$_{\mathrm{SHV}}$ enzymes represent an important subgroup of class A β-lactamases, while chromosomally encoded β-lactamase *bla*$_{\mathrm{OKP}}$ cannot hydrolyse extended-spectrum cephalosporins (65). Homologues of the core AMR genes *oqxA* and *oqxB* (encoding OqxAB, a plasmid-encoded efflux pump that confers resistance to fluoroquinolones) were encoded by #64, #74, #91, #95, #115 and #118. Strains #64 and #95 encoded homologues of FosA6, while #74, #85, #92, #115 and #118 encoded homologues of FosA5 (both gene products confer resistance to fosfomycin, and are core AMR genes (53)).

While the majority of the neonatal *K. pneumoniae* strains did not represent known pathogenic lineages, virulence factors were detected in their genomes using VFDB (**Figure 2b**). The host limits iron availability within the GI tract to prevent colonization by pathogens and bacterial overgrowth. However, *Klebsiella* spp. have evolved numerous mechanisms to circumvent these defences. Thus, we determined whether gene clusters associated with iron uptake and siderophore systems (i.e. enterobactin, yersiniabactin, aerobactin, colibactin, salmochelin) were present in the strains. All strains encoded enterobactin, while only #115 encoded an additional system (yersiniabactin) (**Figure 2b**). All strains encoded *Esc. coli* common pilus, OmpA, Hsp60, type 3 fimbriae, ferric aerobactin receptor *IutA* and the AcrAB efflux pump. All strains except #102 encoded type 1 fimbriae; all strains except *K. quasipneumoniae* #91 encoded typical *K. pneumoniae* lipopolysaccharide (LPS) according to VFDB (>70 % amino acid identity and 90 % query coverage). *K. quasipneumoniae* #91 encoded thin aggregative fimbriae, associated with biofilm formation and adhering to human mucosal or epithelial surfaces. Incomplete coverage of *Klebsiella* capsule genes is likely due to the limited database of VFDB compared with those used to populate Kaptive and Kleborate. Kaptive had shown #91 to be K11 and O3/O3a. The core LPS region of #91 was identified using the *waa* gene cluster (66); WaaL clustering with an 80 % threshold showed the strain had LPS core type 1 (67) **(Supplementary Figure 4)**.

Whole-genome analyses of isolates tentatively identified as *K. oxytoca*

K. oxytoca is a minor member of the human gut microbiota, recovered at low levels from the faeces of 1.6–9 % of healthy adults (68). Toxigenic *K. oxytoca* is a causative agent of antibiotic-associated haemorrhagic colitis, a condition affecting mainly young and otherwise healthy outpatients after brief treatment with penicillin derivatives (69). *K. oxytoca* has been detected in the faeces of a subset of preterm infants via cultivation or shotgun metagenomics, but its association with preterm-associated infections is unknown (1,17,70). At the DNA level, bacteria characterized phenotypically as *K. oxytoca* actually represent six phylogroups/distinct species: Ko1, *K. michiganensis*; Ko2, *K. oxytoca*; Ko3, '*Klebsiella spallanzanii*'; Ko4, '*Klebsiella pasteurii*'; Ko6, *K. grimontii*; Ko8, *K. huaxiensis* (71–74)(75). *K. michiganensis* and *K. oxytoca* are distinguishable based on the *bla*$_{\text{OXY}}$ gene they carry (*bla*$_{\text{OXY-1}}$ and *bla*$_{\text{OXY-2}}$, respectively) (71). *K. grimontii* was recently described to accommodate Ko6 strains based on *rpoB*, *gyrA* and *rrs* gene sequences (73). All six members of the complex can be differentiated by MALDI-TOF (75), but reference databases currently in routine clinical use lack reference spectra of the different species to allow identification beyond *K. oxytoca*. Consequently, reports on complex members other than *K. oxytoca* have only recently begun to appear in the literature (73–77). The colonization of humans with *K. oxytoca* phylogroups has previously been associated with the genetic backgrounds of strains: Ko2 mainly inhabits the lower GI tract, with Ko1 and Ko6 generally associated with respiratory isolates and faecal isolates, respectively (72).

On the basis of API 20E data and initial genome (MetaPhlAn2.6) analysis, five neonatal isolates were identified as *K. oxytoca* (#80, #83, #88, #99, #108). It has recently been shown that API 20E and MALDI-TOF using current clinical reference databases are as effective as one another for characterization of complex members as *K. oxytoca* (78). MetaPhlAn2.6 cannot distinguish among species of the *K. oxytoca* complex. ANI of the genomes against reference genomes showed #88 and #108 to be *K. michiganensis* (both 98.78 and 98.94 % ANI, respectively with GCA_002925905) and #80, #83 and #99 to be *K. grimontii* (99.18, 99.23 and 99.17 % ANI, respectively, with GCA_900200035) (**Supplementary Figure 5a**), with ANI cut-off values well above the ~95 % proposed for species delineation (79–81) and used by Passet & Brisse (73) to separate *K. grimontii* from *K. oxytoca* and *K. michiganensis*. Phylogenetic analysis with a panel of authentic *K. oxytoca*, *K. grimontii* and *K. michiganensis* genomes confirmed the species affiliations of the infant isolates

(**Supplementary Figure 5b**). Similar to the *K. pneumoniae* isolates, no two infants harboured the same strain of *K. michiganensis* or *K. grimontii* (**Figure 3a**).

It is notable that of the publicly available genomes deposited as *K. oxytoca*, 74 were found to represent *K. michiganensis*, 64 were *K. oxytoca* and 24 were *K. grimontii*. This suggests that *K. michiganensis* may be more clinically relevant than *K. oxytoca sensu stricto*. *K. michiganensis* was originally proposed to describe an isolate closely related to *K. oxytoca* recovered from a toothbrush holder (74). The bacterium is now recognised as an emerging pathogen, with this recognition due to improved genomic characterization of clinical isolates that would have previously been described as *K. oxytoca* based on simple phenotypic tests or MALDI-TOF (78,82–84).

Predicted virulence and AMR determinant genes of infant-associated *K. michiganensis* and *K. grimontii*

The *K. michiganensis* and *K. grimontii* strains were examined for the presence of virulence-associated loci found in *K. pneumoniae* strains (53) (**Figure 3b**). Enterobactin was encoded by all strains. Yersiniabactin was predicted to be encoded by *K. grimontii* #99 and *K. michiganensis* #88 and #108. Other siderophore-associated gene clusters (aerobactin, colibactin and salmochelin) found in *K. pneumoniae* were absent. An allantoinase gene cluster (including *allB*/*C*/*R*/*A*/*S* and *ybbW*), which plays a role in *K. pneumoniae* liver infection (85), was identified in the three *K. grimontii* strains.

Due to the clinical importance of AMR in *Enterobacteriaceae*, an *in silico* AMR gene profile was established for the *K. michiganensis* and *K. grimontii* strains. Homologues of 18 AMR determinant genes (acquired AMR genes – *emrB*, *emrR*; core AMR genes *acrB*, *acrD*, CRP, *marA*, *mdtB*, *mdtC*, *baeR*, FosA5, *pmrF*, *oqxA*, *oqxB*, *msbA*, KpnE, KpnF, KpnG, *Escherichia ampH* β-lactamase) were common to the five strains, similar to the *K. pneumoniae* isolates. Both *K. michiganensis* strains encoded homologues of OXY-1-2 (β-lactamase specific to *K. michiganensis* (Ko1; (86)) and *bacA*, while #108 encoded a homologue of *aph(3')-la* (aminoglycoside phosphotransferase). All *K. grimontii* strains encoded *mdtN* (potentially involved in resistance to puromycin, acriflavine and tetraphenylarsonium chloride), while #83 and #99 encoded homologues of OXY-6-2 (β-lactamase specific to *K. grimontii* (Ko6; (86)).

Phenotypic characterization of *K. pneumoniae, K. quasipneumoniae, K. michiganensis* and *K. grimontii* neonatal isolates

Five of the 13 *Klebsiella* strains we characterised were isolated from preterm infants who had been diagnosed with either NEC or sepsis. Thus, we sought to link our genotypic analyses with clinically important virulence traits including the ability to survive and replicate in host immune cells (i.e. macrophages) and the ability to produce iron-acquiring siderophores. We also determined the strains' AMR profiles for a limited set of antimicrobials.

Previous studies have indicated that respiratory infection-associated *K. pneumoniae* are able to survive within macrophages, a critical innate immune cell type required for optimal pathogen clearance (87). However, to date there is limited information relating to this ability in gut-associated strains, and there is no information on other *Klebsiella* species. Thus, all *Klebsiella* strains isolated in this study were tested in PMA-differentiated THP-1 macrophages using a gentamicin protection assay. All strains appeared to persist within macrophages, as bacterial load was either maintained over the time-course or increased or decreased between 1.5 h and 6 h, although these values were not statistically significant (**Figure 4a**). These data suggest that all *Klebsiella* strains tested can reside and persist in macrophages. This ability of all strains to survive, and in some cases potentially replicate, within macrophages indicates their immune evasion capabilities, which may link to increased risk and incidence of NEC and sepsis if these strains translocate from the 'leaky' preterm GI tract to systemic sites contributing to the inflammatory cascades characteristic of these conditions.

Iron is a vital nutrient that performs multiple roles in cellular processes, ranging from DNA replication and cell growth to protection against oxidative stress. In the healthy host, the majority of iron is bound with intracellular proteins and the remaining free iron is extracellular and insoluble, hence difficult to access (88). For invading pathogens, siderophore systems are critical for iron competition and uptake to accomplish colonization and cause infections, this is particularly true in the preterm GI tract. Preterm infants in NICU are heavily supplemented with iron as they receive many red-blood-cell transfusions (increasing hepatic iron stores), iron-supplemented parenteral nutrition, and supplementary oral iron within a few weeks of birth. During infection, *Klebsiella* secretes siderophores to sequester iron and to establish colonization in the host. Enterobactin is the most well-known siderophore produced by *K. pneumoniae* and related species, and was found to be encoded by all our strains (**Figure 2b**, **Figure 3b**). The host innate immune protein lipocalin 2 binds to enterobactin and disrupts bacterial iron uptake (89). *Klebsiella* species have evolved to hoodwink this host response by producing several evasive siderophores (90,91). Siderophore production of *Klebsiella* isolates was monitored using CAS liquid assay. All isolates tested grown in M9 minimal medium were CAS-positive with the estimated siderophore concentration ranging between of 3.5 and 6 nM (**Figure 4b**). There was no significant difference in siderophore production between ‘healthy’ and NEC- and sepsis-associated isolates.

Klebsiella is of concern within an AMR context, particularly in at-risk neonates, due to the increasing emergence of multidrug-resistant isolates that cause severe infection (92). A UK study in which 24 % of all LOS cases were caused by *Enterobacteriaceae* (8.9 % of all caused by *Klebsiella* spp.) showed a high proportion (14 % and 34 %, respectively) of *Enterobacteriaceae* isolates recovered from sick infants were resistant to flucloxacillin/gentamicin and amoxicillin/cefotaxime, the two most commonly used empiric antibiotic combinations (51). Thus, to demonstrate antibiotic-resistance phenotypes in *Klebsiella* spp. correlating to presence of AMR genotypes, we tested the susceptibility of the isolates with three antibiotics commonly prescribed in NICUs; gentamicin, meropenem and benzylpenicillin (**Table 1**). One strain of *K. grimontii* (#80) was potentially sensitive to benzylpenicillin, an aminopenicillin currently not recognized as being clinically useful against *Enterobacteriaceae*.

Isolates #64, #83 and #108 (all encoding KpnG and KpnH; **Figures 2a, 3b**) were resistant to gentamicin, while #80, #88, #95 and #99 (all encoding KpnG and KpnH; **Figures 2a, 3b**) showed intermediate susceptibility to this aminoglycoside.

Presence of a gene in a bacterium’s genome does not mean it is functionally active, nor does it give any indication as to how active the gene is if it is indeed functional: e.g. all nine *K. pneumoniae* isolates encoded KpnG and KpnH (strict CARD matches**; Figure 2a**), but only two showed any resistance to gentamicin upon susceptibility testing.

Isolates #64 (SHV-1), #74 (SHV-36), #85, #88, #91 (OKP-A-2), #92 (SHV-36), #95 (SHV-28), #102 (SHV-164), #115 (SHV-36) and #118 (SHV-36) – which all encoded extended-spectrum β-lactamases (SHV) or *Escherichia ampH* β-lactamase (OKP-A-2) but lacked OmpK35 and OmpK36 – showed intermediate susceptibility to the carbapenem meropenem. Loss of the two porins OmpK35 and OmpK36 is known to confer resistance to carbapenems in strains producing extended-spectrum β-lactamases or plasmid-mediated AmpC-type β-lactamases (93). *K. pneumoniae* #64 was isolated from an infant with clinically diagnosed NEC with confirmed *Klebsiella* colonization. Importantly this preterm infant had previously been treated with benzylpenicillin, gentamicin and meropenem, which may link to the observed phenotypic resistance and corresponding AMR genes *Escherichia ampH* β-lactamase, *bla*$_{\text{SHV-1}}$ and KpnG/KpnH and SHV-1, and lack of OmpK35 and OmpK36, and suggests further treatment with gentamicin and meropenem would have been ineffective in this infant. Indeed, the infant was treated with cefotaxime, metronidazole and vancomycin in a subsequent round of medication (**Supplementary Table 2**). *K. pneumoniae* #115 was isolated from a baby that had confirmed NEC: the strain was resistant to benzylpenicillin (encoded *Escherichia ampH* β-lactamase and *bla*$_{\text{SHV-36}}$) and showed intermediate resistance to meropenem (lacked OmpK35 and OmpK36). *K. michiganensis* #88, also isolated from a baby that had NEC, showed intermediate resistance to both benzylpenicillin (*bla*$_{\text{OXY-1-2}}$, perfect CARD match; **Figure 3b**) and meropenem (lacked OmpK35 and OmpK36): both antibiotics had been administered to the baby at birth. *K. grimontii* #99, isolated from a baby with suspected sepsis, showed intermediate resistance to benzylpenicillin (encoded *Escherichia ampH* β-lactamase and *bla*$_{\text{OXY-6-2}}$). These data indicate that preterm-associated *Klebsiella* have a multi-drug-resistant phenotype that may prove problematic when treatment options are required for sepsis or NEC. Interestingly, other isolates (e.g. #95, recovered from an infant who had received benzylpenicillin and gentamicin; **Supplementary Table 2**) associated with ‘healthy’ preterm infants also harboured AMR genes (#95: **Figure 2a**) and phenotypic resistance profiles suggesting that administration of antibiotics to preterm infants with no signs of clinical infection contributes to the reservoir of AMR genes – the ‘resistome’ (94) – which may increase horizontal gene transfer of AMR determinants to other opportunistic pathogens residing within the GI tract.

Abundance of *K. oxytoca* and related species in metagenomic datasets

We used a published metagenomics dataset (1) to determine the prevalence of *K. oxytoca*, *K. michiganensis* and *K. grimontii* in the preterm infant gut microbiome. These data had previously been used to look at the relationship between NEC and uropathogenic *Esc. coli*, and metadata were available for the samples. Ward *et al.* (1) collected a total of 327 samples at three stages of infant life: stage1, days 3–9; stage2, days 10–16; stage3, days 17–22. Within each life stage samples were collected on more than one day for some infants. In the current study, only samples processed under Protocol A of Ward *et al.* (1) and from the earliest collection day within each life stage were analysed. For those samples for which multiple sets of paired-end data were available, read data were concatenated and used in analyses (**Supplementary Table 3**).

Stage1 comprised samples from 127 infants (105 preterm, 22 term), 16 of whom had been diagnosed with NEC and 10 infants had subsequently died. Stage2 comprised samples from 146 infants (128 preterm and 18 term), 24 of whom later developed NEC with 18 deaths. Stage3 comprised samples from 54 infants (48 preterm, 6 term), including eight NEC patients, six of whom died. Samples were collected from 165 distinct infants (143 preterm, 22 term) but only 41 of them were sequenced at all three life stages. Infants were born either vaginally (*n* = 70) or by caesarean section (*n* = 95). The gestational ages of preterm infants ranged from 23 to 29 weeks (mean 26.1 weeks), while the term babies ranged from 38 to 41 weeks (mean 39.2 weeks).

As we had found that the MetaPhlAn2.6 database contained non-*K. oxytoca* genomes within its *K. oxytoca* dataset (detecting *K. oxytoca*, *K. michiganensis* and *R. ornithinolytica* (refer to Methods)), we used Centrifuge to determine abundance of this species in metagenomes (**Figure 5ab**). Due to their genomic similarity, *K. oxytoca* and *K. michiganensis* could not be readily distinguished using Centrifuge (**Figure 5a**); no genomes assigned to *K. grimontii* were included in the Centrifuge database at the time this study was undertaken. Though it should be noted that, while Centrifuge (and Kraken2) relies on NCBI taxonomy for species identification, there are still many genomes within

GenBank/RefSeq that are assigned to the wrong species (e.g. assemblies GCA_001052235.1 and GCA_000427015.1 within our curated pangenome dataset have been confirmed by detailed analyses to be *K. grimontii* (**Supplementary Figure 2** and (36)), but still assigned as *K. oxytoca* and *K. michiganensis*, respectively, within GenBank as of 28 July 2019; these are by no means the only examples from our current study).

For those samples harbouring *K. oxytoca*, relative abundance of the bacterium increased from stage1 to stage2 and decreased at stage3 (**Figure 5b**).

Metapangenome analysis of preterm infant metagenomic data to detect *K. oxytoca*, *K. michiganensis* and *K. grimontii*

Using a set of 162 *K. oxytoca*-related genomes (**Supplementary Table 4**) and those of the five infant isolates, a pangenome was generated using Roary. The pangenome dataset consisted of 76 *K. michiganensis* (mean ANI among strains 98.55 (0.60) %, range 97.13–100 %), 64 *K. oxytoca* (mean ANI among strains 99.20 (0.30) %, range 98.53–100 %) and 27 *K. grimontii* (mean ANI among strains 98.45 (1.47) %, range 95.70–100 %) strains. A total of 40,605 genes were detected in the open pangenome: 2,769 of them constituted the core gene cluster, while the accessory cluster included 5,108 genes and the remaining 32,728 genes formed the strain-specific cluster. A PCA plot based on the accessory genes clustered strains into the three different species (**Figure 5c**), in agreement with our phylogenetic analysis of the core genes (**Supplementary Figure 5b**) and consistent with the findings of (76), who were able to split the three species (phylogroups) based on a pangenome analysis of fewer genomes.

The Roary-generated pangenome was used as a custom database for PanPhlAn, to detect the presence and absence of core genes and accessory genes in each infant sample. As expected, the proportion of reads that PanPhlAn mapped to the custom database correlated with the Centrifuge-generated abundance data (**Figure 5d**). In stage1, 13 infants (12 preterm, 1 term; 10.2 % of all stage1 samples) were predicted to carry *K. oxytoca*-related species (**Figure 5e**); in stage2, the number was 24 (22 preterm, 2 term; 16.4 % of all stage2 samples) (**Figure 5f**); in stage3 infants, the rate of carriage was much lower, with only three infants (all preterm; 5.6 % of all stage3 samples) potentially harbouring target species (**Figure 5g**). The change in prevalence of *K. oxytoca*-related species across the three stages based on the PanPhlAn analysis was consistent with the *K. oxytoca* abundance data generated with Centrifuge (**Figure 5b**).

The pangenome accessory genes were used in PCA to define which species strains detected by PanPhlAn belonged to. In stage1, samples from six preterm infants (P10111, P10141, P10301, P10451, P11292, P12121) and one term infant (P30221) harboured *K. michiganensis* (**Figure 5e**). In stage2, samples from four preterm infants (P10471, P10472, P11311, P11701) harboured *K. oxytoca*, while those from 14 preterm infants (P10071, P10231, P10301, P10441, P10451, P10501, P10601, P11151, P11202, P11291, P11292, P12121, P12641, P12651) and one term infant (P30221) harboured *K. michiganensis* (**Figure 5f**). Four samples (P11351, P12621, P20241, P30141) could not be assigned a species, while the sample from P11981 located close to *K. oxytoca* (**Figure 5f**). Similarly, two of the stage3 samples from P12221 and P12651 carried *K. michiganensis*, while P11981 located near *K. oxytoca* (**Figure 5g**).

Recovery of *K. oxytoca* and *K. michiganensis* MAGs from metagenomes

Since the abundance of *K. oxytoca*-related species was considerable in some infant samples, we attempted to obtain high-quality MAGs directly from these metagenomes and to assign them to the species. The metagenomic samples were checked and their reads aligned against those of the 167-genome database; reads that mapped were extracted and assembled as ‘original MAGs’. The genome sizes of the 167 genomes ranged from 5.72 Mb to 7.23 Mb (mean 6.35 Mb), thus a genome size of at least 5.5 Mb was used to define a likely complete MAG. After assembly of the reads that mapped to our database, MAGs were generated from the stage 1 (_s1), stage2 (_s2) and stage3 (_s3) samples. ANI and phylogenetic analyses showed these MAGs to be *K. michiganensis* (P10301_s1, P10451_s1, P11292_s1, P12121_s1, P30221_s1, P10071_s2, P10301_s2, P10441_s2, P10451_s2, P10501_s2, P10601_s2, P11151_s2, P11202_s2, P11291_s2, P11292_s2, P12121_s2, P12641_s2, P12651_s2, P30221_s2, P12221_s3 and P12651_s3; mean ANI with GCA_002925905 of 98.61 ± 0.66 %) or *K. oxytoca* (P10472_s2, P11311_s2, P11701_s2, P11981_s2; mean ANI with GCA_900083895 of 99.30

± 0.26 %) (36). No *K. grimontii* MAGs were recovered from any samples.

Prior to checking the completeness and contamination of the *K. michiganensis* and *K. oxytoca* MAGs, contigs <500 nt in length were removed from the assemblies. A high-quality MAG requires a >90 % genome completeness with contamination <5 % (44). According to CheckM results, three stage1 MAGs (P10301_s1, P10451_s1, P30221_s1), seven stage2 MAGs (P10441_s2, P10451_s2, P10501_s2, P10601_s2, P11291_s2, P11292_s2, P30221_s2) and one stage3 MAG (P12221_s3) were of high quality. The rest of the MAGs were ≥90 % complete, but were contaminated (e.g. P12621_2 contained 274.59 % contamination). Thus, we attempted to decontaminate the MAGs using a Diamond BLAST-based approach.

Since we already knew the species each MAG belonged to from PCA and PanPhlAn analyses, scaffolds were mapped against the relevant species-specific genome database under different minimum BLAST identity to report alignments, which were used to generate 'cleaner' MAGs. **Supplementary Figure 6** shows the change in genome completeness and the percentage contamination of stage2 MAGs when different blast identities were applied. The changes were negligible for those MAGs with high-quality-level completeness and lacking contaminants even when the cut-off was set at 99 %. For contaminated MAGs, the percentage contamination decreased markedly as the BLAST identity became stricter and reduced to the bottom when all scaffolds in that MAG could be aligned with 100 % identity. However, 100 % was not an appropriate threshold as the genome completeness was affected greatly at this point (**Supplementary Figure 6**). Instead, a cut-off of 99 % was used to decontaminate MAGs because 13 high-quality level MAGs and 5 medium-quality level MAGs could be obtained when using this identity threshold. Stage2 MAGs that passed PanPhlAn, PCA and ANI analysis reached at least reach medium-quality level using a 99 % identity threshold. This cut-off was also suitable for stage1 MAGs, the quality of which was high. However, for stage3 MAGs, the percentage contamination from P11981_s3 decreased to medium-quality level only at 100 % identity, at which time the genome completeness fell down to 42.40 %. After evaluating their genome completeness and contamination levels, a total of 25 MAGs (**Table 2**) were assessed further.

The presence of tRNAs for the standard 20 amino acids and rRNA was examined as a secondary measure of genome quality. A high-quality MAG requires at least 18 of the 20 possible amino acids (44). P11981_s2 (16 aa) and P12651_s3 (17 aa) had to be classified as medium-quality MAGs even though their genome completeness and contamination reached high-quality levels. 16S rRNA genes were detected in all MAGs except P11151_s2, which was subsequently classified as medium quality. Taking mandatory genome information into consideration (44), a total of 19 high-quality and six medium-quality MAGs were recovered (**Table 2**); the sequences of these MAGs are available from figshare. All of the MAGs had ≥15 standard tRNAs. High-quality MAGs had tRNAs that encoded an average of 19.6 (0.7) of the 20 amino acids, some of them even had a tRNA that encodes an additional amino acid SeC, while medium-quality MAGs had 18 (1.4) basic amino acids encoded by tRNAs. High-quality MAGs consisted of ≤500 scaffolds in 52.6 % of cases (mean 600) and had an average N50 of 121 kb, while only one medium-quality MAG comprised ≤500 scaffolds (mean 1174) and the average N50 was less than half of that of high-quality MAGs (48.3 kb).

Genotyping of MAGs

Comparison of the sequences of the MAGs showed each infant harboured a different strain of *K. oxytoca* (**Figure 6a**) or *K. michiganensis* (**Figure 6b**). In infants where MAGs were recovered across different life stages, the MAGs were highly similar to one another (**Figure 6b**). Similar to what we had seen with our isolates, the MAGs encoded a range of β-lactamase and virulence genes (**Supplementary Figure 7**). It was also notable that two of the MAGs (*K. michiganensis* 10071_s2, *K. oxytoca* 10472_s2) encoded *mcr-9* (perfect match), a plasmid-mediated colistin resistance gene and phosphoethanolamine transferase. However, as noted above for our isolate work, presence of the aforementioned genes in MAGs does not mean they were functionally active in the infants' GI tracts. All the *K. michiganensis* MAGs encoded the siderophore enterobacterin, along with all but one (11981_s2) of the *K. oxytoca* MAGs. The allantoinase gene cluster associated with liver infection was detected in the four *K. oxytoca* MAGs, but only a third of the *K. michiganensis* MAGs. We only detected this cluster in the *K. grimontii* isolates we recovered (**Figure 3b**).

MLST analysis assigned 10 MAGs to eight known STs (**Table 2**). We believe insufficient sequence coverage meant we were unable to ST more MAGs: e.g. *gapA* of MAG P11981_s2 aligned exactly with *gapA* allele 2 sequence but it was only partial, leaving an incomplete match.

The STs of all genomes in the curated genome dataset were also identified (**Supplementary Table 4**). Due to our limited understanding of *K. oxytoca*-related species, many combinations of alleles have not been assigned corresponding STs yet, especially for the newly described species *K. grimontii* (73). MLST identification showed that 54/64 *K. oxytoca sensu stricto* strains had known STs, with some STs being more dominant than others. ST2, the most prevalent ST and represented by 15 strains, belonged to CC2. ST18 and ST19 were also in CC2, with both represented by two strains. ST199 and ST176 were the second and third most common STs, respectively. Among *K. michiganensis* strains, 48/76 of them could be assigned a known ST and comprised 21 distinct STs, with nine represented by more than one isolate. ST11, ST27, ST50, ST85, ST143 and ST202 were the most frequent STs, all of which were represented by at least four strains. *K. michiganensis* #108 was ST157. Only 5/27 *K. grimontii* strains could be assigned an ST (#83 – ST72; #99 – ST76; 10-5250 – ST47, GCA_000247915.1; 1148_KOXY – ST186, GCA_001052235.1; M5al – ST104, GCA_001633115.1).

SUMMARY

Klebsiella spp. encode numerous virulence and antibiotic resistance genes that may contribute to the pathogenesis of NEC and LOS. In this study, we characterized nine *K. pneumoniae*, three *K. grimontii* and two *K. michiganensis* strains isolated prospectively from the faeces of a UK cohort of preterm infants, and have shown these gut isolates are able to reside and persist in macrophages, suggesting they can evade the immune system. These isolates will be used in future studies aiming to replicate aspects of NEC and sepsis in model systems to confirm the role of *Klebsiella* spp. in these diseases.

We have shown that mis-annotated genomes are being used in bioinformatics tools routinely used to characterize the human gut microbiome. By using a carefully curated dataset to undertake metapangenome analyses of the closely related species *K. oxytoca*, *K. michiganensis* and *K. grimontii*, we have demonstrated that *K. michiganensis* is likely to be more clinically relevant to a subset of preterm infants than *K. oxytoca*. Identity of publicly available genomes should be confirmed upon download and linked to accurate taxonomic frameworks prior to analyses of data, especially when attempting to identify and type closely related species in metagenomic data.

AUTHOR STATEMENTS

Authors and contributions

LH and LJH conceived and designed the study. All authors contributed to the writing of the manuscript. TCB did the isolation and initial characterization work, and isolated DNA from bacteria. YC and LH did bioinformatics associated with genome analyses. OL did the anvi'o work. TCB, YC and OL were supervised by LH. LJH is chief investigator for the preterm clinical study from which samples were used, and PC is the clinical lead for the study. CAG performed 16S rRNA gene library preparation from faecal samples, and determined MIC values for strains. IO did macrophage assays. CZS did iron assays, and additional siderophore bioinformatics analysis. SC performed the 16S rRNA gene-associated bioinformatics, SP did 16S rRNA gene sequence analyses, and clinical database management. LJH supervised CA, IO, CZS, SC and SP.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Funding information

Infrastructure support was provided by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre (BRC). This work used the computing resources of the UK MEDical BIOinformatics partnership – aggregation, integration, visualisation and analysis of large, complex data (UK Med-Bio) – which was supported by the Medical Research Council (grant number MR/L01632X/1), CLIMB [76] and the Nottingham Trent University Hamilton High Performance Computing Cluster. TCB was funded by a University of Westminster Faculty PhD studentship, and a Research Visit Grant from the Microbiology Society (RVG16/3). This work was funded via a Wellcome Trust Investigator Award to LJH (100/974/C/13/Z), and support of the BBSRC Norwich Research Park Bioscience Doctoral Training Grant (BB/M011216/1, supervisor LJH, student CAG), and the Institute Strategic Programme Gut Microbes and Health BB/R012490/1, and its constituent project(s) BBS/E/F/000PR10353 and BBS/E/F/000PR10356 to LJH.

Ethics approval

The Ethics Committee of the Faculty of Medical and Health Sciences of the University of East Anglia (Norwich, UK) approved this study. The protocol for faeces collection was laid out by the Norwich Research Park (NRP) Biorepository (Norwich, UK) and was in accordance with the terms of the Human Tissue Act 2004 (HTA), and approved with licence number 11208 by the Human Tissue Authority. Infants admitted to the NICU of the NNUH were recruited by doctors or nurses with informed and written consent obtained from parents.

Consent for publication

All authors approved submission of the manuscript for publication.

Acknowledgements

LH is a member of the ESGHAMI study group (https://www.escmid.org/research_projects/study_groups/host_and_microbiota_interaction/). We thank neonatal research nurses Karen Few, Hayley Aylmer and Kate Lloyd for obtaining parental consents and for collecting the samples at NNUH with the kind assistance of the clinical nursing team. We thank Matthias Scholz for his help in explaining the theory behind PanPhlAn to YC and advice on how to set appropriate parameters.

REFERENCES

1. Ward DV, Scholz M, Zolfo M, Taft DH, Schibler KR, Tett A, et al. Metagenomic sequencing with strain-level resolution implicates uropathogenic *E. coli* in necrotizing enterocolitis and mortality in preterm infants. Cell Rep. 2016 Mar 29;14(12):2912–24.
2. Korpela K, de Vos WM. Early life colonization of the human gut: microbes matter everywhere. Curr Opin Microbiol. 2018 Aug 4;44:70–8.

3. Newburg DS, Walker WA. Protection of the neonate by the innate immune system of developing gut and of human milk. Pediatr Res. 2007 Jan;61(1):2–8.

4. Wandro S, Osborne S, Enriquez C, Bixby C, Arrieta A, Whiteson K. The microbiome and metabolome of preterm infant stool are personalized and not driven by health outcomes, including necrotizing enterocolitis and late-onset sepsis. mSphere. 2018 Jun 27;3(3).

5. Wang Y, Hoenig JD, Malin KJ, Qamar S, Petrof EO, Sun J, et al. 16S rRNA gene-based analysis of fecal microbiota from preterm infants with and without necrotizing enterocolitis. ISME J. 2009 Aug;3(8):944–54.

6. DiBartolomeo ME, Claud EC. The developing microbiome of the preterm infant. Clin Ther. 2016;38(4):733–9.

7. Magne F, Abély M, Boyer F, Morville P, Pochart P, Suau A. Low species diversity and high interindividual variability in faeces of preterm infants as revealed by sequences of 16S rRNA genes and PCR-temporal temperature gradient gel electrophoresis profiles. FEMS Microbiol Ecol. 2006 Jul;57(1):128–38.

8. Mai V, Torrazza RM, Ukhanova M, Wang X, Sun Y, Li N, et al. Distortions in development of intestinal microbiota associated with late onset sepsis in preterm infants. PloS One. 2013;8(1):e52876.

9. Shaw AG, Sim K, Randell P, Cox MJ, McClure ZE, Li M-S, et al. Late-onset bloodstream infection and perturbed maturation of the gastrointestinal microbiota in premature infants. PloS One. 2015;10(7):e0132923.

10. Warner BB, Tarr PI. Necrotizing enterocolitis and preterm infant gut bacteria. Semin Fetal Neonatal Med. 2016;21(6):394–9.

11. Robertson C, Savva G, Clapuci R, Jones J, Maimouni H, Brown E, et al. Incidence of necrotising enterocolitis before and after introducing routine prophylactic *Lactobacillus and Bifidobacterium* probiotics. Arch Child Fetal Neonatal Ed [Internet]. 2019 [cited 2019 Dec 3]; Available from: https://www.ncbi.nlm.nih.gov/pubmed/31666311

12. Lu L, Claud EC. Intrauterine inflammation, epigenetics, and microbiome influences on preterm infant health. Curr Pathobiol Rep. 2018;6(1):15–21.

13. Adeolu M, Alnajar S, Naushad S, S Gupta R. Genome-based phylogeny and taxonomy of the '*Enterobacteriales'*: proposal for *Enterobacterales* ord. nov. divided into the families *Enterobacteriaceae*, *Erwiniaceae* fam. nov., *Pectobacteriaceae* fam. nov., *Yersiniaceae* fam. nov., *Hafniaceae* fam. nov., *Morganellaceae* fam. nov., and *Budviciaceae* fam. nov. Int J Syst Evol Microbiol. 2016 Dec;66(12):5575–99.

14. Pammi M, Cope J, Tarr PI, Warner BB, Morrow AL, Mai V, et al. Intestinal dysbiosis in preterm infants preceding necrotizing enterocolitis: a systematic review and meta-analysis. Microbiome. 2017 09;5(1):31.

15. Olm MR, Bhattacharya N, Crits-Christoph A, Firek BA, Baker R, Song YS, et al. Necrotizing enterocolitis is preceded by increased gut bacterial replication, *Klebsiella*, and fimbriae-encoding bacteria that may stimulate TLR4 receptors. bioRxiv. 2019 Feb 22;558676.

16. Claud EC, Keegan KP, Brulc JM, Lu L, Bartels D, Glass E, et al. Bacterial community structure and functional contributions to emergence of health or necrotizing enterocolitis in preterm infants. Microbiome. 2013 Jul 10;1(1):20.

17. Sim K, Shaw AG, Randell P, Cox MJ, McClure ZE, Li M-S, et al. Dysbiosis anticipating necrotizing enterocolitis in very premature infants. Clin Infect Dis Off Publ Infect Dis Soc Am. 2015 Feb 1;60(3):389–97.

18. Torrazza RM, Ukhanova M, Wang X, Sharma R, Hudak ML, Neu J, et al. Intestinal microbial ecology and environmental factors affecting necrotizing enterocolitis. PloS One. 2013;8(12):e83304.

19. Alcon-Giner C, Dalby MJ, Caim S, Ketskemety J, Shaw A, Sim K, et al. Microbiota supplementation with *Bifidobacterium* and *Lactobacillus* modifies the preterm infant gut microbiota and metabolome. bioRxiv. 2019 Jul 12;698092.

20. Kiu R, Caim S, Alcon-Giner C, Belteki G, Clarke P, Pickard D, et al. Preterm infant-associated *Clostridium tertium, Clostridium cadaveris,* and *Clostridium paraputrificum* strains: genomic and evolutionary insights. Genome Biol Evol. 2017 01;9(10):2707–14.

21. Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010 Jan

22;327(5964):469–74.

22. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012 Aug;9(8):811–4.

23. Abbott JC. BugBuilder - an automated microbial genome assembly and analysis pipeline. bioRxiv. 2017 Jun 11;148783.

24. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018 30;9(1):5114.

25. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 2018;3:124.

26. Lam M, Wyres KL, Duchêne S, Wick RR, Judd LM, Gan Y-H, et al. Population genomics of hypervirulent *Klebsiella pneumoniae* clonal-group 23 reveals early emergence and rapid global dissemination. Nat Commun [Internet]. 2018 Jul 13 [cited 2018 Aug 14];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6045662/

27. Wick RR, Heinz E, Holt KE, Wyres KL. Kaptive Web: User-Friendly Capsule and lipopolysaccharide serotype prediction for *Klebsiella* genomes. J Clin Microbiol. 2018 Jun;56(6).

28. Wyres KL, Wick RR, Gorrie C, Jenney A, Follador R, Thomson NR, et al. Identification of *Klebsiella* capsule synthesis loci from whole genome data. Microb Genomics. 2016 Dec;2(12):e000102.

29. Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. Nucleic Acids Res. 2016 Jan 4;44(D1):D694-697.

30. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 2017 Jan 4;45(D1):D566–73.

31. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. 2015;3:e1319.

32. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010 Mar 8;11:119.

33. van Dongen S, Abreu-Goodger C. Using MCL to extract clusters from networks. Methods Mol Biol Clifton NJ. 2012;804:281–95.

34. Schwyn B, Neilands JB. Universal chemical assay for the detection and determination of siderophores. Anal Biochem. 1987 Jan;160(1):47–56.

35. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016;26(12):1721–9.

36. Chen Y. Genome analysis of Gram-negative bacteria isolated from preterm baby faeces and whole-genome analysis of *Klebsiella oxytoca* [MSc]. Imperial College London; 2018.

37. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinforma Oxf Engl. 2014 Jul 15;30(14):2068–9.

38. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinforma Oxf Engl. 2015 Nov 15;31(22):3691–3.

39. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009 Jul;26(7):1641–50.

40. Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods. 2016;13(5):435–8.

41. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. J Stat Softw [Internet]. 2008 [cited 2018 May 1];25(1). Available from: https://www.jstatsoft.org/article/view/v025i01

42. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol J Comput Mol Cell Biol. 2012 May;19(5):455–77.

43. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the

quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015 Jul;25(7):1043–55.

44. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017 Aug 8;35(8):725–31.

45. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015 Jan;12(1):59–60.

46. Chen Y, Brook T, Alcon-Giner C, Clarke P, Hall L, Hoyles L. Draft genome sequences of *Citrobacter freundii* and *Citrobacter murliniae* strains isolated from the feces of preterm infants. Microbiol Resour Announc. 2019;8:e00494-19.

47. Chen Y, Brook T, Alcon-Giner C, Clarke P, Hall L, Hoyles L. Draft genome sequence of *Raoultella ornithinolytica* P079F W, isolates from the feces of a preterm infant. Microbiol Resour Announc. 2019;8:e00493-19.

48. Dobbler PT, Procianoy RS, Mai V, Silveira RC, Corso AL, Rojas BS, et al. Low microbial diversity and abnormal microbial succession is associated with necrotizing enterocolitis in preterm infants. Front Microbiol. 2017;8:2243.

49. Greenwood C, Morrow AL, Lagomarcino AJ, Altaye M, Taft DH, Yu Z, et al. Early empiric antibiotic use in preterm infants is associated with lower bacterial diversity and higher relative abundance of *Enterobacter*. J Pediatr. 2014 Jul;165(1):23–9.

50. Itani T, Ayoub Moubareck C, Melki I, Rousseau C, Mangin I, Butel M-J, et al. Preterm infants with necrotising enterocolitis demonstrate an unbalanced gut microbiota. Acta Paediatr Oslo Nor 1992. 2018 Jan;107(1):40–7.

51. Vergnano S, Menson E, Kennea N, Embleton N, Russell AB, Watts T, et al. Neonatal infections in England: the NeonIN surveillance network. Arch Dis Child Fetal Neonatal Ed. 2011 Jan;96(1):F9–14.

52. Rodrigues C, Passet V, Rakotondrasoa A, Diallo TA, Criscuolo A, Brisse S. Description of *Klebsiella africanensis* sp. nov., *Klebsiella variicola* subsp. *tropicalensis* subsp. nov. and *Klebsiella variicola* subsp. *variicola* subsp. nov. Res Microbiol. 2019 May;170(3):165–70.

53. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. Proc Natl Acad Sci U S A. 2015 Jul 7;112(27):E3574-3581.

54. Turton JF, Payne Z, Coward A, Hopkins KL, Turton JA, Doumith M, et al. Virulence genes in isolates of *Klebsiella pneumoniae* from the UK during 2016, including among carbapenemase gene-positive hypervirulent K1-ST23 and 'non-hypervirulent' types ST147, ST15 and ST383. J Med Microbiol. 2018 Jan;67(1):118–28.

55. Turton JF, Perry C, Elgohari S, Hampton CV. PCR characterization and typing of *Klebsiella pneumoniae* using capsular type-specific, variable number tandem repeat and virulence gene targets. J Med Microbiol. 2010 May;59(Pt 5):541–7.

56. Fung CP, Hu BS, Chang FY, Lee SC, Kuo BI, Ho M, et al. A 5-year study of the seroepidemiology of *Klebsiella pneumoniae*: high prevalence of capsular serotype K1 in Taiwan and implication for vaccine efficacy. J Infect Dis. 2000 Jun;181(6):2075–9.

57. Nishino K, Senda Y, Yamaguchi A. CRP regulator modulates multidrug resistance of *Escherichia coli* by repressing the *mdtEF* multidrug efflux genes. J Antibiot (Tokyo). 2008 Mar;61(3):120–7.

58. Nagakubo S, Nishino K, Hirata T, Yamaguchi A. The putative response regulator BaeR stimulates multidrug resistance of *Escherichia coli* via a novel multidrug exporter system, MdtABC. J Bacteriol. 2002 Aug;184(15):4161–7.

59. Doménech-Sánchez A, Hernández-Allés S, Martínez-Martínez L, Benedí VJ, Albertí S. Identification and characterization of a new porin gene of *Klebsiella pneumoniae*: its role in beta-lactam antibiotic resistance. J Bacteriol. 1999 May;181(9):2726–32.

60. Gunn JS, Lim KB, Krueger J, Kim K, Guo L, Hackett M, et al. PmrA-PmrB-regulated genes necessary for 4-aminoarabinose lipid A modification and polymyxin resistance. Mol Microbiol. 1998 Mar;27(6):1171–82.

61. Bohn C, Bouloc P. The *Escherichia coli cmlA* gene encodes the multidrug efflux pump Cmr/MdfA and is responsible for isopropyl-beta-D-thiogalactopyranoside exclusion and

spectinomycin sensitivity. J Bacteriol. 1998 Nov;180(22):6072–5.

62. Srinivasan VB, Rajamohan G. KpnEF, a new member of the *Klebsiella pneumoniae* cell envelope stress response regulon, is an SMR-type efflux pump involved in broad-spectrum antimicrobial resistance. Antimicrob Agents Chemother. 2013 Sep;57(9):4449–62.

63. Srinivasan VB, Singh BB, Priyadarshi N, Chauhan NK, Rajamohan G. Role of novel multidrug efflux pump involved in drug resistance in *Klebsiella pneumoniae*. PloS One. 2014;9(5):e96288.

64. Esterly JS, Richardson CL, Eltoukhy NS, Qi C, Scheetz MH. Genetic Mechanisms of antimicrobial resistance of *Acinetobacter baumannii*. Ann Pharmacother. 2011 Feb;45(2):218–28.

65. Tärnberg M, Nilsson LE, Monstein H-J. Molecular identification of (bla)SHV, (bla)LEN and (bla)OKP beta-lactamase genes in *Klebsiella pneumoniae* by bi-directional sequencing of universal SP6- and T7-sequence-tagged (bla)SHV-PCR amplicons. Mol Cell Probes. 2009 Aug;23(3–4):195–200.

66. Regué M, Climent N, Abitiu N, Coderch N, Merino S, Izquierdo L, et al. Genetic Characterization of the *Klebsiella pneumoniae waa* gene cluster, involved in core lipopolysaccharide biosynthesis. J Bacteriol. 2001 Jun;183(12):3564–73.

67. Follador R, Heinz E, Wyres KL, Ellington MJ, Kowarik M, Holt KE, et al. The diversity of *Klebsiella pneumoniae* surface polysaccharides. Microb Genomics [Internet]. 2016 Aug 25 [cited 2020 Mar 3];2(8). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5320592/

68. Beaugerie L, Metz M, Barbut F, Bellaiche G, Bouhnik Y, Raskine L, et al. *Klebsiella oxytoca* as an agent of antibiotic-associated hemorrhagic colitis. Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc. 2003 Sep;1(5):370–6.

69. Högenauer C, Langner C, Beubler E, Lippe IT, Schicho R, Gorkiewicz G, et al. *Klebsiella oxytoca* as a causative organism of antibiotic-associated hemorrhagic colitis. N Engl J Med. 2006 Dec 7;355(23):2418–26.

70. Raveh-Sadka T, Thomas BC, Singh A, Firek B, Brooks B, Castelle CJ, et al. Gut bacteria are

rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. eLife. 2015 Mar 3;4.

71. Brisse S, Verhoef J. Phylogenetic diversity of *Klebsiella pneumoniae* and *Klebsiella oxytoca* clinical isolates revealed by randomly amplified polymorphic DNA, gyrA and parC genes sequencing and automated ribotyping. Int J Syst Evol Microbiol. 2001 May;51(Pt 3):915–24.

72. Herzog KAT, Schneditz G, Leitner E, Feierl G, Hoffmann KM, Zollner-Schwetz I, et al. Genotypes of *Klebsiella oxytoca* isolates from patients with nosocomial pneumonia are distinct from those of isolates from patients with antibiotic-associated hemorrhagic colitis. J Clin Microbiol. 2014 May;52(5):1607–16.

73. Passet V, Brisse S. Description of *Klebsiella grimontii* sp. nov. Int J Syst Evol Microbiol. 2018 Jan;68(1):377–81.

74. Saha R, Farrance CE, Verghese B, Hong S, Donofrio RS. *Klebsiella michiganensis* sp. nov., a new bacterium isolated from a tooth brush holder. Curr Microbiol. 2013 Jan;66(1):72–8.

75. Merla C, Rodrigues C, Passet V, Corbella M, Thorpe HA, Kallonen TVS, et al. Description of *Klebsiella spallanzanii* sp. nov. and of *Klebsiella pasteurii* sp. nov. Front Microbiol. 2019;10:2360.

76. Moradigaravand D, Martin V, Peacock SJ, Parkhill J. Population structure of multidrug-resistant *Klebsiella oxytoca* within hospitals across the United Kingdom and Ireland identifies sharing of virulence and resistance genes with *K. pneumoniae*. Genome Biol Evol. 2017 Mar 1;9(3):574–84.

77. Hu Y, Wei L, Feng Y, Xie Y, Zong Z. *Klebsiella huaxiensis* sp. nov., recovered from human urine. Int J Syst Evol Microbiol. 2019 Feb;69(2):333–6.

78. Shibu P. Investigations of carbapenem-resistant *Klebsiella* species and associated clinical considerations. [London]: Westminster; 2019.

79. Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, da Costa MS, et al. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. Int J Syst Evol Microbiol. 2018 Jan;68(1):461–6.

80. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-

DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 2007 Jan;57(Pt 1):81–91.

81. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009 Nov 10;106(45):19126–31.

82. Zheng B, Xu H, Yu X, Lv T, Jiang X, Cheng H, et al. Identification and genomic characterization of a KPC-2-, NDM-1- and NDM-5-producing *Klebsiella michiganensis* isolate. J Antimicrob Chemother. 2018 01;73(2):536–8.

83. Pedersen T, Sekyere JO, Govinden U, Moodley K, Sivertsen A, Samuelsen Ø, et al. Spread of plasmid-encoded NDM-1 and GES-5 Carbapenemases among extensively drug-resistant and pandrug-resistant clinical *Enterobacteriaceae* in Durban, South Africa. Antimicrob Agents Chemother. 2018;62(5).

84. Seiffert SN, Wüthrich D, Gerth Y, Egli A, Kohler P, Nolte O. First clinical case of KPC-3-producing *Klebsiella michiganensis* in Europe. New Microbes New Infect. 2019 May;29:100516.

85. Chou H-C, Lee C-Z, Ma L-C, Fang C-T, Chang S-C, Wang J-T. Isolation of a chromosomal region of *Klebsiella pneumoniae* associated with allantoin metabolism and liver infection. Infect Immun. 2004 Jul;72(7):3783–92.

86. Fevre C, Passet V, Weill F-X, Grimont PAD, Brisse S. Variants of the Klebsiella pneumoniae OKP chromosomal beta-lactamase are divided into two main groups, OKP-A and OKP-B. Antimicrob Agents Chemother. 2005 Dec;49(12):5149–52.

87. Cano V, March C, Insua JL, Aguiló N, Llobet E, Moranta D, et al. *Klebsiella pneumoniae* survives within macrophages by avoiding delivery to lysosomes. Cell Microbiol. 2015 Nov;17(11):1537–60.

88. Skaar EP. The battle for iron between bacterial pathogens and their vertebrate hosts. PLoS Pathog. 2010 Aug 12;6(8):e1000949.

89. Flo TH, Smith KD, Sato S, Rodriguez DJ, Holmes MA, Strong RK, et al. Lipocalin 2 mediates an innate immune response to bacterial infection by sequestrating iron. Nature. 2004 Dec 16;432(7019):917–21.

90. Bachman MA, Oyler JE, Burns SH, Caza M, Lépine F, Dozois CM, et al. *Klebsiella pneumoniae* yersiniabactin promotes respiratory tract infection through evasion of lipocalin 2. Infect Immun. 2011 Aug;79(8):3309–16.

91. Lawlor MS, O'connor C, Miller VL. Yersiniabactin is a virulence factor for *Klebsiella pneumoniae* during pulmonary infection. Infect Immun. 2007 Mar;75(3):1463–72.

92. Navon-Venezia S, Kondratyeva K, Carattoli A. *Klebsiella pneumoniae*: a major worldwide source and shuttle for antibiotic resistance. FEMS Microbiol Rev. 2017 01;41(3):252–75.

93. Hamzaoui Z, Ocampo-Sosa A, Fernandez Martinez M, Landolsi S, Ferjani S, Maamar E, et al. Role of association of OmpK35 and OmpK36 alteration and blaESBL and/or blaAmpC genes in conferring carbapenem resistance among non-carbapenemase-producing *Klebsiella pneumoniae*. Int J Antimicrob Agents. 2018 Dec;52(6):898–905.

94. Penders J, Stobberingh EE, Savelkoul PHM, Wolffs PFG. The human microbiome as a reservoir of antimicrobial resistance. Front Microbiol. 2013;4:87.

FIGURE LEGENDS

Figure 1. Summary information for UK cohort included in this study. (a) Breakdown of birth mode and sex of preterm neonates. (b) 16S rRNA gene sequence results for *Enterobacteriaceae*-positive samples: upper panel, samples from which lactose-positive isolates were recovered; lower panel, samples from which lactose-negative isolates were recovered. (c) Representation of lactose-negative and lactose-positive *Enterobacteriaceae* isolated from faecal samples. (d) Tentative identities of lactose-positive *Enterobacteriaceae* as determined by using API 20E.

Figure 2. Summary of (a) antibiotic resistance and (b) virulence factor genes found in the *K. pneumoniae* isolates by comparison of protein sequences with those of the CARD and VFDB, respectively. (a) Strict CARD match, not identical but the bit-score of the matched sequence is greater than the curated BLASTP bit-score cut-off; perfect CARD match, 100 % identical to the reference sequence along its entire length. Loose matches are not shown to avoid presenting false positives based on sequences with low homology and bit-scores below CARD BLASTP cut-off recommendations. (b) Identity (%), BLASTP reported only for those proteins sharing >70 % identity and 90 % query coverage with VFDB protein sequences.

Figure 3. Genomic characterization of the *K. michiganensis* and *K. grimontii* isolates recovered from neonates. (a) Anvi'o representation of the genomes of *K. michiganensis* and *K. grimontii* isolates recovered from different infants. It is clear the isolates are different from one another at the genomic level. (b) Virulence factor (left side) and antibiotic resistance (right side) genes encoded by the isolates. Criteria for identity and strict/perfect match with respect to VFDB and CARD, respectively, are the same as those given for **Figure 2**.

Figure 4. Phenotypic assays for the *Klebsiella* isolates recovered from infants. (a) Strains were tested for persistence in PMA-differentiated THP-1 macrophages using a gentamicin protection assay. Intracellular bacteria were enumerated 1.5 h and 6 h after infection to determine persistence (*n*=4). Results are shown as mean (SD). (b) *Klebseilla* strains were grown in minimal medium and at 20 h

bacterial growth (OD_{600}) siderophore production was measured using the CAS assay (*n* = 3). Results are shown as mean (SD).

Figure 5. Identification of *K. oxytoca*-related species in infant faecal metagenomes. (a) Comparison of *K. oxytoca* and *K. michiganensis* abundance (as determined using Centrifuge) in stage2 samples of Ward *et al.* (1). (b) Abundance of *K. oxytoca* (determined using Centrifuge) across stage1, stage2 and stage3 samples of Ward *et al.* (1). (c) Separation of the strains of *K. grimontii* (*n*=27), *K. michiganensis* (*n*=76) and *K. oxytoca* (*n*=64) based on accessory genes (*n* = 5,108) detected in the Roary-generated open pangenome. (d) Relationship between PanPhlAn (overall alignment rate) and Centrifuge (abundance *K. oxytoca* (%)) data. (e, f, g) PCA plots show separation of strains in the pangenome plus PanPhlAn-detected strains based on presence of 500 randomly sampled accessory genes at (e) stage1, (f) stage2 and (g) stage3 of Ward *et al.* (1).

Figure 6. Anvi'o representation of the MAGs recovered from the metagenomes of infants included in the study of Ward *et al.* (1). (a) *K. oxytoca*. (b) *K. michiganensis*. It is notable that MAGs recovered from different life stages from the same infant (e.g. 10301_s1, 10301_s2) are highly similar to one another.

Table 1. Determination of MICs for *Klebsiella* spp. isolates

Isolate ID	Species	Gentamicin (mg/L)*	Meropenem (mg/L)†	Benzylpenicillin (mg/L)‡
#64	*K. pneumoniae*	**6.25#**	3.13	1560
#74	*K. pneumoniae*	1.5625	6.25	3130
#85	*K. pneumoniae*	1.5625	3.13	3130
#91	*K. quasipneumoniae*	1.5625	6.25	3130
#92	*K. pneumoniae*	1.5625	3.13	3130
#95	*K. pneumoniae*	3.125	3.13	3130
#102	*K. pneumoniae*	1.5625	6.25	3130
#115	*K. pneumoniae*	1.5625	3.13	3130
#118	*K. pneumoniae*	1.5625	6.25	3130
#80	*K. grimontii*	3.125	1.56	6.25
#83	*K. grimontii*	**12.5**	1.56	780
#99	*K. grimontii*	3.125	0.78	780
#88	*K. michiganensis*	3.125	3.13	3130
#108	*K. michiganensis*	**6.25**	1.56	3130

**Enterobacteriaceae* EUCAST breakpoint for gentamicin resistance is >4 mg/L, and for sensitivity is ≤2 mg/L.

†*Enterobacteriaceae* EUCAST breakpoint for meropenem is resistance >8 mg/L, and for sensitivity is ≤2 mg/L.

‡No *Enterobacteriaceae* EUCAST data are available for benzylpenicillin.

#Bold type, resistant; underlined, intermediate.

Table 2. Summary statistics for MAGs recovered from preterm infant metagenomes

MAG*	Genome length (bp)	Max. contig length	Coverage†	N50	No. of scaffolds	GC content (%)	Completeness (%)‡	Contamination (%)‡	CDS	No. of tRNAs	No. of rRNAs	Species	Quality	*gapA*	*infB*	*mdh*	*pgi*	*phoE*	*rpoB*	*tonB*	ST
10071_s2	6,577,866	334,901	~131x	70,229	349	52.95	97.59	3.64	6204	54	3	*K. michiganensis*	High	3	9	8	9	20	*	8	*
10301_s1	6,304,211	308,355	~158x	143,120	176	53.15	99.70	0.48	5920	54	4	*K. michiganensis*	High	3	5	21	13	74	6	12	202
10301_s2	6,504,340	160,223	~21x	30,917	609	54.93	98.69	4.21	6191	50	2	*K. michiganensis*	Medium	3	5	21	13	74	6	12	202
10441_s2	6,128,395	396,555	~85x	122,210	176	53.35	100.00	0.71	5691	58	4	*K. michiganensis*	High	3	5	21	13	24	6	*	*
10451_s1	6,135,593	499,885	~79x	130,343	154	53.44	100.00	0.71	5708	53	7	*K. michiganensis*	High	3	5	21	13	24	6	*	*
10451_s2	6,140,759	377,549	~65x	130,433	166	53.53	100.00	0.71	5707	54	4	*K. michiganensis*	High	3	5	21	13	24	6	*	*
10472_s2	6,352,656	271,779	~46x	80,647	252	53.26	99.90	2.59	5889	50	3	*K. oxytoca*	High	1	7	2	1	65	1	2	176
10501_s2	6,179,579	451,918	~39x	133,209	217	54.01	100.00	1.34	5730	55	9	*K. michiganensis*	High	3	5	21	3	24	6	*	*
10601_s2	6,014,627	328,826	~59x	181,417	121	52.63	99.96	0.34	5540	56	4	*K. michiganensis*	High	3	5	21	3	24	6	*	*
11151_s2	6,595,368	114,676	~40x	17,470	949	54.71	97.17	5.81	6252	59	0	*K. michiganensis*	Medium	3	5	21	13	20	*	12	*
11202_s2	6,328,179	267,673	~144x	111,043	224	53.38	98.81	3.22	5864	54	3	*K. michiganensis*	High	*	8	24	33	20	6	23	*
11291_s2	6,301,261	429,108	~45x	146,444	193	53.10	99.70	0.79	5911	68	8	*K. michiganensis*	High	3	8	17	21	40	17	29	84
11292_s1	6,533,348	267,672	~43x	118,981	400	53.17	98.81	2.77	6064	48	3	*K. michiganensis*	High	*	8	24	33	20	6	23	*
11292_s2	6,383,200	267,500	~70x	119,207	202	50.77	99.70	2.49	5891	63	8	*K. michiganensis*	High	*	8	24	33	20	6	23	*
11311_s2	6,427,314	266,393	~137x	66,881	516	52.89	99.97	2.96	5986	46	1	*K. oxytoca*	Medium	2	2	2	3	19	2	2	199
11701_s2	6,230,855	162,114	~60x	50,736	555	54.40	99.85	3.20	5807	50	2	*K. oxytoca*	Medium	1	7	2	1	65	1	2	176
11981_s2	5,658,923	237,765	~10x	60,298	326	53.62	89.43	2.20	5280	30	2	*K. oxytoca*	Medium	*	2	2	3	19	*	2	*
12121_s1	5,843,313	426,086	~121x	101,108	251	54.02	96.40	2.49	5358	53	4	*K. michiganensis*	High	3	33	17	45	20	6	48	149
12121_s2	5,690,611	180,068	~82x	56,100	371	54.36	94.21	2.85	5226	35	3	*K. michiganensis*	High	3	33	17	45	20	*	48	*
12221_s3	6,115,156	469,104	~17x	174,402	147	54.67	99.70	1.60	5667	63	8	*K. michiganensis*	High	3	5	21	20	24	6	30	108
12641_s2	6,383,005	237,677	~55x	80,508	279	53.91	99.44	3.15	6027	56	2	*K. michiganensis*	High	3	5	21	13	74	*	12	*
12651_s2	5,124,388	199,319	~28x	81,423	207	53.59	94.05	3.21	4801	42	3	*K. michiganensis*	High	14	24	15	8	18	*	4	*
12651_s3	5,422,817	363,696	~10x	73,253	207	53.78	95.49	2.40	5047	41	2	*K. michiganensis*	Medium	14	2	15	8	18	*	4	*
30221_s1	6,220,970	480,926	~51x	173,840	132	52.47	99.70	1.42	5770	61	9	*K. michiganensis*	High	3	5	21	20	11	6	20	43
30221_s2	6,224,584	349,605	~55x	170,314	142	52.27	99.70	1.39	5765	58	7	*K. michiganensis*	High	3	5	21	20	11	6	20	43

*E.g. 10301_s1 represents a MAG recovered from infant 10301 at stage1.

†Coverage, based on coverage of longest contig (determined from spades data).

‡Completeness and contamination determined using CheckM (v1.0.18).