MICROBIAL GENOMICS

BIORESOURCE

Cooper *et al.*, *Microbial Genomics* 2024;10:001206

DOI 10.1099/mgen.0.001206

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

A validated pangenome-scale metabolic model for the *Klebsiella pneumoniae* species complex

Helena B. Cooper[1,2,*], Ben Vezina[1,2], Jane Hawkey[1], Virginie Passet[3], Sebastián López-Fernández[3], Jonathan M. Monk[4], Sylvain Brisse[3], Kathryn E. Holt[1,5] and Kelly L. Wyres[1,2,*]

Abstract

The *Klebsiella pneumoniae* species complex (KpSC) is a major source of nosocomial infections globally with high rates of resistance to antimicrobials. Consequently, there is growing interest in understanding virulence factors and their association with cellular metabolic processes for developing novel anti-KpSC therapeutics. Phenotypic assays have revealed metabolic diversity within the KpSC, but metabolism research has been neglected due to experiments being difficult and cost-intensive. Genome-scale metabolic models (GSMMs) represent a rapid and scalable *in silico* approach for exploring metabolic diversity, which compile genomic and biochemical data to reconstruct the metabolic network of an organism. Here we use a diverse collection of 507 KpSC isolates, including representatives of globally distributed clinically relevant lineages, to construct the most comprehensive KpSC pan-metabolic model to date, KpSC pan v2. Candidate metabolic reactions were identified using gene orthology to known metabolic genes, prior to manual curation via extensive literature and database searches. The final model comprised a total of 3550 reactions, 2403 genes and can simulate growth on 360 unique substrates. We used KpSC pan v2 as a reference to derive strain-specific GSMMs for all 507 KpSC isolates, and compared these to GSMMs generated using a prior KpSC pan-reference (KpSC pan v1) and two single-strain references. We show that KpSC pan v2 includes a greater proportion of accessory reactions (8.8%) than KpSC pan v1 (2.5%). GSMMs derived from KpSC pan v2 also generate more accurate growth predictions, with high median accuracies of 95.4% (aerobic, *n*=37 isolates) and 78.8% (anaerobic, *n*=36 isolates) for 124 matched carbon substrates. KpSC pan v2 is freely available at https://github.com/kelwyres/KpSC-pan-metabolic-model, representing a valuable resource for the scientific community, both as a source of curated metabolic information and as a reference to derive accurate strain-specific GSMMs. The latter can be used to investigate the relationship between KpSC metabolism and traits of interest, such as reservoirs, epidemiology, drug resistance or virulence, and ultimately to inform novel KpSC control strategies.

DATA SUMMARY

(1) *Klebsiella pneumoniae* species complex (KpSC) pan v2 metabolic model (available at https://github.com/kelwyres/KpSC-pan-metabolic-model).

(2) All KpSC isolate whole genome sequences used in this work were reported previously and are available under Bioprojects PRJEB6891, PRJNA351909, PRJNA493667, PRJNA768294, PRJNA253462, PRJNA292902 and PRJNA391323. Individual accessions listed in Table S1, available in the online version of this article.

Received 21 December 2023; Accepted 06 February 2024; Published 20 February 2024

Author affiliations: [1]Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria 3004, Australia; [2]Centre to Impact AMR, Monash University, Clayton, Victoria 3800, Australia; [3]Institut Pasteur, Université de Paris, Biodiversity and Epidemiology of Bacterial Pathogens, 75015 Paris, France; [4]Department of Bioengineering, University of California, San Diego, California 92093, USA; [5]Department of Infection Biology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK.

***Correspondence:** Helena B. Cooper, helena.cooper@monash.edu; Kelly L. Wyres, kelly.wyres@monash.edu

Keywords: *Klebsiella*; genome-scale metabolic models; pangenome; genomics; bacterial metabolism.

Abbreviations: BBH, bidirectional best hit; BOF, Biomass Objective Function; FN, false negative; FP, false positive; GPR, Gene–Protein–Reaction; GSMM, genome-scale metabolic model; KpSC, *Klebsiella pneumoniae* species complex; ST, sequence type; TN, true negative; TP, true positive; TR, technical replicate.

All supporting data and protocols have been provided within the article, through supplementary data files, Github or FigShare repositories. The KpSC pan v2 model is available at https://github.com/kelwyres/KpSC-pan-metabolic-model.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Four supplementary figures and seven supplementary tables are available with the online version of this article.

Significance as a BioResource to the community

Klebsiella pneumoniae and its close relatives in the *K. pneumoniae* species complex (KpSC) are priority antimicrobial-resistant pathogens that exhibit extensive genomic diversity. There is growing interest in understanding KpSC metabolism, and genome-scale metabolic models (GSMMs) provide a rapid, scalable option for exploration of whole cell metabolism plus phenotype prediction. Here we present a KpSC pan-metabolic model representing the cellular metabolism of 507 diverse KpSC isolates. Our model is the largest and most comprehensive of its kind, comprising >2400 genes associated with >3500 metabolic reactions, plus manually curated evidence annotations. These data alone represent a key knowledge resource for the *Klebsiella* research community; however, our model's greatest impact lies in its potential for use as a reference from which highly accurate strain-specific GSMMs can be derived to inform in-depth strain-specific and/or large-scale comparative analyses.

(3) Strain-specific genome-scale metabolic models (GSMMs) used for comparative analyses (deposited in Figshare – https://doi.org/10.6084/m9.figshare.24871914), plus their associated MEMOTE reports (indicates completeness and annotation quality), reaction and gene presence–absence matrices across all isolates.
(4) Growth phenotype predictions derived from strain-specific GSMMs (deposited in Figshare – https://doi.org/10.6084/m9.figshare.24871914).
(5) Binarized Biolog growth phenotype data (plates PM1 and PM2) for n=37 isolates in aerobic and anaerobic conditions (deposited in Figshare – https://doi.org/10.6084/m9.figshare.24871914).
(6) Additional growth assay data for six substrates not included on Biolog plates PM1 and PM2 (deposited in Figshare – https://doi.org/10.6084/m9.figshare.24871914).

INTRODUCTION

Klebsiella pneumoniae and closely related organisms in the *K. pneumoniae* species complex (KpSC) are a major source of nosocomial infections globally and among the top three causes of deaths associated with antibiotic resistance [1]. Consequently, the World Health Organization has designated development of novel anti-*Klebsiella* control strategies an urgent global priority [2]. Many existing antimicrobials target or alter cellular metabolic processes, and recent works have highlighted a role for metabolism in *K. pneumoniae* virulence [3, 4], supporting these processes as prime targets for novel therapeutics. For example, production of the GltA citrate synthase appears to facilitate liver and spleen infections in mouse models [5], while a psicose sugar utilization locus has been associated with lung infection [6]. Despite this, our understanding of KpSC metabolism remains comparatively limited. The matter is further complicated by metabolic variability [7, 8], probably driven by substantial genetic variation: the *K. pneumoniae* population consists of hundreds of distinct ancestral lineages ('clones') where gene content varies significantly within and between clones [9]. The total pangenome is estimated to exceed 100 000 protein coding sequences [10], with ~37% of these predicted to be involved in metabolism. However, it can be difficult to interpret gene-level information in the context of metabolic phenotypes, and large-scale laboratory growth phenotype experiments are infeasible due to their time and cost-intensive nature, so the true extent of population variation remains unknown.

Genome-scale metabolic models (GSMMs) are a powerful *in silico* approach to explore the metabolism of individual cells, and predict metabolic phenotypes such as the ability to grow on a given substrate or the outcomes of gene knock-out mutations (Fig. 1). Such analyses have been used to identify novel anti-*K. pneumoniae* drug targets [3, 11] and indicate substrate preferences among clinical isolates [3, 12]. GSMMs are a mathematical representation of an individual strain's metabolism derived from published literature, genome annotations and biochemical data [13, 14]. They contain stoichiometrically balanced reactions that are centred around the Biomass Objective Function (BOF), representing the growth requirements for producing a single cell (Fig. 1). Gene information for every reaction is summarized as a Gene–Protein–Reaction (GPR) rule (Fig. 1), which summarizes the genes required to catalyse a reaction, along with their associated nucleotide and amino acid sequences. Reactions are assigned a confidence score based on the level of evidence available to support their occurrence in the species of interest, with protein characterization being the strongest form of evidence [13].

GSMMs can be built *de novo* by identifying candidate metabolic reactions from genome annotations and experimental data, before assembling and manually refining them into a draft model; this is a very time-consuming process that is not feasible for large genome collections [13]. An alternative approach leverages reference models as templates to extract strain-specific models [15]. We recently implemented this approach in an automated pipeline Bactabolize, which enables rapid and scalable model generation and phenotype prediction [16]. We also described a KpSC pan-reference model (KpSC pan v1) based on 37 strain-specific models [16], each derived from *K. pneumoniae* MGH 78578 model iYL1228 [17] with minimal manual curation [8]. Using Bactabolize and KpSC pan v1, we showed that the reference-based approach can generate highly accurate strain-specific models that equalled or surpassed existing scalable model generation methods [16]. However, models generated for KpSC clones not represented in the KpSC pan v1 model tended to have a lower accuracy compared to those that were present [16], indicating room for improvement.

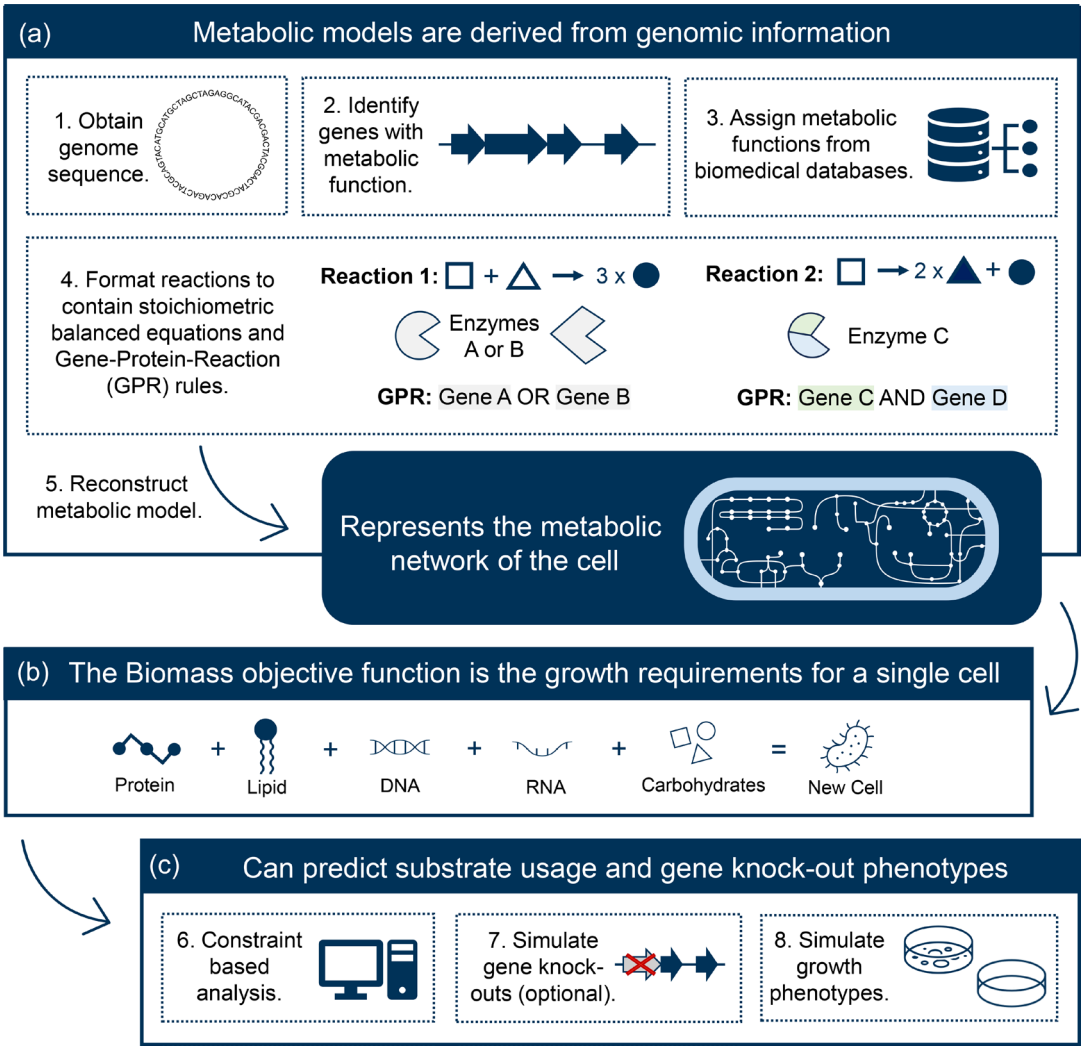

Fig. 1. Genome-scale metabolic model (GSMM) overview. (a) Summary of the typical approach for generating a GSMM. (b) Constituent components of the Biomass Objective Function (BOF), which defines the requirements for production of new daughter cells. Optimization of the BOF via constraint-based analysis (c) can be used to predict growth in different conditions with an optimized objective value ≥0.001 indicating growth whereas an objective value <0.001 indicates no growth. (c) Common applications of GSMMs.

In this study, we generated a novel KpSC pan-reference model using a large, diverse collection of 507 KpSC isolates, which captures more metabolic information than the prior pan-model. When used as a reference, KpSC pan v2 resulted in a more diverse collection of strain-specific GSMMs, with higher growth prediction accuracies as validated with phenotype data. Our data highlight the value of pan-metabolic reference models to support large-scale comparative metabolic modelling analyses of diverse bacterial species, such as *K. pneumoniae* and other priority antimicrobial-resistant pathogens.

METHODS

Genome collection and pangenome construction

Fig. 2 summarizes our approach for developing the KpSC pan v2 reference model. A starting dataset of 510 *Klebsiella* genome assemblies was collected (Table S1), comprising 452 isolate genomes from the *Klebsiella* Acquisition Surveillance Project at Alfred Health (KASPAH) [18], a diverse collection for which all isolates were available in-house for phenotypic validation. In addition, 58 genomes for which draft metabolic models were published previously (n=20/22 multi-drug-resistant *K. pneumoniae* [19], n=37 diverse KpSC [8] and n=1 hypervirulent *K. pneumoniae* [20]), were also included. KASPAH isolates were *de novo* assembled using SPAdes optimized with Unicycler v0.4.7 as reported previously [18, 21] and all remaining assemblies were retrieved from public databases [7, 8, 19, 20]. To ensure the assemblies were of sufficient quality to produce accurate metabolic models, we applied a tiered quality control framework as described previously [16]. Genomes were filtered using a threshold of ≤200 assembly graph

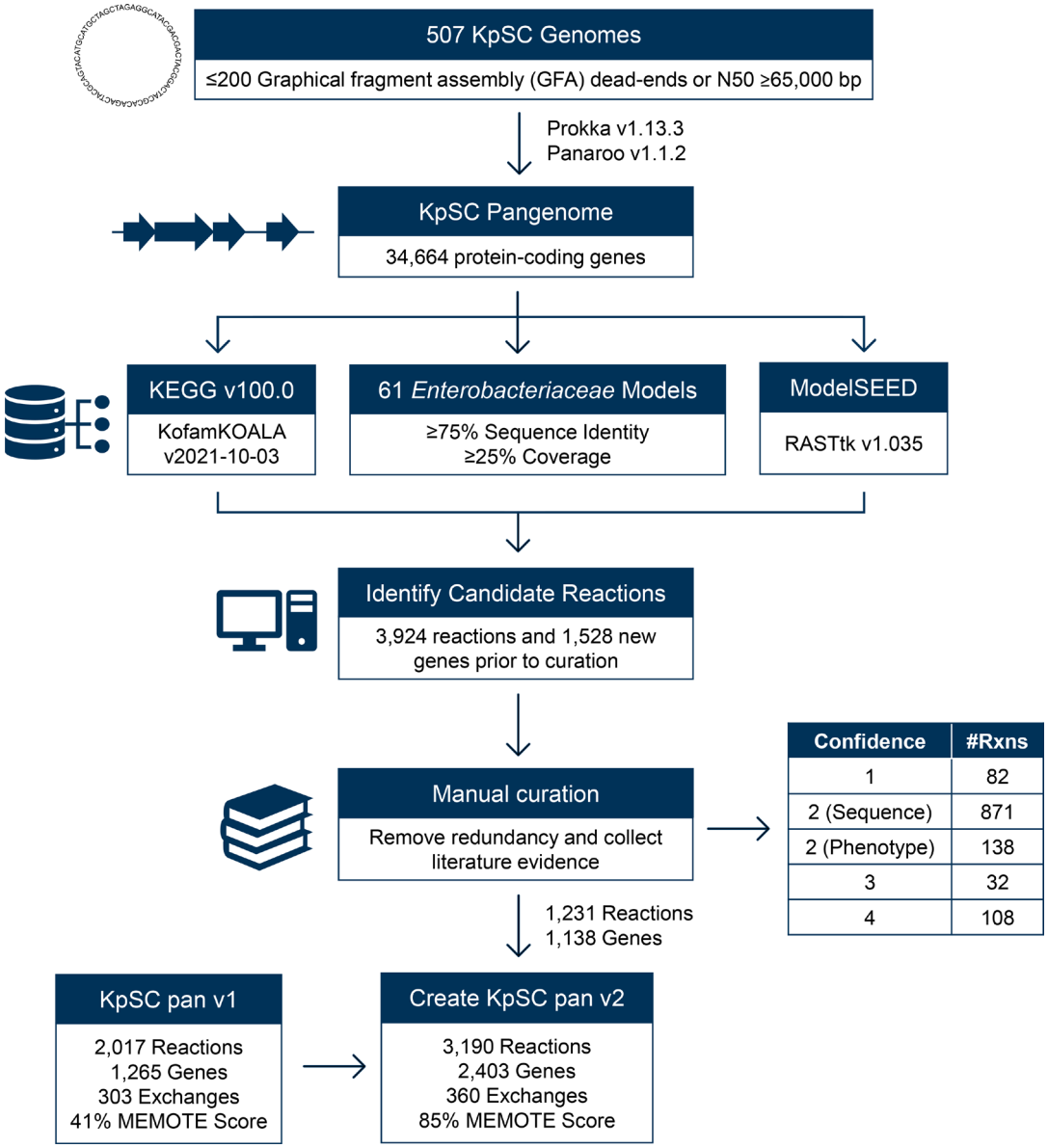

Fig. 2. Approach for generating KpSC pan v2. Blue boxes indicate major analysis steps and data sources, whereas white boxes indicate additional information. The table adjacent to manual curation is a breakdown of the evidence confidence scores (Confidence) and the number of reactions (#Rxns) in each category [13]. The final set of curated gene and reaction information is available in Table S3 and summarized in Table 1.

dead-ends (where assembly graphs were available, *n*=452 KASPAH genomes), and subsequently by N50 ≥65 000 bp (Table S1). The 507 genomes that passed quality control were annotated using prokka v1.13.3 [22] with the following parameters: --gcode 11, --addgenes and --protein to initially annotate proteins using *K. pneumoniae* HS12286 [23]. These were then compiled into a pangenome using panaroo v1.1.2 [24] (default parameters, except for protein family sequence identity and length difference, both of which were set to 90%). Kleborate v2.3.2 was used to determine the seven gene multi-locus sequence types (STs), antibiotic resistance determinants and capsule types [25, 26].

Identification and curation of metabolic reactions

Representative sequences from each of the pangenome orthologue clusters were used to find new candidate metabolic reactions based on sequence homology to proteins identified from three sources (Fig. 2):

(1) **Published *Enterobacteriaceae* GSMMs:** We identified 61 published metabolic models for which corresponding sequence data were publicly available (Table S2), including 50 *Escherichia coli* [27–30], eight *Shigella* [30], one *Yersinia pestis* [31] two

Salmonella enterica [32, 33] and five additional *K. pneumoniae* models [8, 19, 20]. Models were downloaded from BiGG [34], BioModels [35] or KBase (iKp1289 only) [36]. Genomes were downloaded from NCBI [37]. Sequence orthologues were identified by BLASTn and BLASTp bidirectional best hit (BBH) search with coverage ≥25% and identity ≥75% as described previously [8, 15, 38]. Reactions associated with the BBH were subject to manual curation as described below. Reactions for which the GPR comprised n>1 required genes (i.e. Gene X AND Gene Y) were considered for curation if n−1 gene orthologes were identified within the KpSC pangenome with identity ≥75%. Additionally, reactions matching those in the existing KpSC pan v1 model for which novel orthologues were identified at <75% identity were considered for curation via update of the GPR to include the novel divergent orthologue as additional gene sequence variants.

(2) **KEGG:** Metabolic reactions were identified from KEGG v100.0 [39] using kofamscan v2021-09-05 [40] (significance value of ≤0.001).

(3) **ModelSEED:** RASTtk v1.035 was used to assign enzyme commission numbers to each representative KpSC pangenome sequence, and these were matched back to the ModelSEED v2.6.1 database to extract the corresponding metabolism and reaction data [41, 42].

Candidate reactions were filtered to remove redundancy by: (i) comparing their associated protein sequences (80% identity threshold); (ii) merging reactions identified from multiple sources; and (iii) removing information already captured by the KpSC pan v1 model. The remaining reactions were manually curated to ensure they had sufficient evidence for inclusion in the KpSC pan v2 model via a literature search [13] and were formatted according to COBRApy conventions [43]. Each reaction was given a confidence score based on the evidence supporting its inclusion in the model; for example, a score of 2 corresponds to sequence homology evidence alone and/or phenotypic evidence that *Klebsiella* can metabolize the associated substrate(s) [13]. Documented gene knock-out data showing that the absence of a gene causes a loss in phenotype are considered higher level evidence (score 3), as is a documented solved *Klebsiella* protein structure (score 4) [13].

The final set of curated reactions (Table S3) and associated sequence data were combined with those from the KpSC pan v1 model to create the updated KpSC pan-reference model, known as KpSC pan v2. This model, KpSC pan v2.0.1, is available in JSON and SBML formats at github.com/kelwyres/KpSC-pan-metabolic-model, along with accompanying nucleotide (.fna) and protein (.faa) sequence files.

Biomass Objective Function (BOF)

The BOF in KpSC pan v2 was inherited from KpSC pan v1, which in turn was inherited from the original GSMM for *K. pneumoniae* MGH 78578 (iYL1228). The original equation and stoichiometric coefficients (mmol gDW^{-1}) were informed by experimental macromolecular composition analysis of *K. pneumoniae* MGH 78578 [17], and captures the protein, RNA, DNA, membrane and capsule composition requirements for MGH 78578. The BOF was subsequently updated to remove the requirement for rhamnose sugars [8] that are known as variable components of the capsule polysaccharide (present in the particular capsule type expressed by MGH 78578, but not conserved among all *K. pneumoniae* [44]). Here we made a further update to remove the requirement for mannose sugars, that are also known as variable capsule components and for which the associated synthesis genes were absent from multiple genomes in our collection (see Results).

Strain-specific metabolic models and growth phenotype predictions

In order to assess the metabolic content and diversity represented by the KpSC pan v2 model, we compared its performance as a reference model against those of the KpSC pan v1 model [16] and two curated single-strain *K. pneumoniae* models (iYL1228 [17] and iKp1289 [20]). We used Bactabolize v1.0.2 [16] to build four independent strain-specific GSMMs and independently predict growth phenotypes for each of the 507 genome assemblies that passed quality control. All possible sources of carbon, nitrogen, sulphur and phosphorus substrates were tested as the sole source of the respective element in M9 minimal media, in both aerobic and anaerobic conditions. Prior to use as a reference, iYL1228 was updated to add the chemical formulas for every metabolite using BiGG [34], which was required to allow Bactabolize to detect the potential growth sources [16]. A sink reaction in iKp1289 (sink__DASH__dna5mtc_c0) was also modified to be non-reversible to ensure methylated DNA was not being used as a carbon source [20]. Draft models were built using the draft_model command (default parameters) and growth phenotype predictions were performed using the fba command (default parameters except; fba_open_value = −20). When iKp1289 was used as a reference, media_type=m9_SEED_media and fba_spec_name=m9_SEED_spec were set as iKp1289 uses ModelSEED IDs rather than BiGG IDs. Based on a published metabolic modelling protocol [15], simulated growth phenotypes with a predicted biomass value >0.001 were considered positive for growth and those with biomass ≤0.001 were considered negative (Table S4). Comparisons to other metabolic modelling pipelines, such as CarveMe [45], were not performed in this study as the KpSC pan v1 model has been previously benchmarked against several pipelines and was shown to outperform or equal these [16].

Estimating model accuracies

To determine the accuracy of the KpSC pan v2 model, we collected two types of *in vitro* data, the first being Biolog growth phenotypes for 190 carbon substrates (PM1 and PM2 plates) in aerobic and anaerobic conditions (for n=37 KpSC isolates). The

second set of data were derived from independent growth assays for six additional carbon substrates tested in aerobic conditions only (*n*=37–42 KpSC isolates per substrate, Table S5).

The aerobic Biolog data have been described previously by Blin *et al.* [7], with the anaerobic data generated via the same protocol with the following differences: strains were cultured onto Luria-Bertani agar at 37°C for 48 h within an anaerobic chamber. All required reagents and materials were placed inside the anaerobic chamber for 48 h in parallel. Biolog 96-well plates were placed in anaerobic jars containing 2.5 l anaeropack catalysers (Thermo Scientific) for 48 h before being transferred to the anaerobic chamber, where the remainder of the experiment took place. For each strain, IF-0a fluid was inoculated with fresh colonies to reach 40% transmittance. The bacterial preparation was diluted to 1/20 in IF-0a, and then to 1/5 in a solution containing tetrazolium chloride (Dye D), menadione, potassium ferricyanide and water.

The aerobic and anaerobic growth thresholds were defined empirically, with the aerobic growth threshold having been previously determined as a maximum value >150 on the respiration curve [7]. To determine the anaerobic growth threshold, we plotted the distribution of the maximum values from the Biolog respiration curve for every substrate and isolate tested to identify the inflection point between the bimodally distributed data (i.e. growth cut-off points). The distribution was distinct from that of the aerobic data [7], and we hypothesized that this may be driven by species-specific variations (Fig. S1). Plotting the distributions by species confirmed our hypothesis (Fig. S1), and informed species-specific thresholds for *K. pneumoniae* (maximum value >165), *Klebsiella quasipneumoniae* subsp. *quasipneumoniae* (maximum value >155), *Klebsiella quasipneumoniae* subsp. *similipneumoniae* (maximum value >150) and *Klebsiella variicola* subsp. *variicola* (maximum value >160). As *Klebsiella africana, Klebsiella variicola* subsp. *tropica* and *Klebsiella quasivariicola* were only represented by one isolate each, we could not determine species-specific thresholds and therefore used a threshold of 190, derived from the combined distribution (Fig. S1).

Independent phenotype assays were performed for six substrates not present on the Biolog PM1 and PM2 carbon plates, which were chosen because they were newly implemented in the KpSC pan v2 model (either via addition of new exchange reactions or by completion of a metabolic pathway that was partially present in KpSC pan v1), they were available for purchase and import to Australia, and were not prohibitively expensive. All substrates were obtained from Sigma-Aldrich: uracil (CAS 66-22-8), allantoin (CAS 97-59-6), 2-aminoethylphosphonic acid (CAS 2041-14-7), L-ectoine (CAS 96702-03-03), hydrocinnamic acid (CAS 501-52-0) and beta-alanine (CAS 107-95-9). Growth assays were performed for the 37 KpSC isolates with existing Biolog data [7] and up to five additional isolates [18, 46] to confirm the predicted growth variability across the wider population. The protocol was described previously by Hawkey *et al.* [8]. Briefly, isolates were grown overnight in 5 ml of minimal media [2× M9, Minimal Salts (Sigma-Aldrich), 2 mM $MgSO_4$ and 0.1 mM $CaCl_2$] and 20 mM D-glucose (CAS 50-99-7) at 37°C, with shaking at 200 rpm. Overnight cultures were diluted to the McFarland standard of 0.4–0.55 turbidity. Dilutions of 20 µl were inoculated in triplicate to 96-well cell culture plates (Corning) containing M9 minimal media with either 20 mM substrate (4 mM for allantoin due to low solubility [47]), Milli-Q H_2O (negative control) or 20 mM D-glucose (positive control) at pH 7.0 (pH 6.0 for uracil to avoid precipitation). Plates were incubated at 37°C and the OD_{600} values recorded after 24 and 48 h using the Infinite 200 PRO plate reader (Tecan) using Tecan i-control version 2.0.10.0, firmware V_4.31_06/19_Infinite, at 600 nm absorbance after 30 s of shaking at 218.3 rpm (amplitude of 3 mm). Isolates were marked as positive for growth if the adjusted OD_{600} fold change (calculated as indicated below) exceeded the substrate-specific threshold (determined from the empirical distributions, see Fig. S2). Positive growth and fold-change were determined using the following formulae where TR refers to technical replicates:

Positive growth=mean(TR Substrate +Isolate OD_{600})−mean(TR Substrate Only OD_{600})

Fold-change=Substrate positive growth/Water positive growth

We additionally tested two predicted L-histidine auxotrophs and two control isolates for growth on L-histidine (Sigma-Aldrich, CAS 71-00-1). Experiments were performed as described above with the exception that the primary overnight cultures were grown in M9 minimal media plus 20 mM D-glucose and 20 mM L-histidine. Residual histidine was removed by centrifugally washing overnight cultures twice with 1.5 ml 0.9 M NaCl at 8000 ***g*** for 15 min, then a final resuspension in 0.9 M NaCl prior to dilution to the McFarland standard of 0.4–0.55.

In vitro data were compared to *in silico* predictions to determine true positive (TP), true negative (TN), false positive (FP) and false negative (FN) results. Accuracy, sensitivity and specificity were calculated for the strain-specific GSMMs derived from each reference model using the following formulae:

Accuracy=(TP+TN)/(TP+TN+FP+FN)

Sensitivity=(TP)/(TP+FN)

Specificity=(TN)/(TN+FP)

These metrics were calculated using multiple subsets of substrates so that direct comparisons could be made between models generated using different references, which each support different substrates (see Tables S6 and S7). The first comparison was for the set of 89 substrates that could be simulated by all models. Additional comparisons represented the set of 91 substrates that

could be simulated by all models except those generated using the iKp1289 reference, and the set of 105 substrates that could be simulated by all models except those generated using iKp1289 and iYL1228. The final accuracy for KpSC pan v2 was calculated for the maximum number of substrates with matched *in vitro* and *in silico* data (*n*=124).

RESULTS

Diverse collection of isolate genomes

The initial dataset of 510 isolates contained at least one representative from each of the seven KpSC subspecies (including 79.0% *K. pneumoniae*, 15.1% *K. variicola* and 5.3% *K. quasipneumoniae,* Table S1), 268 unique STs and 115 distinct capsule polysaccharide loci (Table S1), notably including representatives of 12 out of 13 global problematic clones [9]. The dataset was diverse in terms of isolate sources, including 344 hospital-acquired infection, 45 community-acquired infection, 126 human gut carriage and 13 environmental isolates, with 23.7% being multidrug resistant (acquired resistance to at least three antimicrobial classes, including 13.7% extended-spectrum beta-lactamases, 6.1% carbapenemases) (Table S1).

Description of KpSC pan v2

Rather than building the reference model *de novo*, we used the KpSC pan v1 model [16] as the starting point to incorporate additional metabolic reactions and/or divergent gene orthologues. Candidate orthologues and their associated metabolic reactions were identified by comparing representative gene sequences derived from the pangenome of 507 KpSC isolates (excludes three genomes which did not pass quality control, see Methods) to those of published *Enterobacteriaceae* GSMMs [20, 27–33], KEGG [39] and ModelSEED [41] databases (Fig. 2). This pangenome contained 34 664 unique genes with 11.4% classed as core (present in ≥95% of isolates) and 88.6% classed as accessory (present in <95% of isolates). Following manual curation (summarized in Table S3), 1141 gene sequences were confirmed as orthologues, corresponding to 1058 unique metabolic reactions (883 catalytic, 175 transport) that were not present in KpSC pan v1. These reactions were formatted as new GPR rules and added to the KpSC pan v2 model along with 173 reactions (18 catalytic, 63 transport, 92 pseudo) that were not associated with a GPR rule.

Most novel GPR-associated reactions were identified from *Enterobacteriaceae* GSMMs only (44.9%) or KEGG only (46.8%). The remaining reactions were identified from *Enterobacteriaceae* GSMMs combined with information from KEGG (6.9%) or ModelSEED (0.2%), and ModelSEED only (1.2%). Each reaction was curated and assigned a confidence score [13], the majority with a confidence score of 2 (81.9%) which corresponded to 70.7% genomic and 11.2% phenotypic evidence (Table S3). The remaining reactions comprised 8.8% confidence score 4 (e.g. protein structures), 2.6% confidence score 3 (e.g. gene knock-outs) and 6.7% confidence score 1 (i.e. required for modelling but without additional supporting evidence). Of these curated reactions, 50.1% were supported by evidence from *Klebsiella,* with the remaining supported by evidence from other *Enterobacteriaceae* or bacterial species (all confidence score ≤2, Table S3).

As an example of curation, a candidate reaction that transports periplasmic lipid A to the extracellular environment (14DENLIPIDAt2ex) requires the products of two genes for each subunit of the LptDE lipopolysaccharide transporter [48]. We successfully identified an orthologue of one of the genes among our KpSC pangenome (>75% identity threshold), but the best match for the second gene (KpSC KPN_00673) had a sequence identity of just 70.4% when compared to that in *E. coli*. Following a literature search, we identified a published *K. pneumoniae* LptDE protein structure [48], and subsequently confirmed that the corresponding amino acid sequence shared 97.2% identity with KPN_00673, supporting the inclusion of the latter in the GPR rule.

The 92 novel pseudo-reactions without a GPR rule corresponded to: 85 exchange reactions, which enable the model to import/export metabolites into the cell; plus seven sink reactions which were required for the model to simulate growth on certain substrates for which the true degradation/export pathways have not been defined (Table S3). There were also 392 reactions (377 catalytic and 15 transport) that already existed in the KpSC pan v1 model that were updated to include additional gene sequence variants in their GPR rule. These changes included 21 reactions that were identified to contain erroneous GPRs inherited from previous *Klebsiella* models, such as a formamidase reaction (FORAMD) which was initially associated with a gene predicted to encode an oxamate carbamoyltransferase (Table S3).

To confirm that the KpSC pan v2 model was biologically viable, we used it as a reference to construct draft strain-specific models for our 507 KpSC isolates and simulate growth on glucose in M9 minimal media conditions – an expected growth phenotype for all *Klebsiella.* During this process we discovered that some of the models were unable to simulate production of D-mannose 1-phosphate (man1p_c), which was inherited as a biomass requirement from the first *K. pneumoniae* GSMM, iYL1228 [17]. However, D-mannose 1-phosphate is a variable capsule component in *Klebsiella* that is not produced by all strains [49], and hence should not be considered as a generalized component for biomass production. We therefore created a new biomass function without the requirement for D-mannose 1-phosphate (BIOMASS_Core_Feb2022). Following this update, all except four isolates were predicted to simulate growth on glucose. The exceptions included *K. variicola* KSB2_2A and *K. pneumoniae* INF263 for which the draft genome assemblies lacked the *hisA* (KPN_02480) and *plsC* (KPN_03436) genes respectively (Table S1), probably absent due to having incomplete assemblies. Two complete assemblies for *K. pneumoniae* INF018 and *K. variicola* INF232 appeared to

Table 1. Features of KpSC pan v2 and previously published KpSC metabolic models

	iYL1228	iKp1289	KpSC pan v1	KpSC pan v2
Reactions (excl. exchanges)	1973	2178	2017	3190
Exchange reactions	289	306	303	360
Genes	1229	1289	1265	2403
MEMOTE score*	84%	61%	41%	85%
Simulatable carbon sources	252	268	266	345
Simulatable nitrogen sources	149	159	153	192
Simulatable phosphorus sources	59	59	59	68
Simulatable sulphur sources	25	25	25	34

*MEMOTE scores indicate the level of completeness of the model and its associated annotations (e.g. nucleotide accessions, KEGG orthologue IDs); full outputs are available on FigShare.

lack the entire *his* operon required for the production of histidine (confirmed as a genuine absence by aligning the reads from these isolate genomes to the *his* operon from *K. pneumoniae* INF149). *In vitro* experiments confirmed that INF018 and INF232 were histidine auxotrophs (Fig. S3) and these isolates were excluded from further analyses.

The final KpSC pan v2 model contains 3550 reactions (including 360 exchange reactions that define the set of substrates for which growth can be simulated) and 2403 genes, an increase of 46.5–61.7% reactions and 86.4–95.5% genes, compared to the iYL1228, iKp1289 and KpSC pan v1 models (Table 1). For exchange reactions, however, there was only an increase of 24.6% compared to the original iYL1228 model (Table 1), reflecting the limited *Klebsiella* growth phenotype data that are available to inform the addition of these exchanges. Nonetheless, the KpSC pan v2 model supports the prediction of 345 carbon, 192 nitrogen, 68 phosphorus and 34 sulphur growth sources in aerobic and anaerobic conditions (Table 1). However, 44 supported growth phenotypes were not considered biologically plausible, such as the use of the carbon component of secreted carbon dioxide as a sole carbon source (Table S4). The KpSC pan v2 model predicts the greatest number of growth phenotypes for sulphur and phosphorus utilization, whereas it is the second highest to iKp1289 for carbon and nitrogen utilization (Fig. 3). Finally, the KpSC pan v2 model had the highest MEMOTE score of the four models tested (85%, Table 1), which indicates that the model is well annotated and stoichiometrically balanced [50].

KpSC pan v2 captures the greatest metabolic variability

We compared strain-specific GSMMs and associated growth predictions generated using each of the four reference models (n=505 KpSC isolates): iYL1228 [17], iKp1289 [20], KpSC pan v1 [16] and KpSC pan v2. The investigation of core (≥95% of isolates) and accessory (<95% of isolates) reactions was only performed for KpSC pan v1 [16] and v2, since iYL1228 was the direct predecessor of KpSC pan v1 [8] and we were unable to satisfactorily harmonize the iKp1289 reaction nomenclature. Within the KpSC pan v1 model, 2263 of 2320 reactions (97.5%) were considered core, while 56 (2.5%) were accessory. In contrast, 3236 of 3550 reactions (91.2%) in the KpSC pan v2 model were core and 314 (8.8%) were accessory. Grouping reactions by KEGG subsystems [39] highlighted those associated with greater population variation with 28.8% (30/104) of xenobiotic biodegradation reactions in KpSC pan v2 considered accessory, as were 19.9% (87/437) of carbohydrate metabolism reactions, indicating that KpSC pan v2 captures more accessory reactions than KpSC pan v1 (Fig. S4).

Pairwise Jaccard gene distances (Fig. 4a) showed that models built using the KpSC pan v2 reference capture greater strain-specific variation compared to those built using the KpSC pan v1 or single-strain references [median pairwise distance of 0.053 versus 0.013 (iYL1228), 0.023 (iKp1289) and 0.020 (KpSC pan v1), respectively, $P<2.2\times10^{-16}$ for all distribution comparisons by Mann–Whitney test]. Two distinct peaks were also observed in the KpSC pan v2 Jaccard distributions (Fig. 4), with the first peak corresponding to comparisons between isolates of the same species and the second to isolates of different species. This trend can also be seen to some extent in the KpSC pan v1 distribution but is notably absent from the single-strain references (Fig. 4), highlighting that the pangenome approach has allowed the model to capture differences between species that would otherwise not be captured by a single-strain reference.

Distributions of pairwise reaction Jaccard distances (Fig. 4b) supported the same trend as those identified for the gene distances [median pairwise distance of 0.014 versus 0.004 (iYL1228), 0.006 (KpSC pan v1), $P<2.2\times10^{-16}$ by Mann–Whitney test]. However, the distribution for models derived from iKp1289 was unexpectedly broad (median=0.025, range=4.1×10^{-4} to 0.17). The longer skewed tail in the iKp1289 reaction distribution was driven by comparisons involving five genomes that were lacking two genes (VK055_1057 and VK055_0013) associated with 241 transport reactions. This disproportionately impacted the pairwise reaction

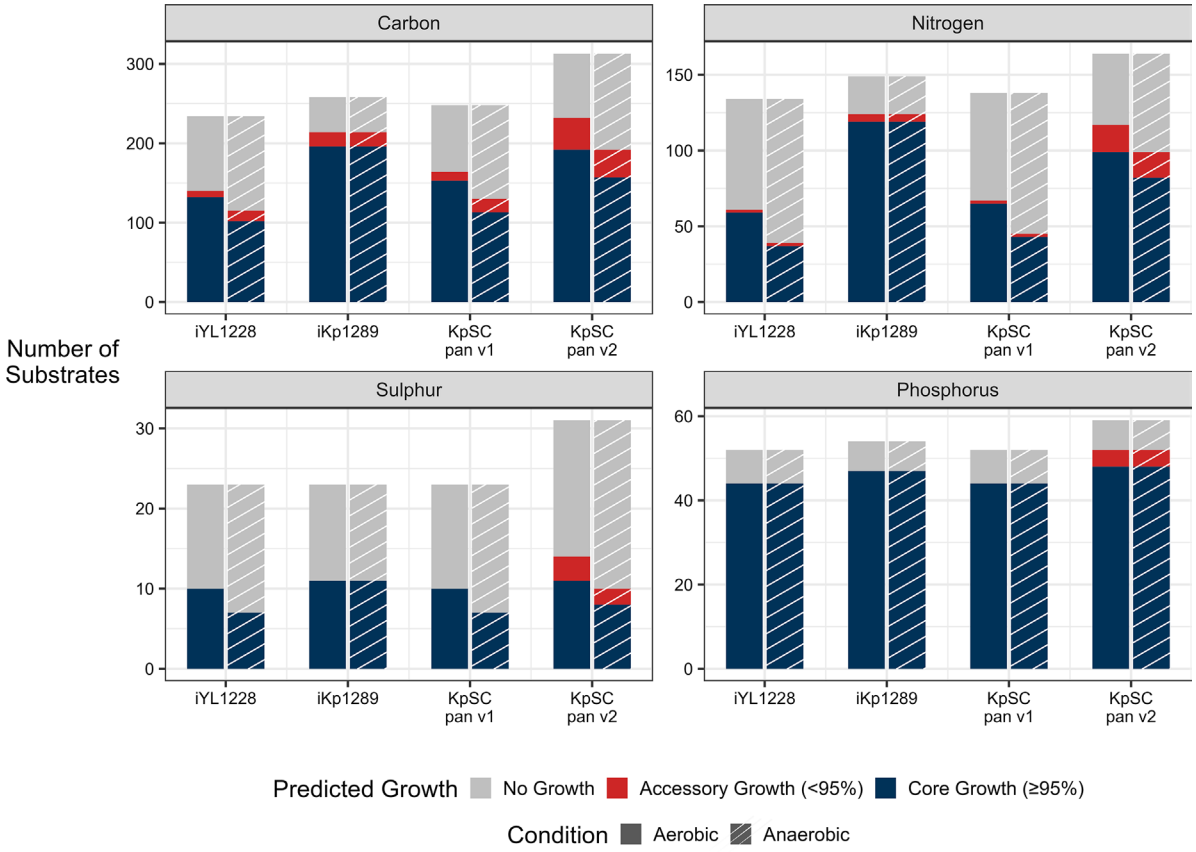

Fig. 3. Growth simulation outcomes for strain-specific genome-scale metabolic models derived from four different references. Simulations represent growth in minimal media supplemented with a single source of carbon, nitrogen, sulphur or phosphorus as indicated. The total number and set of substrates simulated is dependent upon those that can be supported by the reference model (see Table S4; note that substrates supported by the iYL1228 reference are a subset of those supported by KpSC pan v1, which are a subset of those supported by KpSC pan v2. Substrates supported by iKp1289 are an independent and overlapping set). Core growth (blue) refers to substrates for which growth was predicted for ≥95% of the population, whereas accessory growth (red) refers to substrates for which growth was predicted for <95% of the population. 'No growth' (grey) refers to metabolites that are supported by the model, but do not result in positive growth predictions for any strains. These can occur due to incomplete metabolic pathways in the model (e.g. the carbon utilization pathway is complete but the nitrogen pathway is not, resulting in no growth on nitrogen) or where *Klebsiella* is legitimately unable to utilize the substrate for growth.

Jaccard distances (Fig. 4b). Within published GSMMs, these transport reactions are catalysed by the non-specific porins OmpK35 [51, 52], OmpK36 [53, 54], OmpK37 [55, 56] and phosphoporin PhoE [57, 58], that simulate the generic transport of hydrophilic molecules into the cell when the specific enzymes are unknown. Transport of a molecule into the cell will not produce growth unless catalytic enzymes are present, thus sacrificing true molecular accuracy to increase accuracy of simulated phenotypic outcomes. While these porins are treated in the same way in all the reference models, the gene sequences included in the iKp1289 model were not 100% conserved in our KpSC collection, inhibiting their detection and inclusion in the corresponding GSMMs when the default orthologue identity threshold was applied (80%).

As noted above, all KpSC are expected to grow in M9 minimal media supplemented with glucose, and we showed that all but four GSMMs generated using KpSC pan v2 as a reference were able to successfully simulate growth in these conditions. A single additional GSMM generated using the KpSC pan v1 reference was also unable to simulate growth (Table S1). In contrast, 71.0% (360/507) and 38.3% (194/507) of GSMMs generated using the iYL1228 and iKp1289 reference models, respectively, were unable to simulate growth in M9 plus glucose (Table S1). This finding reinforces the notion that single-strain models do not represent sufficient metabolic diversity for use as reference models in population analyses, either because they lack key metabolic pathways that are essential components of metabolism in a subset of the population and/or because they lack the allelic diversity that enables robust orthologue detection. Regardless, in order to support further analysis of growth predictions, we used gap filling to identify and add missing reactions required to simulate growth in minimal conditions [15, 16], restoring positive growth predictions for all models (excluding the known auxotrophs discussed above).

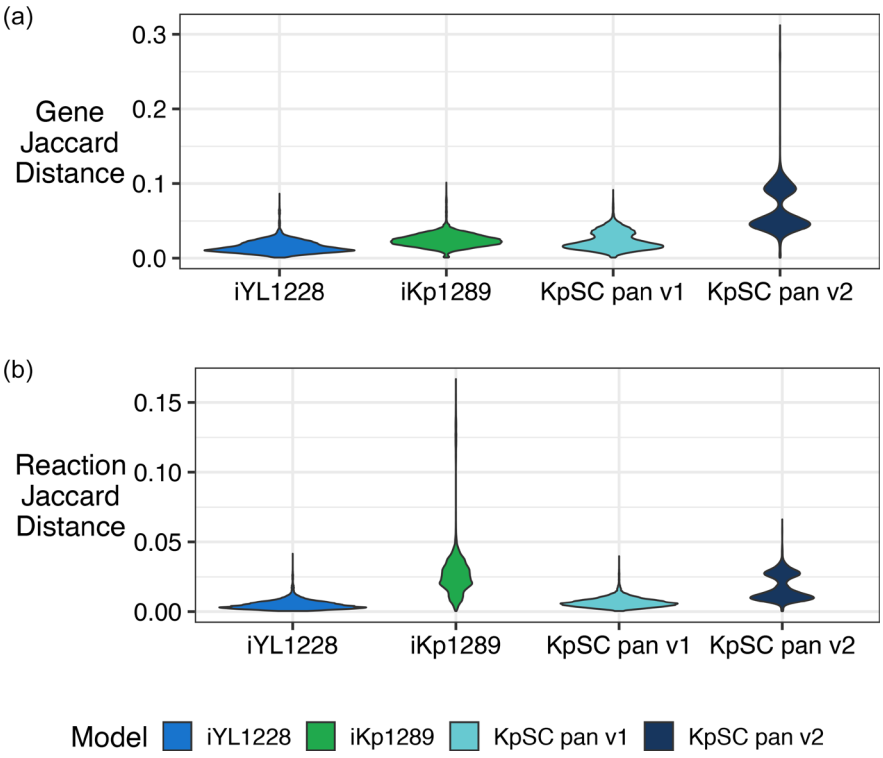

Fig. 4. Pairwise gene and reaction Jaccard distances among strain-specific genome-scale metabolic models (GSMMs) derived from different reference models. Jaccard distances were calculated based on gene (a) and reaction (b) contents of pairs of strain-specific GSMMs derived from each of the iYL1223, iKp1289, KpSC pan v1 and KpSC pan v2 reference models. A Kruskal–Wallis H-test indicated that the distributions were different (Genes: $P<2.2\times10^{-16}$, d.f.=3; Reactions: $P<2.2\times10^{-16}$, d.f.=3). Pairwise comparisons indicated a significant difference between the distributions for KpSC pan v2 and all others (Mann–Whitney test; Genes and Reactions: $P<2.2\times10^{-16}$, Bonferroni correction threshold=0.0167).

We compared the proportion of predicted core growth phenotypes (≥95% of isolates), accessory growth (<95% of isolates) and conserved no growth phenotypes for the collections of GSMMs generated using each of the reference models. Use of the KpSC pan v2 reference resulted in the highest proportion of accessory growth (9.2% versus 2.8, 4.8 and 3.5%, Fig. 3), with sulphur and phosphorus accessory growth phenotypes predicted only for models generated using the KpSC pan v2 reference. In addition, the proportions of conserved no growth phenotypes predicted by models generated with the KpSC pan v2 reference were lower than for those generated with its direct predecessor (KpSC pan v1, 44.6% versus 27.2%, respectively), demonstrating a higher number of completed metabolic pathways for simulatable growth sources.

GSMMs derived from KpSC pan v2 are highly accurate

We compared the accuracy of GSMMs generated using each of the four reference models (KpSC pan v2 and [16, 17, 20]) for growth phenotype predictions in aerobic and anaerobic conditions, for a subset of 37 KpSC isolates with phenotypic growth data for 190 carbon substrates [7]. For anaerobic conditions, we excluded *K. pneumoniae* MGH 78578 for which the data were considered spurious due to a high number of substrates that were not able to support growth, including those that are well characterized and known to generally support *Klebsiella* sp. growth in anaerobic conditions (e.g. glucose [59] and lactose [60]). The specific set of substrates that can be simulated is dependent on the reference model from which each GSMM was derived, and hence in order to facilitate fair comparisons we calculated accuracy metrics for multiple distinct sets of substrates: 89 substrates that were common to all GSMMs, 91 substrates common to GSMMs derived from iYL1228, KpSC pan v1 and KpSC pan v2, 105 substrates common to GSMMs derived from KpSC pan v1 and KpSC pan v2, and finally the full set of 124 substrates that could be simulated by GSMMs derived from KpSC pan v2 only (Fig. 5 and Table 2). The phenotype results contain three substrates (L-lysine [61], glycine [62] and ethanolamine [63]) for which utilization by KpSC is known or expected to be tightly regulated, meaning the standard growth conditions are not expected to stimulate growth. For example, L-lysine utilization is expected to be universal in KpSC [64, 65] and is correctly predicted as such for GSMMs derived from KpSC pan v2 (Table S6), but the required decarboxylase is only active in acidic conditions [61] resulting in no growth in the phenotype data [7]. As GSMMs do not take into account such gene regulation or slow growth rates [13], we have excluded L-lysine, glycine and ethanolamine from the accuracy metrics presented below.

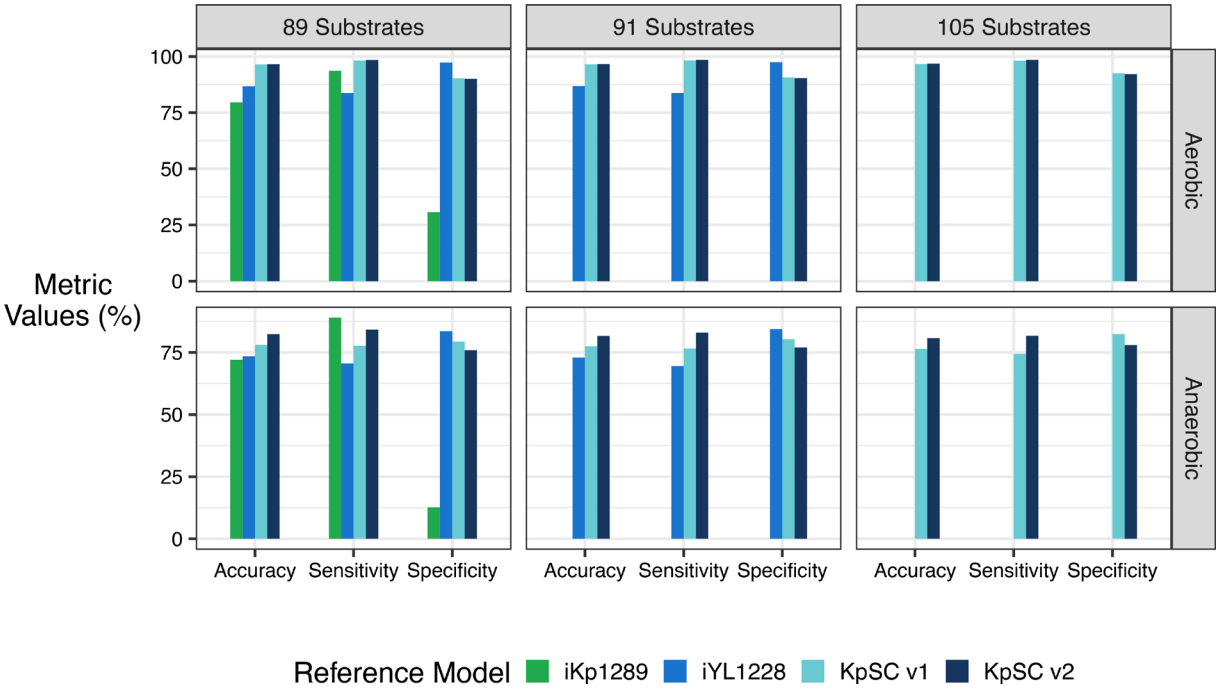

Fig. 5. Comparable growth prediction accuracies for strain-specific genome-scale metabolic models derived from iKp1289, iYL1228, KpSC pan v1 and KpSC pan v2 references. Accuracies, sensitivities and specificities represent combined values for carbon substrate growth predictions compared to true phenotypes generated on the Biolog platform (plates PM1 and PM2) for 37 KpSC isolates. Aerobic conditions are presented in the first row and anaerobic conditions in the second row (raw values for both conditions are available on Figshare). Left panels comprise data for the set of 89 substrates supported by all four models. Middle panels comprise data for the same 89 substrates plus two additional substrates supported by iYL1228, KpSC pan v1 and KpSC pan v2 (91 substrates in total). Right panels comprise data for the same set of 91 substrates plus an additional 14 substrates supported by only KpSC pan v1 and KpSC pan v2. Table 2 shows metrics calculated for the total number of substrates supported by each reference model.

GSMMs derived from the KpSC pan v2 reference showed the highest overall accuracy for each set of comparable growth predictions (96.6% aerobic and 82.3% anaerobic for n=89 substrates, 96.6% aerobic and 81.6% anaerobic for n=91 substrates, 96.8% aerobic and 80.8% anaerobic for n=105 substrates, Fig. 5). This is followed by GSMMs derived from KpSC pan v1 (96.5% aerobic and 78.1% anaerobic for n=89 substrates, 96.5% aerobic and 77.4% anaerobic for n=91 substrates, 96.7% aerobic and 76.4% anaerobic for n=105 substrates, Fig. 5) and those derived from iYL1228 (86.7% aerobic and 73.4% anaerobic for n=89 substrates, 86.9% aerobic and 72.9% anaerobic for n=91 substrates, Fig. 5). GSMMs derived from the iKp1289 reference resulted in the lowest overall accuracy (79.6% aerobic and 72.0% anaerobic for n=89 substrates, Fig. 5), and were associated with particularly low sensitivity (30.7% aerobic and 12.7% anaerobic, Fig. 5). In contrast, there was a trend towards increasing sensitivity and decreasing specificity from GSMMs derived from iYL1228, to those derived from KpSC pan v1, and those derived from KpSC pan v2, implying that as more metabolic content is added to successive versions of these reference models, the ability to accurately predict growth increases at the cost of lower accuracy in predicting no growth (Fig. 5). This behaviour is to be expected, because false positive predictions are assumed to result from regulation that may be preventing growth *in vitro* (not considered by GSMMs) whereas false negative predictions imply that genetic and/or metabolic information is missing from the model [13].

The final accuracies calculated for all 124 substrates that could be simulated by KpSC pan v2-derived GSMMs was 95.4 and 78.8% in aerobic and anaerobic conditions respectively (Table 2), representing a decrease of 1.2 and 2.0% compared to those calculated for the n=105 substrate subset (Tables S6 and S7). Consequently, the final accuracy estimate for KpSC pan v2-derived GSMMs (124 substrates) is lower than those derived from KpSC pan v1 (105 substrates, Table 2), despite the superior accuracy of KpSC pan v2 in the comparative analysis (Fig. 5). This has occurred because most of the substrates newly supported by KpSC pan v2 are less likely to support growth of KpSC strains and are therefore understudied. One such example is 2-deoxy-D-ribose for which the aerobic growth prediction accuracy was 27.0% (Table S6). In *Salmonella* [66], 2-deoxy-D-ribose is degraded by the 2-deoxy-D-ribose utilization operon (*deoQKPX*), with the model encoding *deoK* (deoxyribokinase) and *deoP* (inferred deoxyribose permase). However, neither *deoK* nor *deoP* have been fully characterized and cannot be accurately distinguished from ribokinases (*rbsK*) or L-fucose symporters (*fucP*) respectively at the protein level [39] (both are 100% conserved in our KpSC isolate collection). As the correct genes could not be identified, the associated GPR reactions were left empty

Table 2. Overall model accuracy for aerobic and anaerobic carbon sources using the four *Klebsiella* reference models

		iKp1289	iYL1228	KpSC pan v1	KpSC pan v2
Aerobic	Number of substrates tested*	95	91	105	124
	Overall accuracy (%)	77.92	86.87	96.65	95.38
	Strain-specific range (%)	72.63–82.11	81.32–91.21	92.38–98.10	87.90–97.58
	True positives	2480	2174	2817	3245
	True negatives	259	751	938	1131
	False positives	525	20	76	155
	False negatives	251	422	54	57
	Sensitivity (%)	90.81	83.74	98.12	98.27
	Specificity (%)	33.04	97.41	92.50	87.95
Anaerobic	Overall accuracy (%)	70.23	72.92	76.38	78.76
	Strain-specific range (%)	43.16–81.05	59.34–84.62	61.90–86.67	59.68–87.90
	True positives	2272	1758	2129	2594
	True negatives	130	631	758	922
	False positives	664	117	162	227
	False negatives	354	770	731	721
	Sensitivity (%)	86.52	69.54	74.44	78.25
	Specificity (%)	16.37	84.36	82.39	80.24

*The total number of substrates supported by the model for which matched aerobic and anaerobic phenotype data generated on the Biolog platform were available. Note that substrates supported by iYL1228 are a subset of those supported by KpSC pan v1, which are a subset of those supported by KpSC pan v2. Substrates supported by iKp1289 are an independent but overlapping set.

(i.e. all GSMMs derived from the reference can inherit the reaction), resulting in an incorrect core growth prediction. In spite of this, we have elected to retain the pathway in the model because there is sufficient evidence that *Klebsiella* can utilize this substrate and the pathway is known [66]. The model can easily be updated when further information is available to identify the true sequences.

As previous phenotype experiments do not contain all possible substrates that can be newly simulated by the KpSC GSMMs, we performed independent growth assays for six substrates, for which the estimated accuracies ranged from 32.4 to 90.5% (Table 3). Phenylpropanoate utilization was the most accurately predicted, with high sensitivity (77.8%) and specificity (93.9%, Table 3).

Table 3. Summary of additional phenotypic validation experiments for KpSC pan v2

Substrate*	Predicted growth†	TP	TN	FP	FN	Accuracy (%)
2-Aminoethylphosphonate	31/37	6	6	25	0	32.4
Beta-alanine	37/37	17	0	20	0	46.0
Allantoin	3/37	2	31	1	3	89.2
L-Ectoine	5/41	1	34	4	2	85.4
Phenylpropanoate	10/42	7	31	2	2	90.5
Uracil	37/37	28	0	9	0	75.7

*Substrates were not present on Biolog PM1/PM2 plates, and were tested via an independent culture approach in aerobic conditions only (see Methods).
†Isolates tested include 37 KpSC for which Biolog microarray data were available. Two phenotypes were predicted as rare among this set of 37 KpSC, and hence additional isolates for which growth was predicted were selected from our collection; *n*=4 available for L-ectoine and *n*=5 for phenylpropanoate (see Table S5).

Growth on L-ectoine, allantoin and uracil were also predicted with accuracies >75%; however, we note that the total number of isolates predicted to be able to utilize L-ectoine and/or allantoin was low (≤5). This is to be expected, with allantoinase (required for allantoin utilization) found to be present in only 5% of *K. pneumoniae* and 50% of *K. quasipneumoniae* [10]. The low accuracies for 2-aminoethylphosphonate (32.4%) and beta-alanine (46.0%) may be explained by mechanisms not taken into account by GSMMs [13], such as gene regulation or experimental conditions. In the case of 2-aminoethylphosphonate usage, metabolism in other *Enterobacteriaceae* species only occurs in conditions of phosphate starvation [67], while beta-alanine metabolism may require nitrogen starvation [68]. Alternatively, the discrepancies may be due to limitations with using cellular growth (OD_{600}) instead of tetrazolium dye as a proxy of cellular respiration to determine substrate utilization. Regardless, when these six additional substrates were included, the overall accuracy for aerobic growth conditions for KpSC pan v2-derived GSMMs was 94.3% (sensitivity=98.1% and specificity=85.4%).

DISCUSSION

We present a highly curated and accurate KpSC pan-metabolic model, capturing the highest number of distinct metabolic reactions of any pan-model published to date for any bacterial taxa [69–72]. We show that KpSC pan v2 successfully captures a greater proportion of variable metabolic traits than its predecessors (Figs 3 and 4) and can be used as a reference to generate strain-specific GSMMs with superior completeness (in terms of metabolic pathways supporting core glucose metabolism) and accuracy (for growth phenotype predictions, Fig. 5). When compared to growth phenotype data, strain-specific GSMMs derived from KpSC pan v2 were associated with 95.4 and 78.8% accuracy for aerobic and anaerobic conditions, respectively (n=124 substrates, Table 2). However, six additional substrates tested via our in-house OD_{600} approach, in aerobic conditions only, were associated with a wider range of prediction accuracies (32.4–90.5%, Table 3). The lower accuracy predictions arose from a higher number of false positives (sensitivity=33.3–100%, specificity=0–96.9%, Table 3), probably from a combination of methodological limitations and regulatory effects that are not accounted for by GSMMs, the latter being a key caveat of *in silico* growth predictions that has been recognized previously [13].

The KpSC pan v2 model represents 507 individual KpSC isolates that were selected based on the availability of existing curated GSMMs [8, 16, 17, 20], and to represent broad population diversity. The previously published GSMMs included 37 isolates with matched phenotype data [7] that represented all KpSC subspecies. To further diversify this set we included 452 isolates from published clinical diversity collections [18, 73, 74], which have been extensively characterized and for which the physical isolates were available in our laboratory. These collections represent clinical KpSC isolated from an Australian hospital network over a 12 month period [18], plus isolates from rectal swabs of patients in select wards [73, 74]. Isolates were included regardless of their antibiotic resistance status (23.7% multidrug resistant), and were genetically diverse [18, 73, 74]. Overall, the 507 isolates represented in KpSC pan v2 comprise 268 unique STs and include representatives of 12 out of 13 globally disseminated problematic clones described in the literature [9] (lacking clonal group 307 which was not present in our hospital at the time of isolate collection, Table S1). However, we acknowledge that this collection may not fully represent the true global KpSC population, which comprises hundreds to thousands of unique clones [18, 46, 74]. Additionally, most isolates were either human clinical or gut carriage isolates (97.5% of isolates), with only a small proportion from environmental sources or other host-associated environments (Table S1). Hence, it is likely that the model excludes metabolic traits specific to non-human niches. Furthermore, the inclusion of metabolic traits is limited by the bounds of the existing biochemical and biological knowledge bases, which are undoubtedly biased towards common metabolic processes. As a consequence, we anticipate that rare metabolic traits are underrepresented in the model, and accordingly the relative proportion of accessory versus core reactions (8.8% accessory and 91.2% core) captured in the model is far smaller than that estimated for the total KpSC gene pool [10, 18].

The total pangenome derived from the 507 isolates included in this work comprised 34 664 unique genes and the final curated KpSC pan v2 model includes just 2403 of these (~7%). However, a previous automated analysis indicated that up to 37% of the KpSC pangenome encodes proteins involved with metabolic processes [10], further highlighting the limitations in the underlying knowledge bases that inform metabolic models (in particular where reaction stoichiometries are not defined). In addition, it is important to consider the types of evidence available to support the inclusion of genes and reactions in the model, which in most cases is limited to sequence homology and/or simple growth phenotype data (evidence score 2, 81.9%). Novel proteomic and/or protein structural data generated specifically for KpSC isolates may enable the inclusion of additional reactions with greater confidence, as demonstrated for construction of the *S. enterica* pan-model [75]. Targeted gene-knockout and/or gene expression analyses conducted in aerobic and anaerobic conditions could also shed new light on the associated metabolic differences and inform improvements for anaerobic growth predictions.

Despite its limitations, KpSC pan v2 represents a significant resource for the *Klebsiella* research community, both as a source of curated KpSC-specific metabolic information and as a reference model from which highly accurate strain-specific GSMMs can be readily derived. The model will be updated as additional biochemical data and evidence become available. It is freely available for use, reuse and adaptation under open access licence and we welcome contributions from the community. We anticipate that KpSC pan v2 will inform studies of KpSC virulence and antimicrobial resistance, as well as KpSC ecology

(e.g. within the human gut microbiota), supporting the development of novel therapeutics and containment strategies targeting KpSC. Furthermore, we hope that our data highlighting the benefits of the pan-model approach will inspire similar works for other prominent bacterial pathogens, particularly those known to harbour high levels of genomic diversity such as *E. coli*, *Pseudomonas aeruginosa* and *Acinetobacter baumannii*.

Funding information
This research was supported by an Australian Government Research Training Programme (RTP) Scholarship to H.B.C. and an Australian Research Council Discovery Project award (DP200103364, to K.L.W., K.E.H., S.B. and J.M.M.). K.L.W. is supported by a National Health and Medical Research Council of Australia Investigator Grant (APP1176192). Biolog data generation was supported financially by the French Government's Investissement d'Avenir programme Laboratoire d'Excellence 'Integrative Biology of Emerging Infectious Diseases' (ANR-10-LABX-62-IBEID).

Acknowledgements
The authors would like to thank Kalani Paranagama (Monash University) for discussions and advice regarding this project.

Author contributions
Conceptualization: K.L.W., K.E.H. & J.M.M. Data Curation: H.B.C. Formal Analysis: H.B.C. Funding Acquisition: H.B.C., K.L.W., K.E.H., S.B., J.M.M. Investigation: H.B.C, S.B., V.P., S.L.F. Methodology: H.B.C. & K.L.W. Project Administration: K.L.W. Resources: K.L.W., K.E.H. & S.B. Software: H.B.C. & B.V. Supervision: K.L.W., J.H. & K.E.H. Validation: H.B.C., B.V., V.P. & S.L.F. Visualization: H.B.C. Writing - Original Draft: H.B.C. & K.L.W. Writing - Review & Editing: All authors.

Conflicts of interest
The author(s) declare no conflict of interest.

References

1. Murray CJL, Ikuta KS, Sharara F, Swetschinski L, Robles Aguilar G, *et al*. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 2022;399:629–655.
2. Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, *et al*. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect Dis* 2018;18:318–327.
3. Jenior ML, Dickenson ME, Papin JA. Genome-scale metabolic modeling reveals increased reliance on valine catabolism in clinical isolates of *Klebsiella pneumoniae*. *NPJ Syst Biol Appl* 2022;8:41.
4. Hudson AW, Barnes AJ, Bray AS, Ornelles DA, Zafar MA. *Klebsiella pneumoniae* L-fucose metabolism promotes gastrointestinal colonization and modulates its virulence determinants. *Infect Immun* 2022;90:e0020622.
5. Vornhagen J, Sun Y, Breen P, Forsyth V, Zhao L, *et al*. The *Klebsiella pneumoniae* citrate synthase gene, gltA, influences site specific fitness during infection. *PLoS Pathog* 2019;15:e1008010.
6. Martin RM, Cao J, Wu W, Zhao L, Manthei DM, *et al*. Identification of pathogenicity-associated loci in *Klebsiella pneumoniae* from hospitalized patients. *mSystems* 2018;3:e00015-18.
7. Blin C, Passet V, Touchon M, Rocha EPC, Brisse S. Metabolic diversity of the emerging pathogenic lineages of *Klebsiella pneumoniae*. *Environ Microbiol* 2017;19:1881–1898.
8. Hawkey J, Vezina B, Monk JM, Judd LM, Harshegyi T, *et al*. A curated collection of *Klebsiella* metabolic models reveals variable substrate usage and gene essentiality. *Genome Res* 2022;32:1004–1014.
9. Wyres KL, Lam MMC, Holt KE. Population genomics of *Klebsiella pneumoniae*. *Nat Rev Microbiol* 2020;18:344–359.
10. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, *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;112:E3574–E3581.
11. Ramos PIP, Fernández Do Porto D, Lanzarotti E, Sosa EJ, Burguener G, *et al*. An integrative, multi-omics approach towards the prioritization of *Klebsiella pneumoniae* drug targets. *Sci Rep* 2018;8:10755.
12. Chen X-H, Liu S-R, Peng B, Li D, Cheng Z-X, *et al*. Exogenous l-valine promotes phagocytosis to kill multidrug-resistant bacterial pathogens. *Front Immunol* 2017;8:207.
13. Thiele I, Palsson BØ. A protocol for generating A high-quality genome-scale metabolic reconstruction. *Nat Protoc* 2010;5:93–121.
14. O'Brien EJ, Monk JM, Palsson BO. Using genome-scale models to predict biological capabilities. *Cell* 2015;161:971–987.
15. Norsigian CJ, Fang X, Seif Y, Monk JM, Palsson BO. A workflow for generating multi-strain genome-scale metabolic models of prokaryotes. *Nat Protoc* 2020;15:1–14.
16. Vezina B, Watts SC, Hawkey J, Cooper HB, Judd LM, *et al*. Bactabolize is a tool for high-throughput generation of bacterial strain-specific metabolic models. *eLife* 2023;12:RP87406.
17. Liao Y-C, Huang T-W, Chen F-C, Charusanti P, Hong JSJ, *et al*. An experimentally validated genome-scale metabolic reconstruction of *Klebsiella pneumoniae* MGH 78578, iYL1228. *J Bacteriol* 2011;193:1710–1717.
18. Gorrie CL, Mirčeta M, Wick RR, Judd LM, Lam MMC, *et al*. Genomic dissection of *Klebsiella pneumoniae* infections in hospital patients reveals insights into an opportunistic pathogen. *Nat Commun* 2022;13:3017.
19. Norsigian CJ, Attia H, Szubin R, Yassin AS, Palsson BØ, *et al*. Comparative genome-scale metabolic modeling of metallo-beta-lactamase-producing multidrug-resistant *Klebsiella pneumoniae* clinical isolates. *Front Cell Infect Microbiol* 2019;9:161.
20. Henry CS, Rotman E, Lathem WW, Tyo KEJ, Hauser AR, *et al*. Generation and validation of the iKp1289 metabolic model for *Klebsiella pneumoniae* KPPR1. *J Infect Dis* 2017;215:S37–S43.
21. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13:e1005595.
22. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
23. Liu P, Li P, Jiang X, Bi D, Xie Y, *et al*. Complete genome sequence of *Klebsiella pneumoniae* subsp. pneumoniae HS11286, a multidrug-resistant strain isolated from human sputum. *J Bacteriol* 2012;194:1841–1842.
24. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, *et al*. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* 2020;21:180.
25. Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, *et al*. A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex. *Nat Commun* 2021;12:4188.
26. Lam MMC, Wick RR, Judd LM, Holt KE, Wyres KL. Kaptive 2.0: updated capsule and lipopolysaccharide locus typing for the *Klebsiella pneumoniae* species complex. *Microb Genom* 2022;8:000800.
27. Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, *et al*. iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nat Biotechnol* 2017;35:904–908.

28. Monk JM, Koza A, Campodonico MA, Machado D, Seoane JM, *et al*. Multi-omics quantification of species variation of *Escherichia coli* links molecular features with strain phenotypes. *Cell Syst* 2016;3:238–251.
29. Archer CT, Kim JF, Jeong H, Park JH, Vickers CE, *et al*. The genome sequence of *E. coli* W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of *E. coli*. *BMC Genomics* 2011;12:9.
30. Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, *et al*. Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc Natl Acad Sci U S A* 2013;110:20338–20343.
31. Charusanti P, Chauhan S, McAteer K, Lerman JA, Hyduke DR, *et al*. An experimentally-supported genome-scale metabolic network reconstruction for *Yersinia pestis* CO92. *BMC Syst Biol* 2011;5:163.
32. Thiele I, Hyduke DR, Steeb B, Fankam G, Allen DK, *et al*. A community effort towards a knowledge-base and mathematical model of the human pathogen *Salmonella* Typhimurium LT2. *BMC Syst Biol* 2011;5:8.
33. Raghunathan A, Reed J, Shin S, Palsson B, Daefler S. Constraint-based analysis of metabolic capacity of *Salmonella* Typhimurium during host-pathogen interaction. *BMC Syst Biol* 2009;3:38.
34. King ZA, Lu J, Dräger A, Miller P, Federowicz S, *et al*. BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res* 2016;44:D515–D522.
35. Malik-Sheriff RS, Glont M, Nguyen TVN, Tiwari K, Roberts MG, *et al*. BioModels-15 years of sharing computational models in life science. *Nucleic Acids Res* 2020;48:D407–D415.
36. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, *et al*. KBase: the United States department of energy systems biology knowledgebase. *Nat Biotechnol* 2018;36:566–569.
37. Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, *et al*. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res* 2016;44:D73–D80.
38. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
39. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;45:D353–D361.
40. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, *et al*. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* 2020;36:2251–2252.
41. Seaver SMD, Liu F, Zhang Q, Jeffryes J, Faria JP, *et al*. The ModelSEED biochemistry database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Res* 2021;49:D575–D588.
42. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, *et al*. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci Rep* 2015;5:8365.
43. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-based reconstruction and analysis for python. *BMC Syst Biol* 2013;7:74.
44. Wyres KL, Wick RR, Gorrie C, Jenney A, Follador R, *et al*. Identification of *Klebsiella* capsule synthesis loci from whole genome data. *Microb Genom* 2016;2:e000102.
45. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res* 2018;46:7542–7553.
46. Hawkey J, Wyres KL, Judd LM, Harshegyi T, Blakeway L, *et al*. ESBL plasmids in *Klebsiella pneumoniae*: diversity, transmission and contribution to infection burden in the hospital setting. *Genome Med* 2022;14:97.
47. Becker LC, Bergfeld WF, Belsito DV, Klaassen CD, Marks JG, *et al*. Final report of the safety assessment of allantoin and its related complexes. *Int J Toxicol* 2010;29:84S–97S.
48. Botos I, Majdalani N, Mayclin SJ, McCarthy JG, Lundquist K, *et al*. Structural and functional characterization of the LPS transporter LptDE from gram-negative pathogens. *Structure* 2016;24:965–976.
49. Shu H-Y, Fung C-P, Liu Y-M, Wu K-M, Chen Y-T, *et al*. Genetic diversity of capsular polysaccharide biosynthesis in *Klebsiella pneumoniae* clinical isolates. *Microbiology*2009;155:4170–4183.
50. Lieven C, Beber ME, Olivier BG, Bergmann FT, Ataman M, *et al*. MEMOTE for standardized genome-scale metabolic model testing. *Nat Biotechnol* 2020;38:272–276.
51. Doménech-Sánchez A, Martínez-Martínez L, Hernández-Allés S, del Carmen Conejo M, Pascual A, *et al*. Role of *Klebsiella pneumoniae* OmpK35 porin in antimicrobial resistance. *Antimicrob Agents Chemother* 2003;47:3332–3335.
52. Acosta-Gutiérrez S, Ferrara L, Pathania M, Masi M, Wang J, *et al*. Getting drugs into gram-negative bacteria: rational rules for permeation through general porins. *ACS Infect Dis* 2018;4:1487–1498.
53. Albertí S, Rodríquez-Quiñones F, Schirmer T, Rummel G, Tomás JM, *et al*. A porin from *Klebsiella pneumoniae*: sequence homology, three-dimensional model, and complement binding. *Infect Immun* 1995;63:903–910.
54. Wong JLC, Romano M, Kerry LE, Kwong H-S, Low W-W, *et al*. OmpK36-mediated carbapenem resistance attenuates ST258 *Klebsiella pneumoniae* in vivo. *Nat Commun* 2019;10:3957.
55. 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;181:2726–2732.
56. Rocker A, Lacey JA, Belousoff MJ, Wilksch JJ, Strugnell RA, *et al*. Global trends in proteome remodeling of the outer membrane modulate antimicrobial permeability in *Klebsiella pneumoniae*. *mBio* 2020;11:e00603-20.
57. Hu G, Chen X, Chu W, Ma Z, Miao Y, *et al*. Immunogenic characteristics of the outer membrane phosphoporin as a vaccine candidate against *Klebsiella pneumoniae*. *Vet Res* 2022;53:5.
58. Cowan SW, Schirmer T, Rummel G, Steiert M, Ghosh R, *et al*. Crystal structures explain functional properties of two *E. coli* porins. *Nature* 1992;358:727–733.
59. Vanhooren PT, Baets S, Bruggeman G, Vandamme EJ. Klebsiella. In: Robinson RK (ed). *Encyclopedia of Food Microbiology*. Oxford: Elsevier; 1999. pp. 1107–1115.
60. Cabelli VJ. Lactose utilization in *Klebsiella pneumoniae*: the slow utilization of lactose by resting cells of lactose fermenting strains. *J Bacteriol* 1955;70:15–22.
61. Kanjee U, Gutsche I, Ramachandran S, Houry WA. The enzymatic activities of the *Escherichia coli* basic aliphatic amino acid decarboxylases exhibit a pH zone of inhibition. *Biochemistry* 2011;50:9388–9398.
62. Hammes W, Schleifer KH, Kandler O. Mode of action of glycine on the biosynthesis of peptidoglycan. *J Bacteriol* 1973;116:1029–1053.
63. Liu Y, Zhu S, Wei L, Feng Y, Cai L, *et al*. Arm race among closely-related carbapenem-resistant *Klebsiella pneumoniae* clones. *ISME Commun* 2022;2:76.
64. Hansen DS, Aucken HM, Abiola T, Podschun R. Recommended test panel for differentiation of *Klebsiella* species on the basis of a trilateral interlaboratory evaluation of 18 biochemical tests. *J Clin Microbiol* 2004;42:3665–3669.
65. EdwardsPR, Ewing WH. Identification of enterobacteriaceae. Identification of Enterobacteriaceae; 1962. https://www.cabdirect.org/cabdirect/abstract/19632702221
66. Christensen M, Borza T, Dandanell G, Gilles A-M, Barzu O, *et al*. Regulation of expression of the 2-deoxy-D-ribose utilization regulon, deoQKPX, from *Salmonella enterica* serovar typhimurium. *J Bacteriol* 2003;185:6042–6050.
67. Kim AD, Baker AS, Dunaway-Mariano D, Metcalf WW, Wanner BL, *et al*. The 2-aminoethylphosphonate-specific transaminase of the 2-aminoethylphosphonate degradation pathway. *J Bacteriol* 2002;184:4134–4140.

68. Kurihara S, Oda S, Kato K, Kim HG, Koyanagi T, *et al*. A novel putrescine utilization pathway involves gamma-glutamylated intermediates of *Escherichia coli* K-12. *J Biol Chem* 2005;280:4602–4608.
69. Seif Y, Kavvas E, Lachance J-C, Yurkovich JT, Nuccio S-P, *et al*. Genome-scale metabolic reconstructions of multiple *Salmonella* strains reveal serovar-specific metabolic traits. *Nat Commun* 2018;9:3771.
70. Seif Y, Monk JM, Machado H, Kavvas E, Palsson BO. Systems biology and pangenome of *Salmonella* O-antigens. *mBio* 2019;10.
71. Blázquez B, San León D, Rojas A, Tortajada M, Nogales J. New insights on metabolic features of *Bacillus subtilis* based on multistrain genome-scale metabolic modeling. *Int J Mol Sci* 2023;24:7091.
72. Ardalani O, Phaneuf P, Mohite OS, Nielsen LK, Palsson BO. Pangenome reconstruction of *Lactobacillaceae* metabolism predicts species-specific metabolic traits. *Syst Biol* 2023.
73. Gorrie CL, Mirceta M, Wick RR, Edwards DJ, Thomson NR, *et al*. Gastrointestinal carriage is a major reservoir of *Klebsiella pneumoniae* infection in intensive care patients. *Clin Infect Dis* 2017;65:208–215.
74. Gorrie CL, Mirceta M, Wick RR, Judd LM, Wyres KL, *et al*. Antimicrobial-resistant *Klebsiella pneumoniae* carriage and infection in specialized geriatric care wards linked to acquisition in the referring hospital. *Clin Infect Dis* 2018;67:161–170.
75. Steeb B, Claudi B, Burton NA, Tienz P, Schmidt A, *et al*. Parallel exploitation of diverse host nutrients enhances *Salmonella* virulence. *PLoS Pathog* 2013;9:e1003301.