A metabolic atlas of the *Klebsiella pneumoniae* species complex reveals lineage-specific metabolism that supports persistent co-existence of diverse lineages

Authors

Ben Vezina[1,2*], Helena B. Cooper[1,2], Martin Rethoret-Pasty[3], Sylvain Brisse[3], Jonathan M. Monk[4], Kathryn E. Holt[1,5] and Kelly L. Wyres[1,2*]

[1] Department of Infectious Diseases, School of Translational Medicine, Monash University, Melbourne, Victoria, Australia

[2] Centre to Impact AMR, Monash University, Clayton, Victoria, Australia

[3] Institut Pasteur, Université Paris Cité, Biodiversity and Epidemiology of Bacterial Pathogens, Paris, France

[4] Department of Medicine, University of California, San Diego School of Medicine, San Diego, CA, United States of America

[5] Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, UK

*Corresponding authors Ben Vezina (benjamin.vezina@monash.edu); Kelly L. Wyres (Kelly.wyres@monash.edu)

Abstract

The *Klebsiella pneumoniae* species complex inhabits a wide variety of hosts and environments. Genomics has revealed the population comprises multiple species/subspecies and hundreds of distinct co-circulating sub-lineages that are associated with distinct gene complements. These data are consistent with metabolic differentiation as a driver of population structure, but this has so far remained unsubstantiated. Here we used comparative genomics and genome-scale metabolic modelling to systematically explore metabolic diversity across the population (n=7,835 genomes). We simulated growth outcomes for carbon, nitrogen, phosphorus and sulfur sources under aerobic and anaerobic conditions (n=1,278), and confirmed that sub-lineages exhibit unique metabolic profiles. *In vitro* co-culture experiments demonstrated reciprocal commensalistic cross-feeding between sub-lineages, effectively extending the range of conditions supporting individual growth. We propose that these substrate specialisations promote the existence and persistence of co-circulating sub-lineages by reducing nutrient competition and facilitating commensal interactions.

Introduction

The *Klebsiella pneumoniae* species complex (*Kp*SC) is a group of closely related bacteria with a diverse gene pool, spread across seven taxa (hereafter species) and divided into thousands of phylogenetically distinct sub-lineages [1, 2]. A portion of these sub-lineages are clinically significant clones contributing substantially to disease burden, either as multi-drug resistant agents of healthcare-associated infections or as 'hypervirulent' community-acquired pathogens (usually drug-susceptible) [1]. As gene content differences are known to play a role in driving these differential epidemiologies, we hypothesise that this diverse gene pool also explains the cosmopolitan distribution of *Kp*SC, which are found on every continent across a wide variety of environments and hosts [1, 3]. However, individual sub-lineages can be identified from multiple sources including human, animals and environmental sources, and even highly localised studies typically identify substantial diversity [4, 5]. This raises the question of how so many distinct sub-lineages can be maintained in the population - an outstanding puzzle across many bacterial species. A key unmapped feature to this question is the distribution of metabolic traits and how these intersect with global pervasiveness and epidemiological patterns.

A population-metabolism link has been elegantly proposed [6], whereby metabolic diversity may be partially responsible for maintaining bacterial population structure. Accordingly, genomic studies have predicted the non-random distribution of metabolic traits among *Vibrio parahaemoliticus* [7], *Yersinia enterocolitica* [8], *Campylobacter jejuni* [9], *Enterococcus faecium* [10] and *Escherichia coli* [11]. In addition, multi-drug resistant *E. coli* lineages have been shown to have different metabolic genes compared to non-multi-drug resistant counterparts [11]. However, the complexities of cellular metabolism make systematic prediction of phenotypes from genome data difficult, limiting capacity for large-scale analyses of population metabolism.

Genome-scale metabolic models attempt to encapsulate the total metabolic potential of an organism and can be used to predict metabolic phenotypes [12]. Comparative metabolic modelling analyses can be utilised to probe diversity within bacterial species populations, and recent larger-scale studies (n = 50-2,446 isolates) have highlighted the power of these approaches to identify industrially relevant, ecology-, drug-resistance- and pathogenicity-associated metabolic traits [13-18]. However, none have attempted to investigate the interplay between metabolic diversity and fine-scale population structure.

Here, we describe a population-level analysis of 7,835 *Kp*SC genomes to evaluate the distribution and diversity of metabolism in the context of the population structure. We performed quantitative analyses of metabolic gene content, constructed metabolic models and predicted 1,278 metabolic phenotypes for each isolate, a subset of which were validated via experimental assays. Our data reveal considerable variation between and within species that is non-randomly distributed across sub-lineages. Notably, we show that sub-lineages, including multi-drug resistant and hypervirulent clones, are associated with distinct substrate usage profiles that facilitate commensal interactions and likely promote co-existence.

Results

Metabolic traits are bimodally distributed across the population

The study dataset included 7,835 high-quality *Kp*SC genomes covering all seven taxa in the complex (**Fig. 1A**) and 1,931 sequence types (STs). Genomes were obtained from 29 medium to large-scale studies and dereplicated to limit sampling biases (see **Methods, Table S1**). Pan-genome analysis [19] identified 61,595 gene clusters, 11,184 (18.2%) of

which were matched to a total of 2,601 unique KEGG Orthologs [20] with associated substrate-reaction data. This represented a lower bound estimate of the proportion of 'metabolic genes' in the *Kp*SC pan-genome, as 38.9% of gene clusters could not be assigned functional annotations (hypothetical proteins). Metabolism-associated KEGG Orthologs (hereafter 'metabolic orthologs') were bimodally distributed (**Fig. 1B, Table S2**), with 1,499 orthologs (57.4%) present in ≥95% of genomes and 899 rare orthologs present in <15% of genomes.

To analyse the phenotypic diversity across the *Kp*SC, a strain-specific metabolic model was generated for each genome using Bactabolize [21], which produces models with 93.2% and 78.1% predictive accuracy for aerobic and anaerobic growth phenotypes, respectively [22]. Growth phenotypes were predicted for 345 carbon, 192 nitrogen, 68 phosphorus and 34 sulfur substrates as sole sources in minimal media under aerobic and anaerobic conditions (total 1,278 conditions). Individual strain-specific models comprised a mean of 1677.0 genes ± 17.3 (standard deviation, SD) and 3330.1 reactions ± 13.5 (SD). Individual isolates were predicted to grow in a mean of 678.2 ± 17.9 (SD) conditions (**Table S3**, **Fig. S1A, Supplemental Results**).

The distribution of positive growth predictions mirrored that for metabolic gene content. At the population level, 643 conditions (47.6%) were predicted to support growth for ≥95% isolates ('core', **Fig. 1C**, **Table S3**) and were comprised of 349 aerobic and 294 anaerobic conditions, which systematically confirmed the facultative anaerobe status of the *Kp*SC. These 643 core conditions consisted of 207 total substrates (note that some substrates can be tested as sole sources of multiple elements e.g. carbon and nitrogen), with at least 204 of these being human metabolites, 134 diet-related compounds and 203 gut microbiome metabolites (**Table S3**). In contrast, 509 modelled growth conditions (43.6%) did not support growth of any isolates (224 aerobic and 285 anaerobic conditions). Carbon substrate usage displayed the greatest variability, with 41 aerobic and 37 anerobic conditions predicted to support growth in <95% isolates (**Fig. 1C, Fig. S1B**).

<FIG. 1>

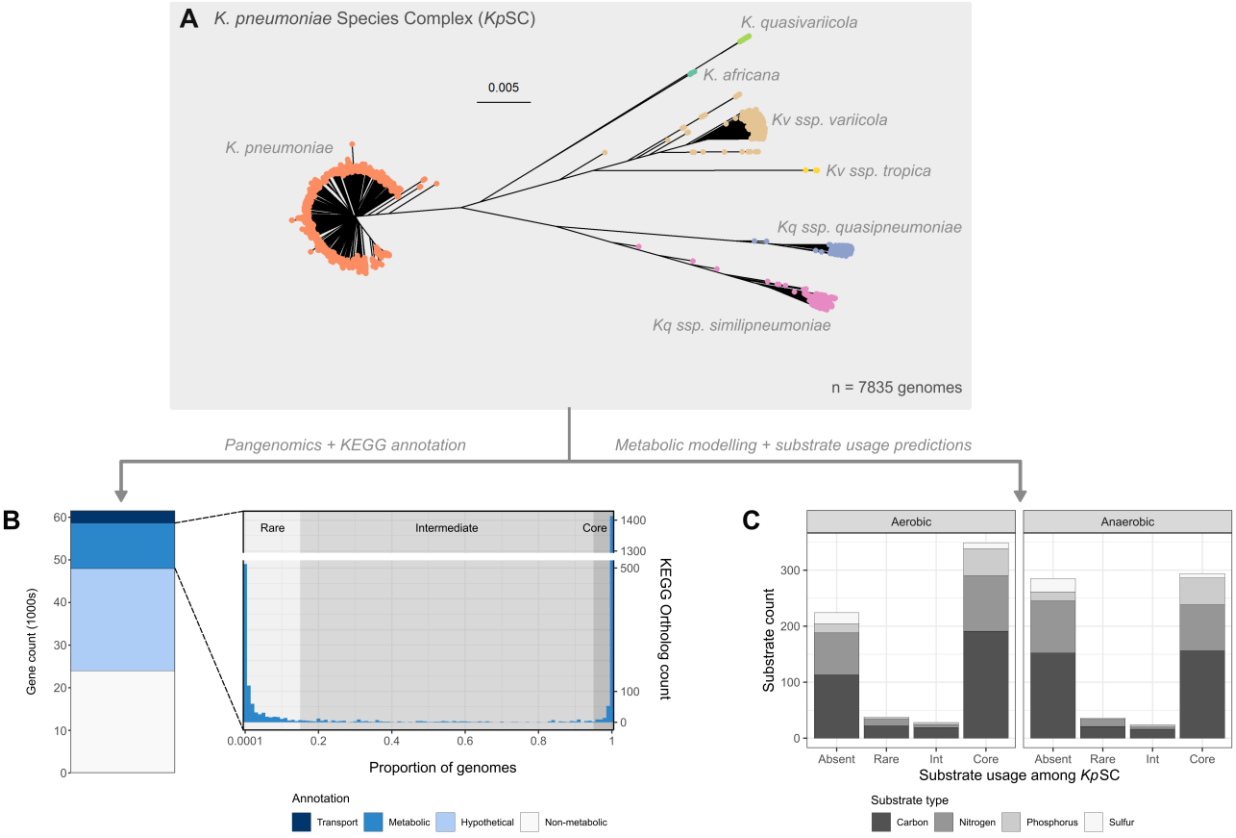


Fig. 1: Population level metabolic diversity in the *K. pneumoniae* Species Complex

A: Neighbour-joining tree of the 7,835 *Kp*SC genomes used in this study. Tip colours indicate species as labelled. An interactive version can be found at https://microreact.org/project/pcTMwQZAGCsqqhbEZzaj11-a-metabolic-atlas-of-the-klebsiella-pneumoniae-species-complex. **B:** Stacked bars show pangenome content by metabolic gene (KEGG) category as indicated in legend (left). Population distribution of KEGG orthologs associated with metabolic genes ('metabolic orthologs', right). The light grey background shows rare orthologs (>0 to 15%), the medium grey shows intermediate orthologs (>15% to <95%) while the dark grey shows core orthologs (≥95%). **C:** Stacked bars show population frequency categories for predicted substrate usage, coloured by substrate type as indicated in legend.

Species employ differential metabolic specialisation

The set of core growth conditions for each *Kp*SC taxon (species or subspecies) was greater than the species complex as a whole (11-39 additional conditions each) and each taxon was associated with a distinct core growth profile (**Fig. 2A, Table S3**), consistent with previous findings [23, 24]. Four taxa were represented by sufficient genomes for comparisons (n>200 each); *K. pneumoniae* (hereafter *Kp,* n=6,652), *Klebsiella quasipneumoniae* subsp. *quasipneumoniae* (*Kqq*, n=201), *K. quasipneumoniae* subsp. *similipneumoniae* (*Kqs*, n=285) and *Klebsiella variicola* subsp. *variicola* (*Kvv,* n=672). Among these taxa, a total of 52 distinct growth conditions (corresponding to 18 distinct substrates) were core to at least one but variable and/or absent (not predicted to support growth of any isolates) from at least one other taxon. This included 31 growth conditions that were uniquely core to a single taxon (range 0-16).

The predicted growth capabilities were generally consistent with biochemical tests described in formal species definitions [25-28]. There were several cases that we defined as minor discrepancies where biochemical testing indicated complete absence or conservation of a capability within the KpSC or a specific taxon, and our predictions indicated <5% or >95%

conservation, respectively (**Table 1**). These differences likely reflect rare strain variations that have been captured by our genome collection but were not present among the biochemical test data. Seventeen larger-scale discrepancies were also identified; two may reflect inaccuracies in the model predictions for L-carnitine usage, which we previously estimated at ≤30% accuracy [22]. Twelve discrepancies were associated with substrates for which we have previously confirmed high predictive accuracy from metabolic models (≥94.6%, **Table 1**). It is most likely that these discrepancies were driven by under-sampling in the biochemical testing set, resulting in inaccurate conservation estimates and indicating that the formal species definitions should be revisited. The final three discrepancies were associated with D-lactic acid methyl ester for which we currently have no prediction accuracy estimate, but for which the biochemically-derived and predicted conservation levels were consistent for taxa represented by ≥200 genomes each.

Table 1: Biochemical tests described in formal species definitions and consistency with model predictions

Biochemical test	Taxa (phylogroup)							Growth prediction accuracy (%) [22]
	Kp (Kp1)	*Kqq* (Kp2)	*Kvv* (Kp3)	*Kqs* (Kp4)	*Kvt* (Kp5)	*Kqv* (Kp6)	*Ka* (Kp7)	
Substrates for which growth was thought to be conserved among KpSC [25]								
D-Glucose	(100)	(100)	(100)	(100)	(100)	(100)	(100)	100
L-Arabinose	(100)	(100)	(100)	(100)	(100)	(100)	(100)	100
D-Mannitol	(100)	(100)	(100)	(100)	(100)	(100)	(100)	92
D-Mannose	(100)	(100)	(100)	(98)	(100)	(100)	(100)	100
Melibiose	(100)	(100)	(100)	(98)	**(60)**	(100)	(100)	100
Raffinose	(100)	(100)	(100)	(100)	(100)	(100)	(100)	100
L-Rhamnose	(100)	(100)	(100)	(100)	(100)	(100)	(100)	97
Salicin	(100)	(100)	(100)	(100)	(100)	(100)	(100)	100
D-Sorbitol	(100)	(100)	(100)	(98)	(100)	(100)	(100)	100
Sucrose	(99)	(100)	(100)	(99)	(100)	(100)	(100)	100
Trehalose	(100)	(100)	(100)	(100)	(100)	(100)	(100)	100
D-Xylose	(100)	(100)	(100)	(100)	(100)	(100)	(100)	78
D-Arabitol	(99)	(100)	(99)	(100)	(100)	(100)	**(0)**	97
Cellobiose	(100)	(100)	(100)	(100)	(100)	(100)	(100)	100
Lactose	(100)	(100)	(100)	(100)	(100)	(100)	(100)	100
Maltose	(97)	(100)	(100)	(98)	(100)	(100)	**(0)**	95
Glycerol	(100)	(100)	(100)	(100)	(100)	(100)	(100)	97
Urea (nitrogen)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	Not tested
L-Tartrate	**(67)**	**(74)**	(100)	**(30)**	(100)	(100)	(100)	95
myo-Inositol	(100)	(100)	(100)	(98)	(100)	(100)	(100)	100
Substrates that differentiate KpSC taxa [28]								
Dulcitol	~ (48)	- (1)	~ (58)	~ (34)	~ (80)	+ (100)	+ (100)	100.0
Adonitol	+ (99)	~ (63)	- (1)	+ (99)	- (0)	- (0)	- (0)	100.0
Tricarballylic acid	- (0)	+ (100)	~ (100)	+ (97)	+ (60)	~ (93)	+ (100)	94.6
N-acetyl-neuraminic acid	- (0)	- (0)	- (0)	+ (97)	- (0)	- (0)	- (0)	100.0

D-arabitol	+ (99)	+ (100)	+ (99)	+ (100)	+ (100)	+ (100)	- (0)	100.0
L-sorbose	~ (53)	- (33)	+ (100)	- (0)	+ (100)	+ (100)	+ (100)	97.3
D-tagatose	~ (48)	- (1)	~ (58)	~ (34)	~ (80)	+ (100)	+ (100)	100.0
5-keto-D-gluconic acid	- (0)	~ (41)	+ (100)	- (3)	+ (100)	+ (100)	+ (100)	100.0
D-lactic acid methyl ester	+ (100)	+ (100)	+ (100)	+ (99)	- (100)	- (100)	- (100)	Not tested
4-hydroxyl-L-proline	- (11)	+ (75)	+ (100)	+ (98)	+ (100)	+ (100)	+ (29)	100.0
L-carnitine	~ (0)	- (0)	+ (0)	+ (0)	+ (0)	- (0)	+ (0)	29.7
Sample sizes								
N isolates (biochemical tests) [28]	10	5	5	4	6	5	1	N/A
N genomes (growth predictions) (This study)	6652	201	672	285	5	13	7	N/A

Substrate usage conservation as determined by biochemical testing reported in [25, 28] and predicted in this study (parentheses) is shown for each KpSC taxon: Kp = *K. pneumoniae,* Kqq = K. *quasipneumoniae* subsp. *quasipneumoniae*, Kvv = *K. variicola* subsp. *variicola,* Kqs = K. *quasipneumoniae* subsp. *similipneumoniae*, Kvt = *K. variicola* subsp. *tropica*, Kqv = *K. quasivariicola,* Ka = *K. africana* (labels in parentheses are phylogroup designations reported in [25]). '+' indicates 100% conservation in biochemical testing, '-' indicates 0% conservation in biochemical testing, '~' indicates >0% and <100% conservation in biochemical testing (absolute values not reported). Values in parentheses indicate the percentage of isolates predicted to be able to utilise the substrate in this work. Formatting indicates the level of concordance between the conservation values derived from biochemical tests and those derived from growth predictions; none = conservation levels are consistent; underline = minor discrepancies i.e. predicted conservation level <5% different from biochemical conservation level (applied only where biochemical testing indicated 0% or 100% conservation); bold and grey shading = conservation levels are not consistent. Substrate level growth prediction accuracies are also indicated, as reported in [22].

Predicted growth phenotype variability is driven by variation in gene and reaction content in the metabolic models; however, these represent only the subset of true metabolic genes for which the relevant reaction stoichiometries and supporting literature evidence is available for inclusion in the reference model [22]. Hence, we also explored the distribution of the broader set of metabolic orthologs. This analysis replicated the trends described above, showing that each taxon was associated with a distinct core metabolic ortholog profile (**Fig. S2, Table S2**). While all taxa appeared to dedicate similar proportions of their genomes to metabolic orthologs (51.3-54%, **Fig. S3**), there were statistically significant differences in the proportions assigned to key macromolecular functions as determined by analysis of Clusters of Orthologous Gene (COG) categories (**Fig. 2B**, **Supplemental Results**).

<FIG. 2>

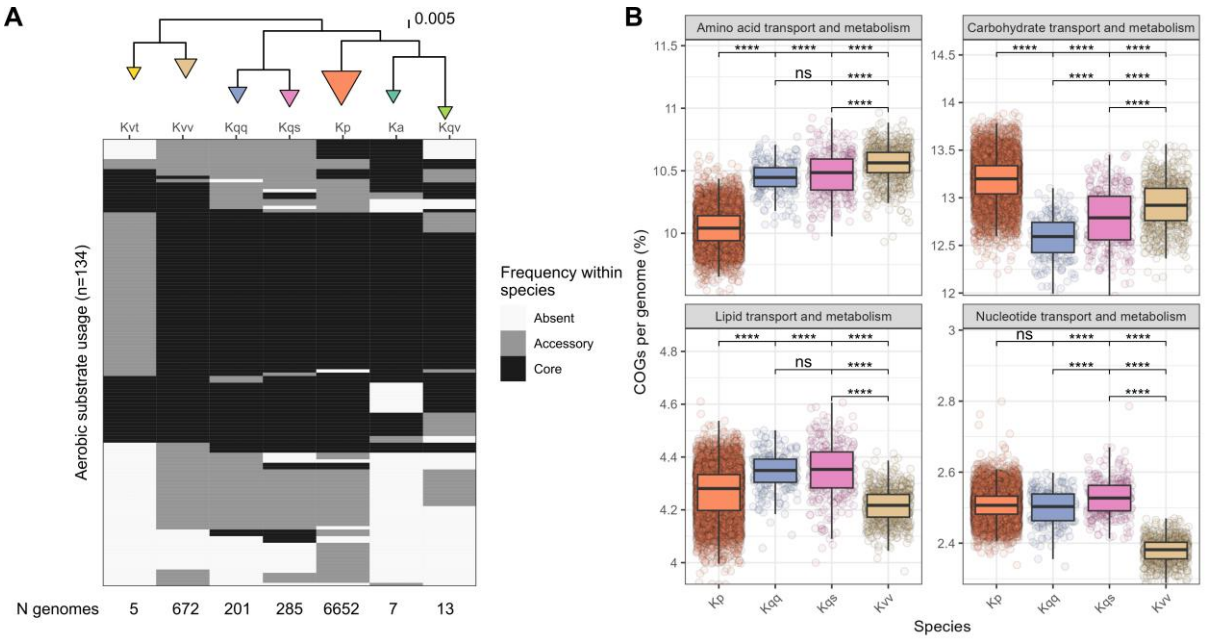


Fig. 2: Species-specific metabolism among the *Kp*SC

A: Heatmap showing model-predicted species-specific substrate usage. Only substrates for which usage frequencies differed in aerobic conditions are shown. Columns are arranged by species-phylogeny and the number of genomes per species is shown along the bottom of the x-axis and arrowhead size. Species names shortened for brevity: Ka = *K. africana*. Kp = *K. pneumoniae*. Kqq = K. *quasipneumoniae* subsp. *quasipneumoniae*. Kqs = K. *quasipneumoniae* subsp. *similipneumoniae*. Kqv = *K. quasivariicola*. Kvt = *K. variicola* subsp. *tropica*. Kvv = *K. variicola* subsp. *variicola*. **B:** Box and whisker array showing the proportion of each genome assigned to Clusters of Orthologous Gene categories (COG) by species coloured as per the phylogeny in panel A. Dots show individual genomes. Significance indicated such that the first row shows all comparisons to Kp, the second row shows comparisons to Kqq and third row shows comparison to Kvv. Significance calculated using a non-parametric Kruskal–Wallis test with Holm correction, followed by Dunn's *post-hoc* test (****: $p < 0.0001$, ns: not significant). Only species with n>200 genomes were analysed.

Metabolic variation is structured within species

While individual *Kp*SC taxa are clearly defined by distinct metabolic profiles, there was also clear evidence of variation within taxa (**Fig. 2A, Fig. S2**). In *Kp,* 938 metabolic orthologs were variably present in <95% genomes (544, 586 and 738 orthologs variable among *Kqq, Kqs* and *Kvv*, respectively) and 79 conditions were predicted to support variable growth (81, 78 and 67 for *Kqq, Kqs* and *Kvv*). We hypothesised that this variation is not randomly distributed in the population, and indeed our analysis of all-vs-all genome pairs within taxa showed statistically significant positive relationships between pairwise Average Nucleotide Identity (ANI, generally indicative of ancestral relatedness) and pairwise Jaccard similarity across metabolic genes (general linear hypothesis test, $p<2x10^{-16}$ for each of the four well sampled taxa, **Fig. 3A**). Comparison of the regression gradients indicated significant differences between taxa; however, the R^2 values were low (range 0.04-0.48), indicating a large amount of unexplained variation.

Next we used the recently established core-genome Life Identifier Number (LIN) code scheme to assign genomes to sub-lineages (SL) [29]. These sub-lineages represent deep-branching phylogenetic clades, separated by hundreds of years of evolutionary divergence (with notable exceptions discussed below). Of the 7,835 genomes, 7,833 were assigned to 1,418 distinct sub-lineages and two were unassigned due to insufficient alleles detected.

Pairwise Jaccard similarities for metabolic ortholog profiles were significantly higher within sub-lineages (median 0.956, IQR 0.947-0.965) than between sub-lineages within the same species (median 0.923, IQR 0.914-0.932), $p < 0.0001$ (Kruskal–Wallis test). Mirroring this, pairwise Jaccard similarities for predicted substrate usage profiles were significantly higher for pairs within sub-lineages (median 0.990, IQR 0.983-0.997) than between sub-lineages (median 0.97, IQR 0.96-0.98) for both aerobic and anaerobic substrate usage ($p < 0.0001$ for all comparisons, Kruskal–Wallis test with Holm correction, followed by Dunn's *post-hoc*, **Fig. 3B**).

Next we sought to explore structuring among accessory metabolic traits, using a co-occurrence analysis (see **Methods**); however, just 12 pairs of co-occurring substrate usages were identified. Of these, 11 pairs were mechanistically-linked e.g. the same substrate tested as different element sources (n=5 pairs) or substrates for which the catabolic pathways are overlapping (n=4 pairs, see **Supplemental Results**). Just one co-occurring phenotype pair (formamide and Fe(III)dicitrate as sources of carbon) was not obviously mechanistically linked, indicating that they may be co-selected or encoded by genes that are physically linked in the genome.

<FIG. 3>

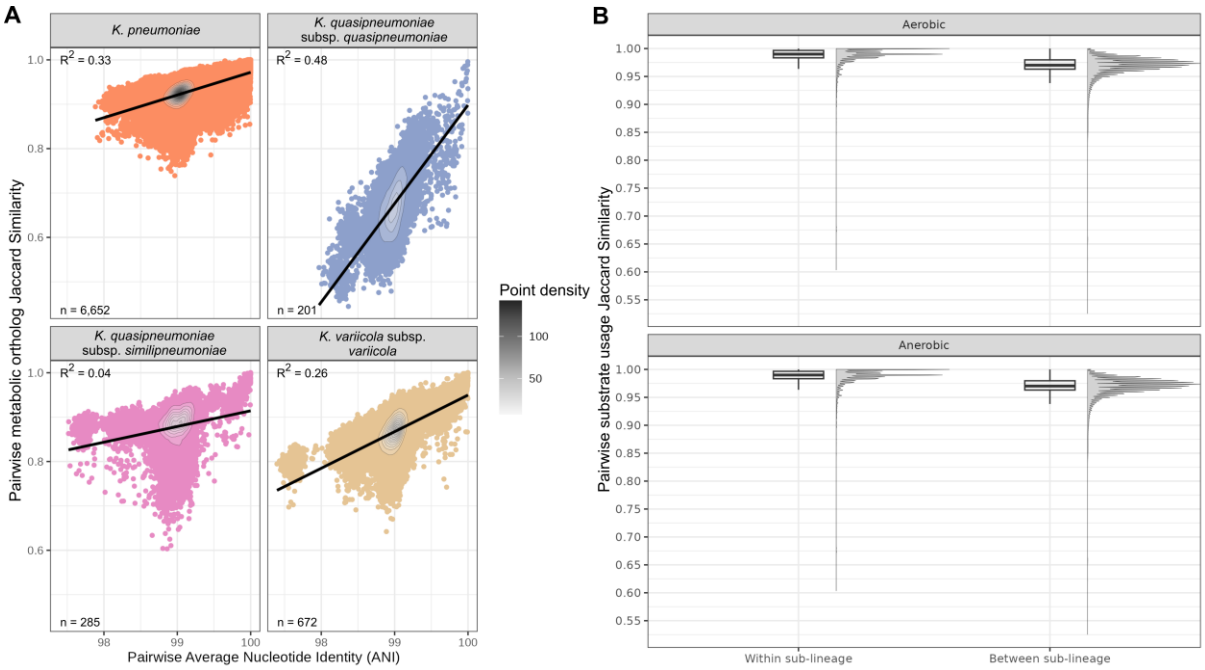


Fig. 3: Metabolic heterogeneity within species is non-randomly distributed

A: Pairwise Average Nucleotide Identity (ANI) vs pairwise Jaccard similarity of metabolic orthologs within each of the well sampled species (n>200 genomes). **B:** Distributions of pairwise (all vs all) Jaccard Similarity of simulated substrate usage for pairs of genomes within vs between sub-lineages in the same species. Significance asterisks excluded as each group was significantly different from every other group ($p < 2x10^{-16}$), calculated using a non-parametric Kruskal–Wallis test with Holm correction, followed by Dunn's *post-hoc* test.

Major sub-lineages are defined by unique metabolic profiles

To explore the association between metabolism and population structure at higher-resolution, we calculated the relative prevalence of predicted growth phenotypes within 48 *Kp* sub-lineages that were each represented by genomes from broad geographies and time-spans (≥15 genomes, 3-12 geographic regions, 5-86 years). These included sub-lineages corresponding to globally distributed multi-drug resistant (SL14, SL15, SL17, SL29, SL37,

SL101, SL147, SL258, SL307) and hypervirulent clones (SL23, SL25, SL65, SL86) [1]. The data clearly showed that each sub-lineage is associated with a distinct, core metabolic fingerprint, indicating metabolic specialisation (**Fig. 4**, 253-316 core aerobic and 198-223 core anaerobic conditions each, **Table S3**). Note that in some cases the sub-lineage specific core was smaller than the *Kp* species core reported above using the ≥95% genome threshold (363 aerobic, 307 anaerobic conditions), because most sub-lineages represent far less than 5% of the total *Kp* genome collection. This contradiction highlights the influence of sample size and diversity on core definitions, which should be considered as indicative of the general trends in the population. No pair of sub-lineages shared the same core growth profile, even pairs that are thought to share relatively recent common ancestry such as SL14 and SL15 (common ancestor approximately 45 years ago [2]) and SL65 and SL25 (common ancestor yet to be dated) [30] (**Fig. 4**, **Fig. S4**).

However, there was also clear evidence of diversity within sub-lineages, with 3-102 variable growth phenotypes each, a pattern that was mirrored by metabolic ortholog distributions (**Fig. S2, Supplemental Results**).

After excluding growth conditions predicted to support core growth among ≥95% (46/48) *K. pneumoniae* sub-lineages (n=615 conditions, 340 aerobic and 275 anerobic, corresponding to 202 total substrates) and those that did not support any growth among the 48 sub-lineages (n=525 conditions), the data support a broad classification of phenotypes as follows: i) 89 that were core in some sub-lineages (range=1-45 sub-lineages) but variably present in others (range=1-17) which we call 'sub-lineage-specific core traits'; ii) 13 that were not core in any sub-lineage (accessory in 3-46 sub-lineages, including the only pair of co-occurring phenotypes that was not mechanistically linked) (**Fig. 4**, **Fig. S4**). The sub-lineage-specific core traits could be further divided into those that were core to the majority (≥66%) of sub-lineages (n=57 phenotypes that we call 'majority sub-lineage-specific core traits'), those that were core to an intermediate number (≥25% to <66%) of sub-lineages (n=13 phenotypes, corresponding to six substrates, that we call 'common sub-lineage-specific core traits') and those that were core in only very few (<25%) sub-lineages (n=13 'rare sub-lineage-specific core traits', eight aerobic and five anerobic, corresponding to seven total substrates). We propose that the majority of sub-lineage specific core traits are unlikely to play a major role in defining sub-lineage specific evolutionary trajectories or ecological interactions, and hence we focussed our follow-up investigations on the common and rare sub-lineage specific core traits.

We have previously confirmed predictive accuracies ≥97% for five of the six common sub-lineage-specific core substrates; L-sorbose, L-tartrate, galactitol, D-tagatose and L-galactonate as sole sources of carbon in aerobic conditions [22] (tested using the Omnilog phenotyping microarray). The sixth, fructoselysine, is not present in the Omnilog array and could not be tested here due to its high cost. However, we tested seven of the rare sub-lineage-specific core growth substrates under aerobic conditions, confirming high accuracy for L-hydroxyproline as a carbon source (84.6%) and variable accuracies for the remainder (**Table S4, Supplemental Results**), which was not unexpected as rare substrate usage is comparatively understudied.

<FIG. 4>

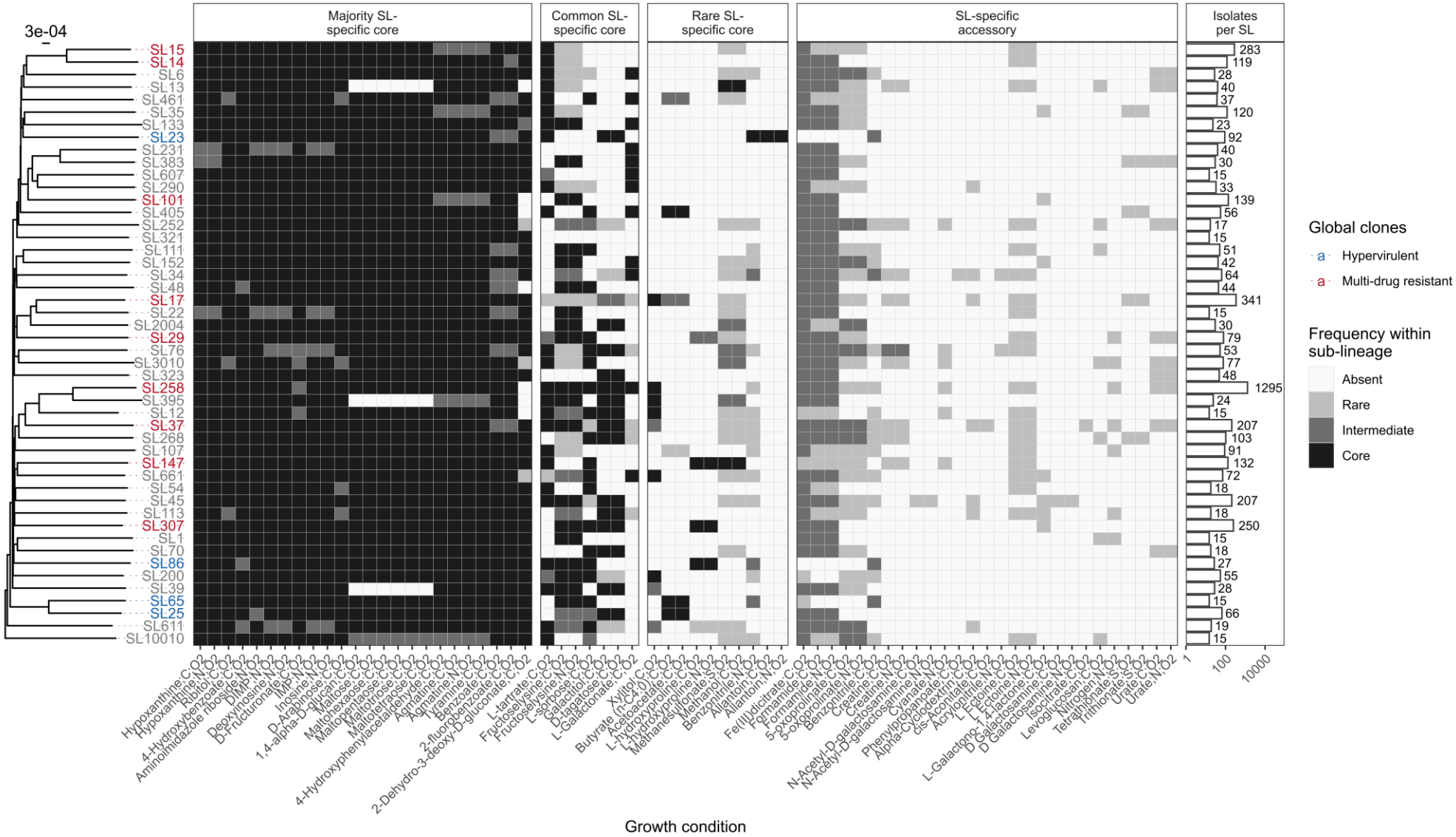

3e-04
Majority SL-specific core
Common SL-specific core
Rare SL-specific core
SL-specific accessory
Isolates per SL
SL15
SL14
SL6
SL13
SL461
SL35
SL133
SL23
SL231
SL383
SL607
SL290
SL101
SL405
SL252
SL321
SL111
SL152
SL34
SL48
SL17
SL22
SL2004
SL29
SL76
SL3010
SL323
SL258
SL395
SL12
SL37
SL268
SL107
SL147
SL661
SL54
SL45
SL113
SL307
SL1
SL70
SL86
SL200
SL39
SL65
SL25
SL611
SL10010
283
119
28
40
37
120
23
92
40
30
15
33
139
56
17
15
51
42
64
44
341
15
30
79
53
77
48
1295
24
15
207
103
91
132
72
18
207
18
250
15
18
27
55
28
15
66
19
15
Hypoxanthine:C:O2
Hypoxanthine:N:O2
Ribitol:C:O2
4-Hydroxybenzoate:C:O2
Aminoimidazole riboside:N:O2
DMP:N:O2
Deoxyinosine:N:O2
D-Fructuronate:C:O2
IMP:N:O2
Inosine:N:O2
D-Arabinose:C:O2
1,4-alpha-D-glucan:C:O2
Maltose:C:O2
Maltohexaose:C:O2
Maltopentaose:C:O2
Maltotriose:C:O2
Maltotetraose:C:O2
4-Hydroxyphenylacetaldehyde:C:O2
Agmatine:C:O2
Agmatine:N:O2
Tyramine:C:O2
Benzoate:C:O2
2-fluorobenzoate:C:O2
2-Dehydro-3-deoxy-D-gluconate:C:O2
L-tartrate:C:O2
Fructoselysine:C:O2
Fructoselysine:N:O2
L-sorbose:C:O2
Galactitol:C:O2
D-tagatose:C:O2
L-Galactonate:C:O2
Xylitol:C:O2
Butyrate (n-C4:0):C:O2
Acetoacetate:C:O2
L-hydroxyproline:C:O2
L-hydroxyproline:N:O2
Methanesulfonate:S:O2
Methanol:C:O2
Benzonitrile:N:O2
Allantoin:C:O2
Allantoin:N:O2
Fe(II)dicitrate:C:O2
Formamide:C:O2
Formamide:N:O2
5-oxoprolinate:C:O2
5-oxoprolinate:N:O2
Benzonitrile:C:O2
Creatine:C:O2
Creatine:N:O2
N-Acetyl-D-galactosamine:C:O2
N-Acetyl-D-galactosamine:N:O2
Cyanate:N:O2
Phenylpropanoate:C:O2
Alpha-Cyclodextrin:C:O2
cis-Aconitate:C:O2
Acrylonitrile:N:O2
L-Ectoine:N:O2
L-Galactono-1,4-lactone:C:O2
D-Galactosamine:C:O2
D-Galactosamine:N:O2
Isocitrate:C:O2
Levoglucosan:C:O2
Nitrogen:N:O2
Tetrathionate:S:O2
Trithionate:S:O2
Urate:C:O2
Urate:N:O2
1
100
10000
Growth condition
Global clones
a Hypervirulent
a Multi-drug resistant
Frequency within sub-lineage
Absent
Rare
Intermediate
Core

Fig. 4: Distinct metabolic fingerprints of sub-lineages within species

Heatmap showing frequency of variable aerobic substrate usage across 48 global *K. pneumoniae* sub-lineages (SL). Rows are ordered by phylogeny. Sub-lineage labels are coloured to indicate the globally-distributed clones described in [1]: Blue shows hypervirulent while red shows multidrug resistant. Substrate shown along X-axis. The element source of each substrate is semi-colon separated and abbreviated for brevity: C = Carbon, N = Nitrogen and S = Sulfur. O2 indicates aerobic conditions. Frequency of substrate usage indicated by shading as shown in legend. Number of isolates per each sub-lineage shown in bars. Anaerobic growth predictions shown in **Fig. S4**

Sub-lineages co-operate through cross-feeding metabolites

We have shown substrate usage varies across the *Kp*SC and some of this variation is sub-lineage-dependant. We hypothesised that such diversity promotes the maintenance of distinct sub-lineages by limiting inter-lineage competition and/or promoting commensalism. We tested this hypothesis using *in vitro* co-culture experiments.

We selected four carbon substrates (galactitol, L-tartrate, L-sorbose, L-hydroxyproline) for use in three independent substrate combinations and corresponding sub-lineage pairs for which predicted substrate-usages were mutually exclusive sub-lineage-specific core traits. One isolate from each sub-lineage was selected for co-culture aerobically in each of two conditions (minimal media plus each of the sub-lineage-specific core substrates), such that each experiment contained one isolate that was able to utilise the substrate as a sole carbon source (the prototroph) and one that was not (the auxotroph). Growth of the auxotroph was measured using OD_{600} and compared to the same isolate growing in monoculture in the same media. Co-cultured isolates were physically separated by a Transwell membrane but had shared access to the growth medium (**Fig. 5A**). A unique aspect of this experiment was the use of independent isolates from the same sub-lineages as biological replicates rather than replicates of the same isolate. This allowed us to limit the impact of strain-specific growth biases while demonstrating conserved behaviour across a greater number of independent isolates. Sub-lineages representing clinically-relevant clones were selected (multi-drug resistant SL37 vs hypervirulent SL29, hypervirulent SL86 vs hypervirulent SL23 and multi-drug resistant SL17 vs SL29).

Our data showed that co-culturing auxotrophs with prototrophs in the presence of a sub-lineage-specific core substrate significantly increased the OD_{600} of the auxotrophs (**Fig. 5B**) with mean $\log_2$-fold change 3.78 ± 2.64 SD at 24 hours, $p = 0.0128$; 4.02 ± 2.34 SD at 36 hours, $p = 0.0001$; 4.94 ± 1.14 SD at 48 hours, $p = 0.0039$; and 4.35 ± 1.27 at 72 hours $p = 0.0039$; Kruskal-Wallis + Dunn's *post-hoc* with Holm correction (**Fig. 5C**). Corresponding *in silico* co-culture simulations indicated that prototrophs were likely cross-feeding the auxotrophs by catabolising the sole carbon substrates, then exporting metabolites that were capable of supporting auxotroph growth (see **Supplemental Results**).

To confirm this phenomenon only occurred in the presence of a cross-feeding partner and eliminate the possibility of contamination or horizontal transfer of the genes conferring the prototroph phenotype, aliquots from each co-culture well were plated onto LB agar to confirm a single, distinct colony morphology. Colonies were then inoculated into M9 minimal media plus the auxotrophic substrate as a sole carbon source, confirming the original growth behaviour of each isolate (**Table S5**). Together these results indicated that cross-feeding was occurring from the prototrophs to the auxotrophs across: i) multiple genetic backgrounds (sub-lineages); ii) to and from each strain partner within each pair under different conditions; iii) across independent isolates within sub-lineages, confirming that commensal interactions can occur between co-habiting *Kp* sub-lineages.

<FIG. 5>

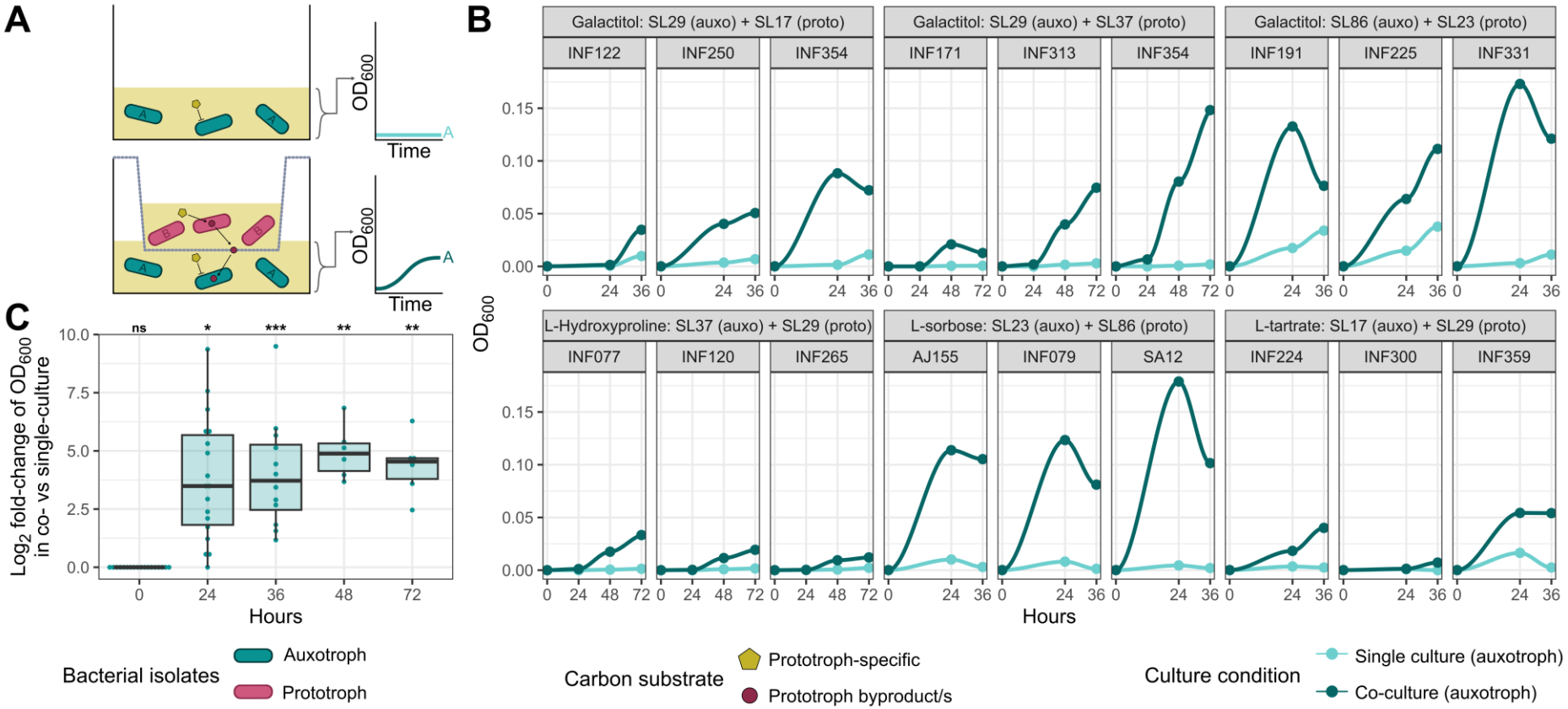


Fig. 5: Nutrient prototrophic lineages support growth of auxotrophic lineages

Co-culture growth assays between prototrophs and auxotrophs under various substrate usage conditions. Time points captured were at 0, 24, 36, 48 and 72 hours depending upon the substrate pair (**Methods**). **A:** Graphic demonstrating the experimental design measuring commensalism between an auxotrophic strain unable to utilise a carbon source and a prototrophic strain which can. The growth of the auxotroph was compared in single vs co-culture conditions. **B:** Growth results (OD_{600}) over time for each auxotrophic strain in the bottom well (n=3 strains per sub-lineage as biological triplicates) in single and co-culture with a prototroph for the substrate. **C:** Log_2-fold change distribution of growth results (OD_{600}) at each timepoint under different culture conditions. Statistical significance comparing the OD_{600} values of co-culture to single culture is shown above plot (Dunn's *post-hoc* with Holm correction, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Raw data can be found in **Table S5**.

Discussion

Here we present a bacterial population metabolism study of the *K. pneumoniae* species complex, in which we used quantitative metabolic modelling to predict and compare growth phenotypes for >7,000 members of a single species complex, and contextualised the results against an established population genomics framework [1]. This approach allowed us to identify key metabolic differences that are indicative of metabolic fingerprints at both the species and sub-species (sub-lineage) levels. Our subsequent *in vitro* co-culture experiments using multiple independent isolates from multiple sub-lineages and multiple substrate combinations (**Fig. 5**) provide evidence that these metabolic differences can facilitate commensal interactions between sub-lineages.

Our growth predictions implicated 643 distinct conditions, corresponding to 207 unique growth substrates, were able to support growth of all or almost all members of the species complex (≥95% isolates). A further 74 conditions corresponding to 33 substrates were predicted to support growth of a subset of 0 - 15% of *Kp*SC, while only 114 conditions supported growth at intermediate levels. This bimodal distribution mirrored the distribution of metabolic KEGG orthologs and is consistent with the general distribution of accessory genes in the *Kp*SC [2], and other bacterial species with similarly large effective population sizes [31, 32].

Beyond the high-level distribution, our data shows that growth capabilities are structured in the population: while there was clearly variation within and between sub-lineages, each of the 48 sub-lineages we examined was associated with a distinct set of core traits ('sub-lineage-specific core,' metabolic orthologs and growth phenotypes conserved in ≥95% of all members of the sub-lineage, **Fig. 4**, **Fig S4**), that most likely represent those of the sub-lineage common ancestor propagated via vertical descent. This finding is consistent with a model of negative frequency-dependant selection as a driver for the maintenance of diverse sub-lineages (entities) in the population [33, 34]. Where two sub-lineages compete for limited resources, e.g. food, the ability to consume a substrate that cannot be consumed by a competitor provides a selective advantage resulting in a relative frequency increase. This in turn results in a depletion of the food resource and shift in selective pressure to favour the competitor. Although comparatively understudied, commensal interactions such as the metabolite cross-feeding implicated by our co-culture experiments (**Fig. 5**), can drive similar frequency-dependent selective shifts. Together, these fluctuations prevent the exclusion of sub-lineages and drive the population to a stable equilibrium of co-existence.

Niche range no doubt further complicates the eco-evolutionary dynamics within *Kp* populations e.g. where sub-lineages are exposed to fluctuating selective pressures due to transient migration between niches, and/or where sub-lineages undergo niche adaptation. Recent analyses of contemporaneous *Kp* collections have shown that sub-lineages are not equally distributed between different hosts and environmental sources, which may indicate some level of niche-adaptation [4, 5]. However, the same studies also highlighted significant diversity within niches and near-identical strains from multiple sources, which is indicative of niche migration. Unfortunately, we were unable to directly test for associations between sub-lineage, metabolism and ecological niche with our convenience sample of genomes comprising primarily human-derived isolates (6,465/7,835 genomes). Future studies will be required to probe these complex multi-way interactions and to systematically test the model of negative-frequency dependent selection, which is notoriously difficult to detect [34].

While genome-scale metabolic models are powerful approaches for exploring cellular metabolism, they are not without caveats. Firstly, model construction is reliant on curated

databases that are limited by current biochemical knowledge [20], and while our pan-metabolic reference model was built from a large diverse collection of strains [22], it cannot possibly capture all metabolism in the *Kp*SC population. Hence, we are likely underestimating the metabolic diversity within the highly variable *Kp*SC. Secondly, growth prediction accuracies can vary by substrate, and are generally lower for anaerobic conditions for which the metabolic processes are less well understood [22]. While many of the core substrates and common sub-lineage specific core traits have been confirmed with high accuracy in previous work, our phenotypic testing showed variable accuracy for rarer substrates. Of note, our assays showed that allantoin usage, a trait previously associated with hypervirulent infections [35] and predicted by our models as a sub-lineage specific core trait in the dominant hypervirulent SL23, may in fact be core to *Kp*SC (see **Supplemental Results**). These and other inaccurate predictions highlight gaps in our current knowledge that are noted for future investigations. Finally, metabolic models do not consider the impact of sequence variations within metabolic genes nor variations in gene regulation that will also play an important role in determining metabolic phenotypes and ecological adaptation.

Notwithstanding the caveats of *in silico* modelling, our analyses clearly challenge the assumption that *Kp* performs a defined and homogenous metabolic function. This has important implications for how we understand and model microbial communities containing this organism. Notably, the *K. pneumoniae* sub-lineage-specific core traits we identified include those implicated in driving competitive interactions and colonisation resistance in the mammalian gut e.g. fructoselysine and galactitol usage [36, 37], as well as those that are recognised animal metabolites (e.g. L-hydroxyproline, a component of collagen) [38]. These sub-lineage-specific differences likely drive variations in fitness and virulence under different conditions that must be considered when designing and interpreting *Kp* studies, particularly those seeking to design or evaluate novel control strategies targeting metabolic processes or harnessing microbial competitive interactions.

Online Methods

Genome acquisition, quality filtering and genotyping

Genomes from 29 medium- to large-scale *Kp*SC studies [2, 39-65] were selected to maximise diversity of *Kp*SC species and STs, and to reduce sampling biases (i.e. outbreak investigations were not included). All accessions and collated strain metadata can be found in **Table S1**. Reads were acquired from European Nucleotide Archive using the enaDataGet and enaGroupGet scripts from enaBrowserTools v1.6 [66].

Reads were trimmed with Trim Galore v0.5.0 using -q 20 option [67] then assembled with Unicycler v0.4.7 with the --keep 0 option [68]. Assemblies that failed the KlebNET-GSP (*Klebsiella* Genomic Surveillance Platform) inclusion criteria were excluded (i.e. genomes with >500 contigs, total length <4,969,898 or >6,132,846 bp). The Graphical Fragment Assembly (GFA) files were then analysed for the number of GFA dead ends per assembly using the custom getUnicyclerGraphStats.py from Unicycler [68]. Assemblies with >200 GFA dead ends were excluded from the study as per previous quality control recommendations for construction of genome-scale metabolic models [21].

Assemblies were annotated with Bakta v1.1.1 [69] with the '--gram –' option and Bakta database (date accessed: 1/09/2021) and Kleborate v2.0.4 [70] was used to identify species and Sequence Types (STs). Pathogenwatch (https://pathogen.watch/) and BIGSdb

(https://bigsdb.pasteur.fr/) [71, 72] were used to assign LIN codes [29] and sub-lineage numbers. Pairwise ANI was calculated using FastANI v1.33 [73] using --fragLen 3000.

Genome dereplication

While we deliberately excluded isolate genomes from studies focussed on outbreak investigations, there remain a subset of sub-lineages that were vastly overrepresented (e.g. dense sampling in specific geographies or the inclusion of transmission clusters within larger collections). To further subsample these overrepresented sub-lineages, we used a dereplication similar to that described previously [70]. First, popPUNK v2.4.0 [74] was used to finely cluster isolates. The 'create-db' function was used with the following options: '--sketch-size 1000000 --min-k 15 --max-k 29 --qc-filter prune'. Then the 'fit-model' function was used with the following options: 'dbscan --ranks 1,2,3,5 --graph-weights'. Finally, the 'fit-model' function was used again, with the options: 'refine --graph-weights --unconstrained'. Additionally, the 'poppunk_visualise' function was used, with the '--distances' and '--previous-clustering' utilising the refined model fit, to output a neighbour-joining core tree.

Isolates which shared the same combination of 'dbscan popPUNK cluster', 'source of isolation' and 'country of isolation' were sorted using newfangled_untangler.py v1.0.4 (https://github.com/bananabenana/newfangled_untangler/) then the highest quality assemblies were selected using Assembly-Dereplicator v0.1.0 [75] with the following options: '--threshold 0.0003'.

Pangenome analysis

The 7,835 filtered and dereplicated *Kp*SC genomes were split into species, then run separately using Panaroo v1.2.8 [19] with the following options: '--clean-mode sensitive -a core --aligner mafft --no_clean_edges --core_threshold 0.98 --merge_paralogs --remove-invalid-genes'. Due to computational limits, the 6,676 *K. pneumoniae* genomes were split into two groups of 3,338 by extracting the tips from the neighbour-joining tree from popPUNK, to ensure closely-related *K. pneumoniae* genomes were run together. These were run separately on Panaroo. The Panaroo output graphs for all species were then merged using the panaroo-merge command with the following options: '--merge_paralogs' to obtain the final matrix of gene presence and absence per genome, and representative gene sequences for each Panaroo cluster.

Quantitative COG analysis

Clusters of Orthologous Gene (COG) annotations [76, 77] were extracted from Bakta output files then tallied and expressed as a proportion of total COGs per genome (https://dx.doi.org/10.6084/m9.figshare.24503737). COGs were amalgamated into the 26 functional categories. Statistical differences between species were calculated using a Kruskal-Wallis ANOVA then a Dunn's post-hoc test with Holm correction via the stat_compare_means() function from ggpubr v0.5.0 [78], with the following options: ' test = "kruskal.test, p.adjust.method = "holm" '

Metabolic gene identification and analysis

To analyse metabolic pathways, the pan_genome_reference.fa file from Panaroo was translated using the transeq tool from EMBOSS v6.6.0 [79] with the following options: '-table 11 -frame 1'. Translated sequences were analysed using kofam_scan v1.3.0 (https://github.com/takaram/kofam_scan), the command line version of KEGG's [20] kofamKOALA [80]. KEGG release 94.0 ver. 2020-04-02.

Genes with an e-value of ≤0.001 were retained and analysed using KEGG Mapper (https://www.genome.jp/kegg/mapper/reconstruct.html) to identify pathway information. Using the BRITE tab, genes which fell under the Protein families: metabolism subheading were used to classify genes as metabolic or non-metabolic.

KEGG ortholog assignments were used to convert the Panaroo gene presence/absence matrix into a metabolic ortholog matrix. This allowed grouping of functional isozymes which would otherwise have an identical KEGG function. Pairwise metabolic ortholog Jaccard similarities were calculated using the vegdist function from R package vegan v2.5-7 [81].

Distribution of metabolism genes in the population

The Twilight protocol [31] was used to classify the distribution of metabolic orthologs among *Kp*SC taxa and 48 common *K. pneumoniae* sub-lineages (n≥15 genomes each, defined by LIN codes) using the -s 5 and -s 20 commands, respectively. Twilight distribution definitions were used in this study. 'Core' refers to ≥95% prevalence within a group. 'Intermediate' refers to <95% and >15% prevalence. 'Rare' refers to ≤15% and >0% prevalence. 'Absent' refers to 0% prevalence.

Metabolic model construction and analysis

Simulated metabolic models were generated via Bactabolize v1.0.1 [21] using the *draft_model* command with the following options: '--media_type m9 --atmosphere_type aerobic --min_coverage 25 --min_pident 80'. The *Kp*SC-pan v2.0 model [22] was used as the input reference. One-hundred and sixty of 7,835 models (2.0%) required gap-filling to enable simulation of growth on M9 minimal media with glucose (median: 1 reaction added, range: 1 – 23), using flux balance analysis to optimise the most recent version of the biomass objective function 'BIOMASS_Core_Feb2022'. (Note - we have previously shown that models gap-filled to simulate growth in minimal media plus glucose can accurately predict a range of other substrate growth phenotypes [21].)

Flux balance analysis was used to predict growth phenotypes for all possible carbon, nitrogen, phosphorus and sulfur sources supported by the reference model as sole sources of carbon, nitrogen, phosphorus or sulfur in M9 minimal media, in both aerobic and anaerobic conditions. These analyses were performed using Bactabolize's *fba* command with the following option: '--fba_spec_name m9'. All growth lower bounds and default substrate sources used in this analysis are available at https://github.com/kelwyres/Bactabolize. Models were considered capable of simulating growth when the optimised biomass value was ≥0.0001. Pairwise Jaccard similarities and population distribution of positive growth phenotypes were calculated as described above for metabolic orthologs.

To evaluate co-occurrence of metabolic traits and control for population structure, genomes were randomly subsampled to a maximum of 10 per sub-lineage (n=2,427 genomes), then predicted aerobic growth phenotypes analysed using Coinfinder v1.2.0 [82] with the --associate option.

In vitro growth experiments

To validate substrate usage predictions, 13 isolates (**Table S4**) were selected from our in-house collections (described in [2, 35, 45]) and tested in triplicate on each of nine substrates in minimal media; seven substrates as sole sources of carbon (methanol, L-hydroxyproline, xylitol, butyrate, acetoacetate, D-glycerol, D-glucose), nitrogen (allantoin) and sulfur (methanesulfonate) in aerobic conditions. Two different experiments were performed: i)

measuring OD_{600} at 24- and 48-hour endpoints; ii) 3-day substrate coaxing. Previous studies have demonstrated that strains will undergo adaptation to achieve *in silico* predicted growth [83]. The coaxing experiment was performed for substrates where one or more isolates that were predicted to grow were not positive for growth in the endpoint experiment, which was predicted to be a regulatory issue. To address this, isolates were coaxed onto the testing substrate via subculturing into progressively lower amounts of D-glycerol, which all isolates can use as a carbon source (10 mM on subculture 1, 5 mM on subculture 2, 0 mM on subculture 3). See **Supplemental Results** for more details.

For endpoint experiments, isolates were grown overnight in 2 mL M9 media + 10 mM D-glycerol. M9 media was prepared by using 1X M9 Minimal Salts (Sigma) and adding 2 mM $MgSO_4$ and 0.1 mM $CaCl_2$ after autoclaving. 1 mL of culture was pelleted at 8000 x g for 10 minutes, supernatant discarded, then centrifugally washed using 1 mL 0.9% NaCl to remove excess carbon/nitrogen/sulfur. Isolates were then diluted 1:100 using 0.9% NaCl and 5 µL sub-cultured into 96-well plates (CLS3603, Corning) containing 200 µL M9 + substrate. Substrates were tested at the following concentrations: 400 mM methanol, 20 mM L-hydroxyproline, 20 mM xylitol, 20 mM sodium butyrate, 20 mM acetoacetate, 20 mM D-glucose, 10 mM D-glycerol 20 mM methanesulfonate, 60 mM allantoin, all at pH 7.0 and 0.2 µm filter sterilised. In the case of allantoin as a nitrogen source, custom nitrogen-negative M9 was prepared (34 mM NaH_2PO_4, 64 mM K_2HPO_4, 1 µM $FeSO_4$, 2 mM $MgSO_4$ and 0.1 mM $CaCl_2$). In the case of methanesulfonate as a sulfur source, sulfur-negative M9 was prepared (2 mM $MgCl_2$ instead of $MgSO_4$). Isolate negative controls were performed for each media. All substrates and media were made fresh. D-glucose or D-glycerol positive controls were performed for each isolate. Plates were incubated at 37°C, shaking at 150 RPM and optical density read every 24 hours at 600 nm absorbance using an infinite 200Pro and i-control v2.0.10.0 (Tecan). Each isolate experiment was done in biological triplicate. Each well was blanked with a no isolate control, then the mean used to determine growth, with a cut-off of 0.05 due to limits of detection.

Coaxing experiments were performed as above, over three days. Isolates were added to media containing the tested substrate plus D-glycerol 10 mM on subculture day 1, 5 mM on subculture day 2, 0 mM on subculture day 3. Prior to sub-culturing, cells were centrifugally washed as above and 5 µL transferred to 200 µL total media. Only subcultures on day 3 were used for analysis (0 mM D-glycerol).

Co-culture experiments

We selected three pairs of sub-lineages and three corresponding pairs of substrates (four total substrates) for which one was predicted as a sub-lineage specific core trait in sub-lineage A and the other predicted as a sub-lineage specific core trait in sub-lineage B. For each sub-lineage, three isolates were selected from our in-house collections (described in [2, 24, 45]) and treated as biological replicates. To rule out strain-strain antagonism prior to co-culturing, a primary inoculum of *K. pneumoniae* isolates (**Table S5**) were streaked onto LB agar plates and grown for eight hours at 37°C. Then, primary inoculums of additional isolates were streaked perpendicularly. Plates were incubated for 16 hours at 37°C and no growth inhibition of the perpendicular streaks was observed.

Co-culture results were compared to single cultures using OD_{600} of pairs of isolates (**Table S5**) that were physically separated by 0.4 µm pores using 24-well Transwells (Sigma). All cultures were grown at 37°C, 100 RPM. See **Methods** for additional details. Substrates were used in 1X M9 media at the following concentrations: 20 mM for galactitol, L-sorbose, L-tartrate, and 5 mM for L-hydroxyproline.

M9 media plus the tested substrate was added to the Transwell (100 µL) and the main well (800 µL). Isolates were grown overnight in 1X M9 containing 10 mM D-glycerol, then centrifugally washed as above. Cultures were diluted 1:100 in 0.9% NaCl, and 5 µL added to wells. The prototrophic isolate was added to the Transwell while the auxotroph was added to the main well. Cultures were grown and 75 µL was taken from wells at timepoint for OD_{600} measurements. For isolate-substrate combinations where we observed slow growth rates, 24, 48 and 72-hour timepoints were used, for remaining substrates, timepoints at 0, 24 and 36 hours were used. OD_{600} values were blanked using no-isolate media controls. For log_2 fold-change analysis, values of 0 were transformed to $1x10^{-5}$, but kept raw for other statistical analyses. To confirm auxotroph growth in co-culture was not due to any passed-on genetic or other characteristic, but solely due to cross-feeding from the prototrophic isolate, co-cultured isolates were then plated onto LB agar, confirmed uniform, distinct colony morphology, then re-grown under the same nutrient conditions in single culture.

Co-culture simulations

Co-culture simulations were performed using MICOM v0.32.2 [84] with the relevant *K. pneumoniae* metabolic models. Community models were built using the build function, then communities were grown in growth media matching co-culture experiments (M9 minimal media with various carbon sources) with tradeoff set to 0.5. Metabolite and reaction flux was then analysed to determine metabolite sharing from prototrophs to auxotrophs. All MICOM code is available at Figshare (https://dx.doi.org/10.6084/m9.figshare.24503737).

Data visualisation

All code used to generate figures and perform statistical analysis can be found at (https://dx.doi.org/10.6084/m9.figshare.24503737).

Statistical analysis and graphical visualisation were performed using R v4.0.3 [85], RStudio v1.3.1093 [86], with the following software packages: tidyverse v1.3.1 [87], viridis v0.5.1 [88], RColorBrewer v1.1-2 [89], ggpubr v0.4.0 [78] ggpmisc v0.4.4 [90], aplot v0.1.6 [91], colorspace v2.0-2 [92], ggtree v2.4.1 [93], vegan v2.5-7 [81], rstatix v0.7.0 [94] and ggnewscale v0.4.5 [95].

Author Contributions

Conceptualization: BV, SB, JMM, KEH, KLW.

Methodology: BV, SB, JMM, KEH, KLW

Validation: BV.

Formal analysis: BV, HBC, KLW.

Investigation: BV, HBC.

Resources: KLW, SB, KEH.

Data Curation: BV, MR-P

Writing: BV, HBC, MR-P, SB, JMM, KEH, KLW.

Visualization: BV, KLW.

Supervision: SB, KLW.

Project administration: KLW.

Funding acquisition: SB, JMM, KEH, KLW.

Data availability

All genomic data used is publicly available - accessions and genomes used can be found in **Table S1**. All raw data can be found in **Supplemental Tables**. R and MICOM code can be found at Figshare (https://dx.doi.org/10.6084/m9.figshare.24503737). Interactive phylogeny can be found at https://microreact.org/project/pcTMwQZAGCsqghbEZzaj11-a-metabolic-atlas-of-the-klebsiella-pneumoniae-species-complex.

Funding

This work was funded by the Australian Research Council Discovery Project (DP200103364, awarded to KLW, KEH, SB and JMM). KLW is supported by a National Health and Medical Research Council of Australia Investigator Grant (APP1176192).

Competing interests

The authors declare no competing interests

Acknowledgements

This work was supported by the MASSIVE HPC facility (www.massive.org.au).

Supplemental material

Fig. S1: Substrate usage frequencies across *Kp*SC dataset. **A:** Number of substrates predicted to support growth for each isolate, stratified by substrate type and atmosphere. **B:** Population frequency of predicted substrate usage by substrate type. Core refers to substrate usage conserved in ≥95% of isolates. 'Int' refers to intermediately present (>15% to <95%). Rare refers to >0 to <15%, while 'Absent' indicates no predicted growth. Raw data can be found in **Table S3**.

Fig. S2: Species are associated with unique KEGG ortholog profiles. Heatmap showing species-specific KEGG orthologs. Only orthologs for which frequencies differed between taxa are shown. Columns are arranged by species-phylogeny as in **Fig. 2A** and names shortened for brevity: Ka = *K. africana*. Kp = *K. pneumoniae*. Kqq = K. *quasipneumoniae* subsp. *quasipneumoniae*. Kqs = K. *quasipneumoniae* subsp. *similipneumoniae*. Kqv = *K. quasivariicola*. Kvt = *K. variicola* subsp. *tropica*. Kvv = *K. variicola* subsp. *variicola*. Frequencies are indicated by colours as shown in the legend: Absent = not present in any genomes; Accessory = present in >0% and <95% genomes; Core = present in ≥95% genomes. Raw data can be found in **Table S2**.

Fig. S3: Metabolic genes outnumber non-metabolic genes within genomes. Ridgeline plot showing the ratio of metabolic:non-metabolic genes per individual genome. Distributions of each taxa are shown, along with number of genomes per taxa taxon-specific distributions of the (i.e. The ratio of metabolic to non-metabolic genes).

Fig. S4: Distinct anerobic metabolic fingerprints of sub-lineages within species. Heatmap showing frequency of variable anaerobic substrate usage across 48 global *K. pneumoniae* sub-lineages (SL). Rows are ordered by phylogeny. Sub-lineage labels are coloured to indicate the globally-distributed clones described in [1]: Blue shows hypervirulent while red shows multidrug resistant. Substrate shown along X-axis. The element source of each substrate is semi-colon separated and abbreviated for brevity: C = Carbon, N =

Nitrogen and S = Sulfur. AnO2 indicates anaerobic conditions. Frequency of substrate usage indicated by shading as shown in legend. Number of isolates per each sub-lineage shown in bars. Anaerobic version of **Fig. 4**.

Table S1: List of all 7,835 genomes used for this study along with accessions and metadata. Also contains results of simulated substrate usage analysis from metabolic models, proportional Clusters of Orthologous Genes (COG) data.

Table S2: Metabolic gene (metabolic KEGG Ortholog) presence/absence table for entire dataset 7,835 genome dataset.

Table S3: Binarised growth predictions from metabolic modelling for entire dataset 7,835 genome dataset.

Table S4: Results of *in vitro* growth experiments for metabolic model validation.

Table S5: Raw data and statistical analysis of co-culture experiments.

References

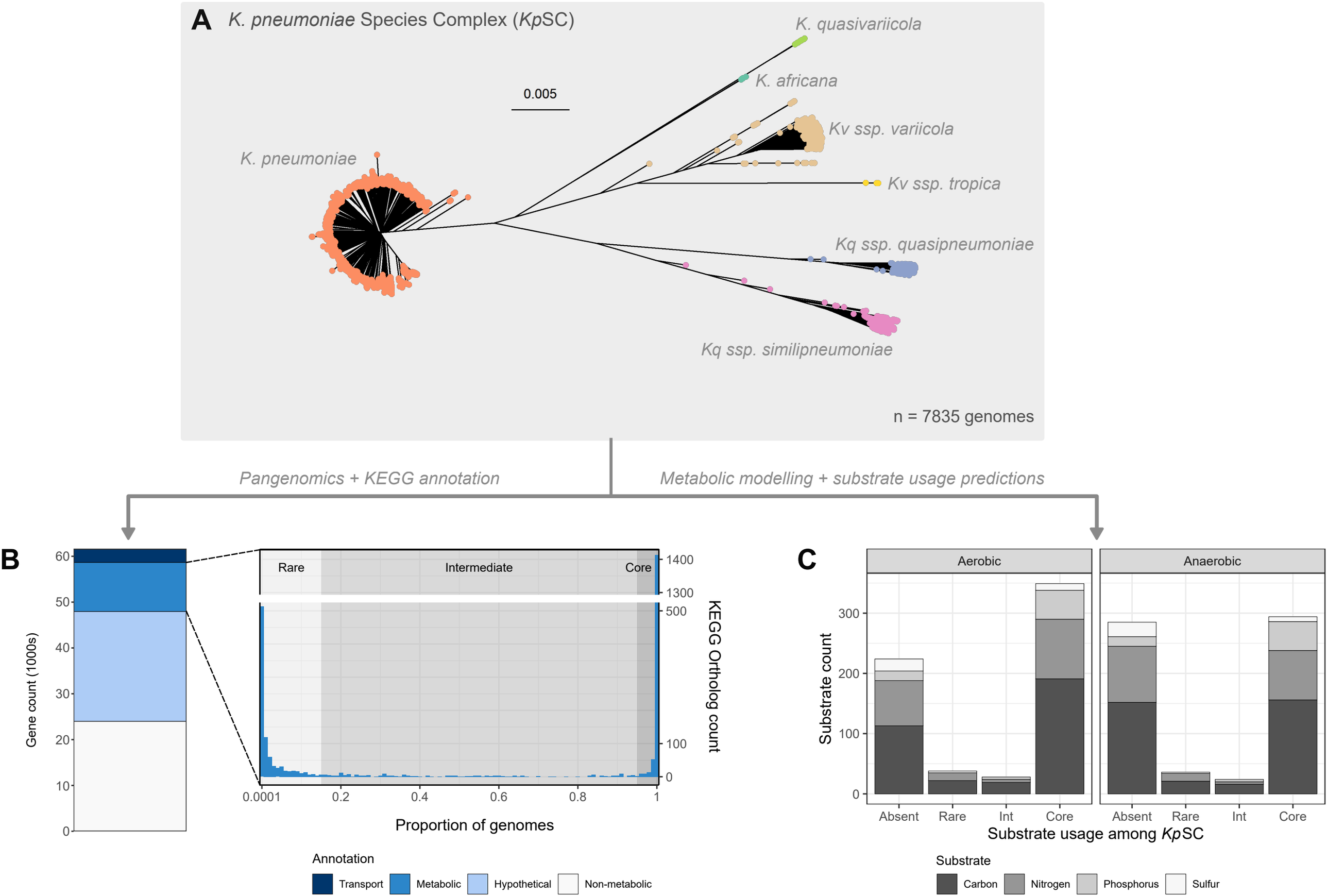
1. Wyres, K.L., M.M.C. Lam, and K.E. Holt, *Population genomics of Klebsiella pneumoniae.* Nature Reviews Microbiology, 2020. **18**(6): p. 344-359.
2. Holt, K.E., et al., *Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance Klebsiella pneumoniae, an urgent threat to public health.* Proceedings of the National Academy of Sciences, 2015. **112**(27): p. E3574.
3. Podschun, R. and U. Ullmann, *Klebsiella spp. as Nosocomial Pathogens: Epidemiology, Taxonomy, Typing Methods, and Pathogenicity Factors.* Clinical Microbiology Reviews, 1998. **11**(4): p. 589-603.
4. Thorpe, H.A., et al., *A large-scale genomic snapshot of Klebsiella spp. isolates in Northern Italy reveals limited transmission between clinical and non-clinical settings.* Nature Microbiology, 2022. **7**(12): p. 2054-2067.
5. Calland, J.K., et al., *Population structure and antimicrobial resistance among Klebsiella isolates sampled from human, animal, and environmental sources in Ghana: a cross-sectional genomic One Health study.* The Lancet Microbe, 2023. **4**(11): p. e943-e952.
6. Watkins, E.R., M.C.J. Maiden, and S. Gupta, *Metabolic competition as a driver of bacterial population structure.* Future Microbiology, 2016. **11**(10): p. 1339-1357.
7. Cui, Y., et al., *Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen Vibrio parahaemolyticus.* Molecular Biology and Evolution, 2015. **32**(6): p. 1396-1410.
8. Reuter, S., et al., *Directional gene flow and ecological separation in Yersinia enterocolitica.* Microb Genom, 2015. **1**(3): p. e000030.
9. Sheppard, S.K., et al., *Cryptic ecology among host generalist Campylobacter jejuni in domestic animals.* Mol Ecol, 2014. **23**(10): p. 2442-51.
10. Top, J., et al., *Functional characterization of a gene cluster responsible for inositol catabolism associated with hospital-adapted isolates of Enterococcus faecium.* Microbiology, 2021. **167**(9).
11. Connor, C.H., et al., *Multidrug-resistant E. coli encoding high genetic diversity in carbohydrate metabolism genes displace commensal E. coli from the intestinal tract.* PLOS Biology, 2023. **21**(10): p. e3002329.
12. Bordbar, A., et al., *Constraint-based models predict metabolic and associated cellular functions.* Nature Reviews Genetics, 2014. **15**(2): p. 107-120.
13. Omid, A., et al., *Pangenome reconstruction of Lactobacillaceae metabolism predicts species-specific metabolic traits.* bioRxiv, 2023: p. 2023.09.18.558222.

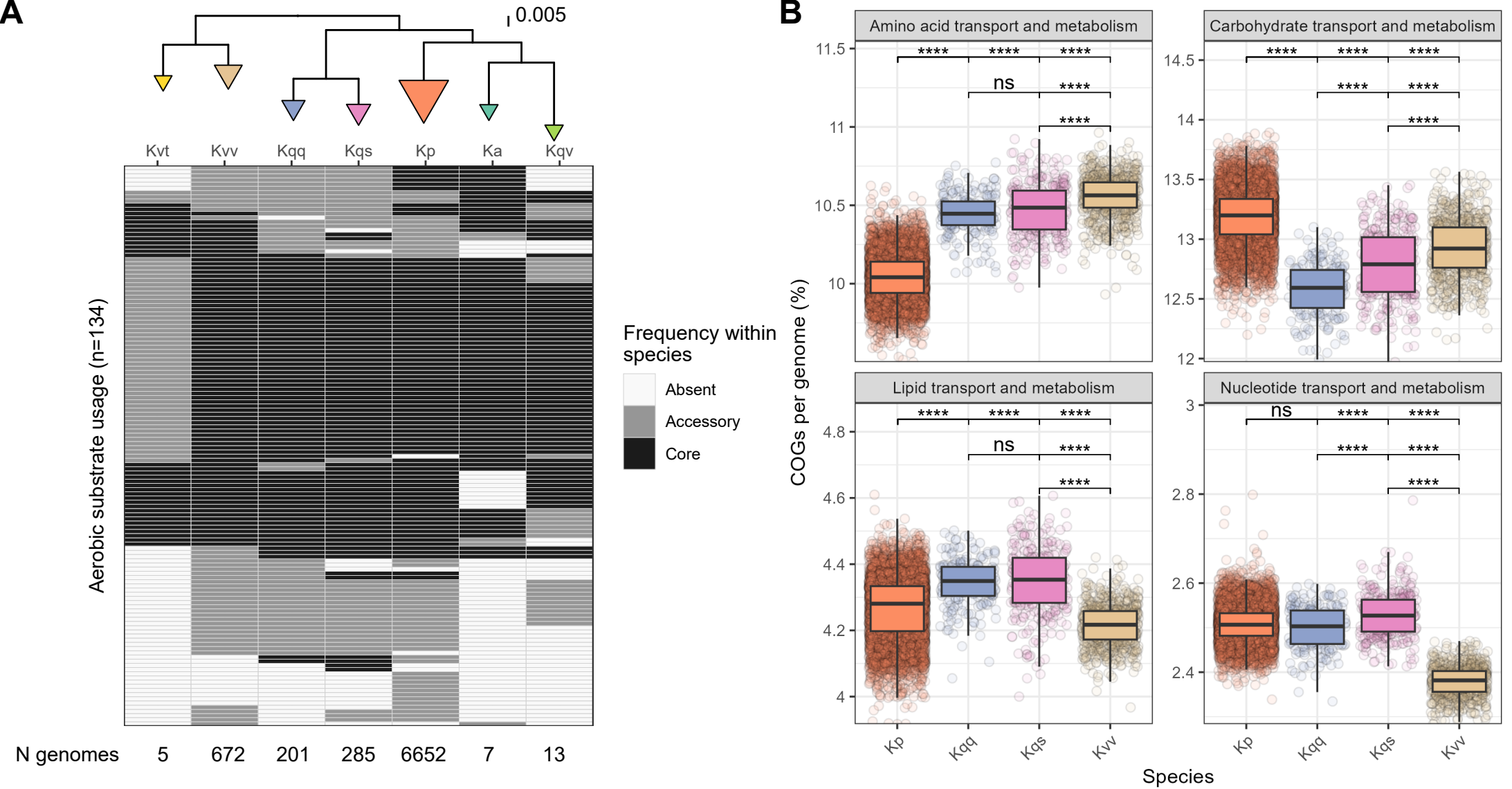
14. Seif, Y., et al., *Genome-scale metabolic reconstructions of multiple Salmonella strains reveal serovar-specific metabolic traits.* Nature Communications, 2018. **9**(1): p. 3771.
15. Blázquez, B., et al. *New Insights on Metabolic Features of Bacillus subtilis Based on Multistrain Genome-Scale Metabolic Modeling*. International Journal of Molecular Sciences, 2023. **24**, DOI: 10.3390/ijms24087091.
16. Norsigian, C.J., et al., *Comparative Genome-Scale Metabolic Modeling of Metallo-Beta-Lactamase–Producing Multidrug-Resistant Klebsiella pneumoniae Clinical Isolates.* Frontiers in Cellular and Infection Microbiology, 2019. **9**(161).
17. Monk, J.M., et al., *Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments.* Proceedings of the National Academy of Sciences, 2013. **110**(50): p. 20338-20343.
18. Bosi, E., et al., *Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity.* Proceedings of the National Academy of Sciences, 2016. **113**(26): p. E3801-E3809.
19. Tonkin-Hill, G., et al. *Producing polished prokaryotic pangenomes with the Panaroo pipeline*. Genome biology, 2020. **21**, 180 DOI: 10.1186/s13059-020-02090-4.
20. Kanehisa, M. and S. Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Research, 2000. **28**(1): p. 27-30.
21. Vezina, B., et al., *Bactabolize is a tool for high-throughput generation of bacterial strain-specific metabolic models.* eLife, 2023. **12**: p. RP87406.
22. Cooper, H.B., et al., *A validated pangenome-scale metabolic model for the Klebsiella pneumoniae species complex.* Microbial Genomics, 2024. **10**(2).
23. Hawkey, J., et al., *ESBL plasmids in Klebsiella pneumoniae: diversity, transmission and contribution to infection burden in the hospital setting.* Genome Medicine, 2022. **14**(1): p. 97.
24. Blin, C., et al., *Metabolic diversity of the emerging pathogenic lineages of Klebsiella pneumoniae.* Environmental Microbiology, 2017. **19**(5): p. 1881-1898.
25. Imhoff, J., *Bergey's Manual® of Systematic Bacteriology*. 2005. p. 587-850.
26. Rosenblueth, M., et al., *Klebsiella variicola, A Novel Species with Clinical and Plant-Associated Isolates.* Systematic and Applied Microbiology, 2004. **27**(1): p. 27-35.
27. Brisse, S., V. Passet, and P.A.D. Grimont, *Description of Klebsiella quasipneumoniae sp. nov., isolated from human infections, with two subspecies, Klebsiella quasipneumoniae subsp. quasipneumoniae subsp. nov. and Klebsiella quasipneumoniae subsp. similipneumoniae subsp. nov., and demonstration that Klebsiella singaporensis is a junior heterotypic synonym of Klebsiella variicola.* Int J Syst Evol Microbiol, 2014. **64**(Pt 9): p. 3146-3152.
28. Rodrigues, C., et al., *Description of Klebsiella africanensis sp. nov., Klebsiella variicola subsp. tropicalensis subsp. nov. and Klebsiella variicola subsp. variicola subsp. nov.* Research in Microbiology, 2019. **170**(3): p. 165-170.
29. Hennart, M., et al., *A Dual Barcoding Approach to Bacterial Strain Nomenclature: Genomic Taxonomy of Klebsiella pneumoniae Strains.* Molecular Biology and Evolution, 2022. **39**(7): p. msac135.
30. Wyres, K.L., et al., *Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of Klebsiella pneumoniae.* PLOS Genetics, 2019. **15**(4): p. e1008114.
31. Horesh, G., et al., *Different evolutionary trends form the twilight zone of the bacterial pan-genome.* Microbial Genomics, 2021. **7**(9).
32. McInerney, J.O., *Prokaryotic Pangenomes Act as Evolving Ecosystems.* Molecular Biology and Evolution, 2023. **40**(1): p. msac232.
33. Harrow, G.L., et al., *Negative frequency-dependent selection and asymmetrical transformation stabilise multi-strain bacterial population structures.* The ISME Journal, 2021. **15**(5): p. 1523-1538.
34. Christie, M.R. and G.G. McNickle, *Negative frequency dependent selection unites ecology and evolution.* Ecol Evol, 2023. **13**(7): p. e10327.

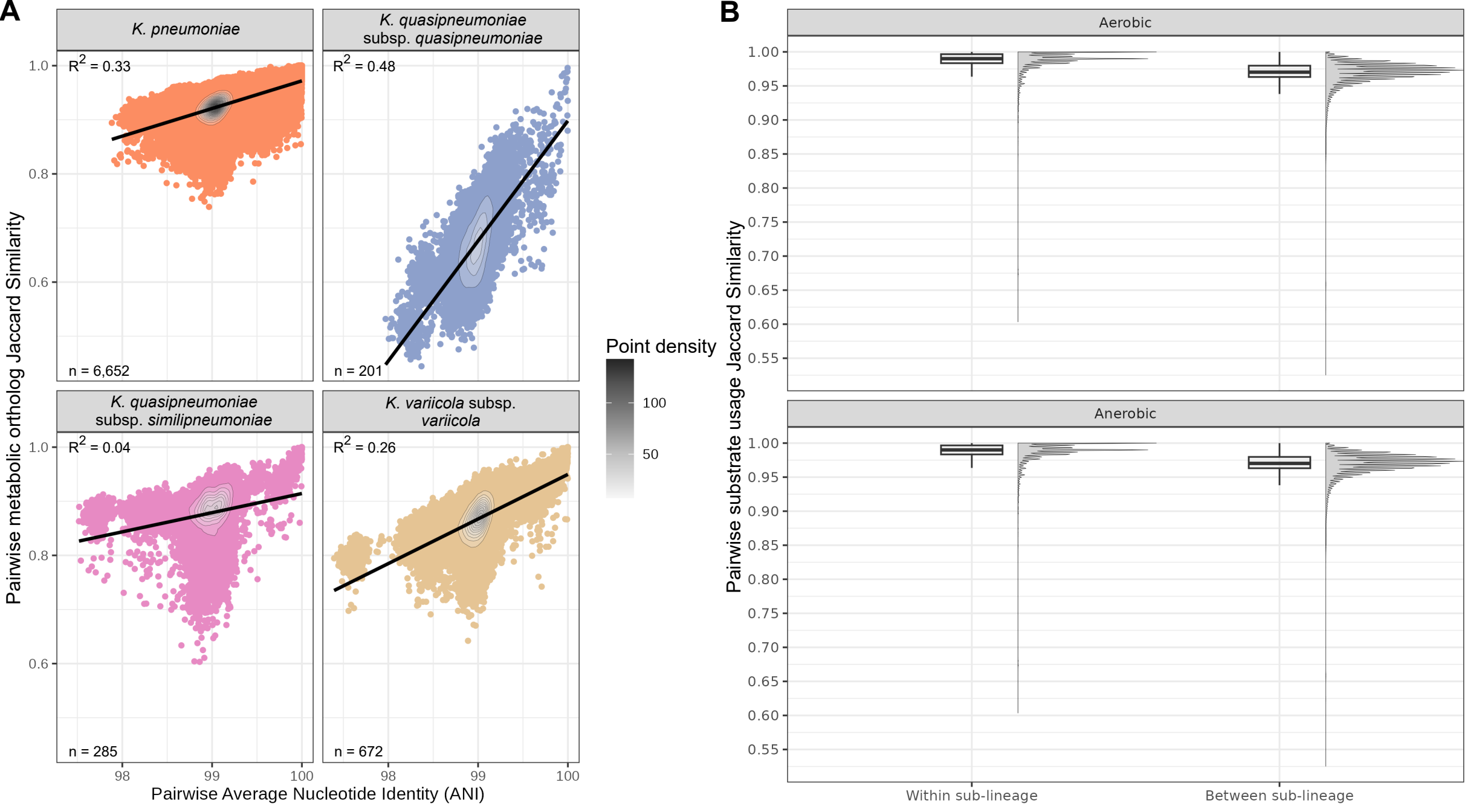
35. Chou, H.C., et al., *Isolation of a chromosomal region of Klebsiella pneumoniae associated with allantoin metabolism and liver infection.* Infect Immun, 2004. **72**(7): p. 3783-92.
36. Eberl, C., et al., *E. coli enhance colonization resistance against Salmonella Typhimurium by competing for galactitol, a context-dependent limiting carbon source.* Cell Host & Microbe, 2021. **29**(11): p. 1680-1692.e7.
37. Gül, E., et al., *Differences in carbon metabolic capacity fuel co-existence and plasmid transfer between Salmonella strains in the mouse gut.* Cell Host & Microbe, 2023. **31**(7): p. 1140-1153.e3.
38. Koike, K., et al., *Free 4-Hydroxyproline Content in Serum of Bedridden Aged People Is Elevated Due to Fracture.* Biological & Pharmaceutical Bulletin, 2000. **23**(1): p. 101-103.
39. Bialek-Davenet, S., et al., *Genomic definition of hypervirulent and multidrug-resistant Klebsiella pneumoniae clonal groups.* Emerging infectious diseases, 2014. **20**(11): p. 1812-1820.
40. Wang, M., et al., *Clinical outcomes and bacterial characteristics of carbapenem-resistant Klebsiella pneumoniae complex among patients from different global regions (CRACKLE-2): a prospective, multicentre, cohort study.* The Lancet Infectious Diseases, 2021.
41. Davis, G.S., et al., *Intermingled Klebsiella pneumoniae Populations Between Retail Meats and Human Urinary Tract Infections.* Clinical Infectious Diseases, 2015. **61**(6): p. 892-899.
42. Ellington, M.J., et al., *Contrasting patterns of longitudinal population dynamics and antimicrobial resistance mechanisms in two priority bacterial pathogens over 7 years in a single center.* Genome Biology, 2019. **20**(1): p. 184.
43. Grundmann, H., et al., *Occurrence of carbapenemase-producing Klebsiella pneumoniae and Escherichia coli in the European survey of carbapenemase-producing Enterobacteriaceae (EuSCAPE): a prospective, multinational study.* The Lancet Infectious Diseases, 2017. **17**(2): p. 153-163.
44. Franklin-Alming, F.V., et al., *Exploring Klebsiella pneumoniae in Healthy Poultry Reveals High Genetic Diversity, Good Biofilm-Forming Abilities and Higher Prevalence in Turkeys Than Broilers.* Frontiers in Microbiology, 2021. **12**(2609).
45. Gorrie, C.L., et al., *Gastrointestinal Carriage Is a Major Reservoir of Klebsiella pneumoniae Infection in Intensive Care Patients.* Clinical Infectious Diseases, 2017. **65**(2): p. 208-215.
46. Hawkey, J., et al., *A curated collection of Klebsiella metabolic models reveals variable substrate usage and gene essentiality.* Genome Research, 2022.
47. Henson, S.P., et al., *Molecular epidemiology of Klebsiella pneumoniae invasive infections over a decade at Kilifi County Hospital in Kenya.* International Journal of Medical Microbiology, 2017. **307**(7): p. 422-429.
48. Huynh, B.-T., et al., *Klebsiella pneumoniae carriage in low-income countries: antimicrobial resistance, genomic diversity and risk factors.* Gut Microbes, 2020. **11**(5): p. 1287-1299.
49. Leangapichart, T., et al., *Characterization of Klebsiella pneumoniae complex isolates from pigs and humans in farms in Thailand: population genomic structure, antibiotic resistance and virulence genes.* Journal of Antimicrobial Chemotherapy, 2021.
50. Lee, I.R., et al., *Differential host susceptibility and bacterial virulence factors driving Klebsiella liver abscess in an ethnically diverse population.* Scientific Reports, 2016. **6**(1): p. 29316.
51. Lewis, J.M., et al., *Genomic and antigenic diversity of colonizing Klebsiella pneumoniae isolates mirrors that of invasive isolates in Blantyre, Malawi.* Microbial Genomics, 2022. **8**(3).
52. Ludden, C., et al., *A One Health Study of the Genetic Relatedness of Klebsiella pneumoniae and Their Mobile Elements in the East of England.* Clinical Infectious Diseases, 2020. **70**(2): p. 219-226.

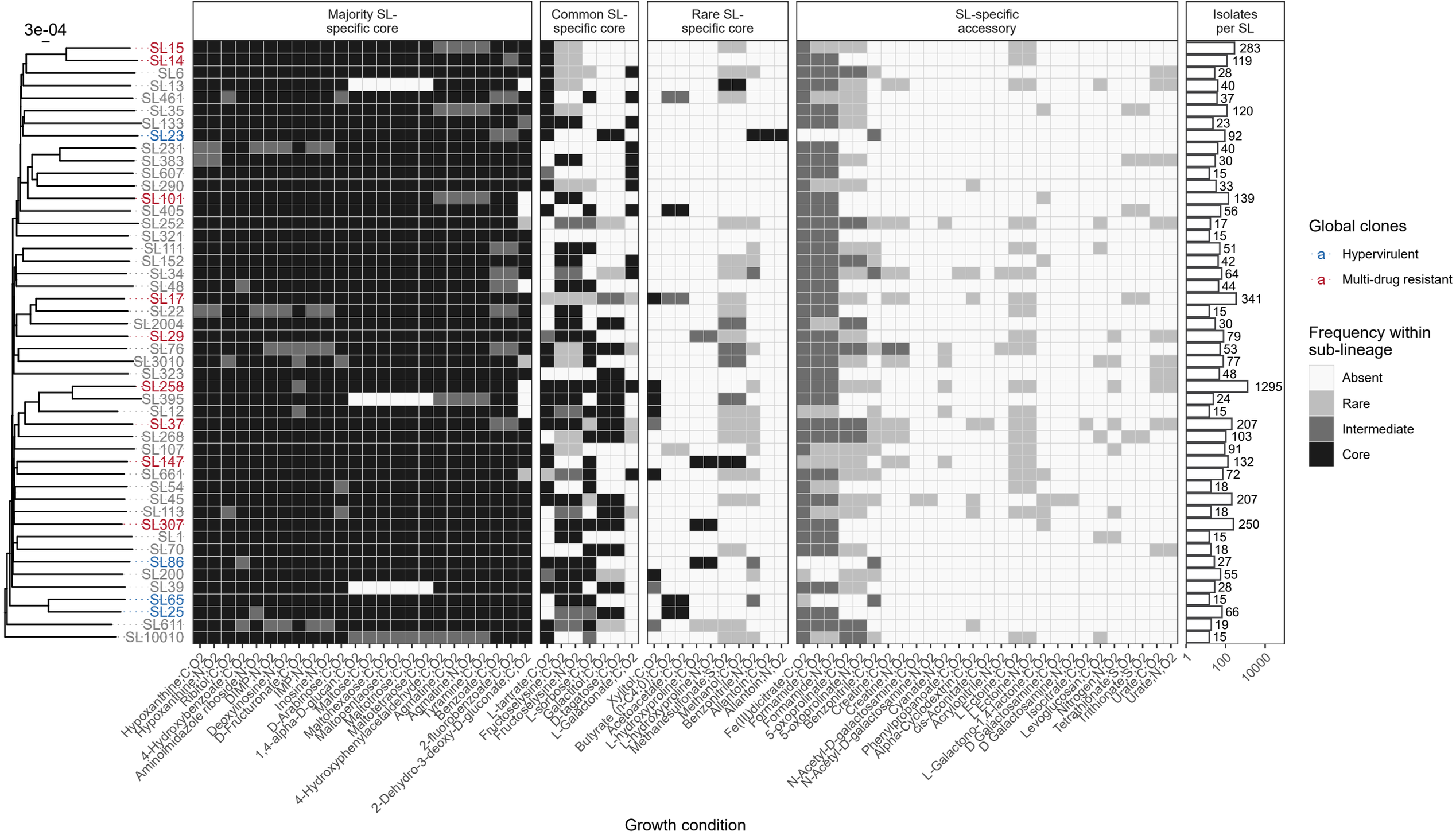
53. McDougall, F.K., et al., *Novel strains of Klebsiella africana and Klebsiella pneumoniae in Australian Fruit Bats (Pteropus poliocephalus).* Research in Microbiology, 2021: p. 103879.
54. Musicha, P., et al., *Genomic analysis of Klebsiella pneumoniae isolates from Malawi reveals acquisition of multiple ESBL determinants across diverse lineages.* Journal of Antimicrobial Chemotherapy, 2019. **74**(5): p. 1223-1232.
55. Potter, R.F., et al., *Population Structure, Antibiotic Resistance, and Uropathogenicity of Klebsiella variicola.* mBio, 2018. **9**(6): p. e02481-18.
56. Raffelsberger, N., et al., *Gastrointestinal carriage of Klebsiella pneumoniae in a general adult population: a cross-sectional study of risk factors and bacterial genomic diversity.* Gut Microbes, 2021. **13**(1): p. 1939599.
57. Rodrigues, C., et al., *High Prevalence of Klebsiella pneumoniae in European Food Products: a Multicentric Study Comparing Culture and Molecular Detection Methods.* Microbiology Spectrum, 2022. **10**(1): p. e02376-21.
58. Smit, P.W., et al., *Transmission Dynamics of Hyper-Endemic Multi-Drug Resistant Klebsiella pneumoniae in a Southeast Asian Neonatal Unit: A Longitudinal Study With Whole Genome Sequencing.* Frontiers in Microbiology, 2018. **9**.
59. Thorpe, H., et al., *One Health or Three? Transmission modelling of Klebsiella isolates reveals ecological barriers to transmission between humans, animals and the environment.* bioRxiv, 2021: p. 2021.08.05.455249.
60. Stoesser, N., et al., *Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data.* Journal of Antimicrobial Chemotherapy, 2013. **68**(10): p. 2234-2244.
61. Stoesser, N., et al., *Genome Sequencing of an Extended Series of NDM-Producing Klebsiella pneumoniae Isolates from Neonatal Infections in a Nepali Hospital Characterizes the Extent of Community- versus Hospital-Associated Transmission in an Endemic Setting.* Antimicrobial Agents and Chemotherapy, 2014. **58**(12): p. 7347-7357.
62. Vezina, B., et al., *Transmission of Klebsiella strains and plasmids within and between grey-headed flying fox colonies.* Environmental Microbiology, 2022. **24**(9): p. 4425-4436.
63. Wand Matthew, E., et al., *Characterization of Pre-Antibiotic Era Klebsiella pneumoniae Isolates with Respect to Antibiotic/Disinfectant Susceptibility and Virulence in Galleria mellonella.* Antimicrobial Agents and Chemotherapy, 2015. **59**(7): p. 3966-3972.
64. Wyres, K.L., et al., *Genomic surveillance for hypervirulence and multi-drug resistance in invasive Klebsiella pneumoniae from South and Southeast Asia.* Genome Medicine, 2020. **12**(1): p. 11.
65. Zheng, Z., et al., *Whole-genome analysis of Klebsiella pneumoniae from bovine mastitis milk in the U.S.* Environmental Microbiology, 2021. **n/a**(n/a).
66. European-Nucleotide-Archive, *enaBrowserTools*. 2020: https://github.com/.
67. Krueger, F. *Trim Galore*. 2012; Available from: https://github.com/FelixKrueger/TrimGalore.
68. Wick, R.R., et al., *Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads.* PLOS Computational Biology, 2017. **13**(6): p. e1005595.
69. Schwengers, O., et al., *Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification.* Microbial Genomics, 2021. **7**(11).
70. Lam, M.M.C., et al., *A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex.* Nature Communications, 2021. **12**(1): p. 4188.
71. Argimón, S., et al., *Rapid Genomic Characterization and Global Surveillance of Klebsiella Using Pathogenwatch.* Clin Infect Dis, 2021. **73**(Suppl_4): p. S325-s335.
72. Jolley, K.A. and M.C.J. Maiden, *BIGSdb: Scalable analysis of bacterial genome variation at the population level.* BMC Bioinformatics, 2010. **11**(1): p. 595.

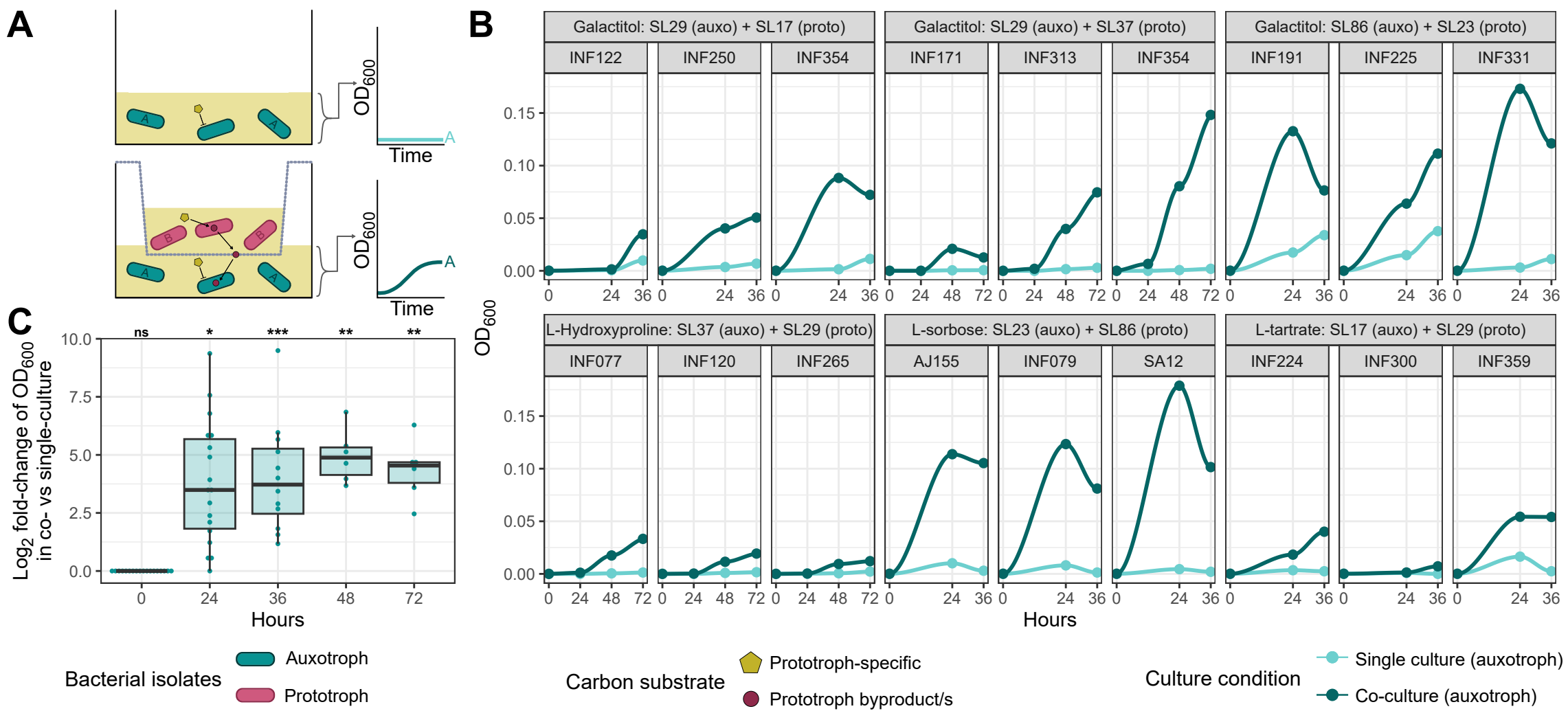
73. Jain, C., et al., *High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.* Nature Communications, 2018. **9**(1): p. 5114.
74. Lees, J.A., et al., *Fast and flexible bacterial genomic epidemiology with PopPUNK.* Genome Research, 2019. **29**(2): p. 304-316.
75. Wick, R.R. and K.E. Holt, *Assembly Dereplicator*. 2019, https://github.com/rrwick/Assembly-Dereplicator.
76. Galperin, M.Y., et al., *COG database update: focus on microbial diversity, model organisms, and widespread pathogens.* Nucleic Acids Research, 2021. **49**(D1): p. D274-D281.
77. Tatusov, R.L., et al., *The COG database: a tool for genome-scale analysis of protein functions and evolution.* Nucleic Acids Research, 2000. **28**(1): p. 33-36.
78. Kassambara, A., *ggpubr: 'ggplot2' Based Publication Ready Plots*. 2022.
79. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: The European Molecular Biology Open Software Suite.* Trends in Genetics, 2000. **16**(6): p. 276-277.
80. Aramaki, T., et al., *KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold.* Bioinformatics, 2019. **36**(7): p. 2251-2252.
81. Oksanen, J., et al., *vegan: Community Ecology Package*. 2020.
82. Whelan, F.J., M. Rusilowicz, and J.O. McInerney, *Coinfinder: detecting significant associations and dissociations in pangenomes.* Microbial Genomics, 2020. **6**(3).
83. Ibarra, R.U., J.S. Edwards, and B.O. Palsson, *Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth.* Nature, 2002. **420**(6912): p. 186-189.
84. Diener, C., M. Gibbons Sean, and O. Resendis-Antonio, *MICOM: Metagenome-Scale Modeling To Infer Metabolic Interactions in the Gut Microbiota.* mSystems, 2020. **5**(1): p. e00606-19.
85. R-Core-Team, *R: A Language and Environment for Statistical Computing*. 2020, R Foundation for Statistical Computing.
86. RStudio-Team, *RStudio: Integrated Development for R*. 2020, RStudio: Boston, MA
87. Wickham, H., et al., *Welcome to the Tidyverse.* Journal of Open Source Software, 2019. **4**(43): p. 1686.
88. Garnier, S., *viridis: Default Color Maps from 'matplotlib'*. 2018.
89. Neuwirth, E., *RColorBrewer: ColorBrewer Palettes*. 2014.
90. Aphalo, P.J., K. Slowikowski, and S. Mouksassi, *ggpmisc: Miscellaneous Extensions to 'ggplot2'*. 2021.
91. Yu, G., *aplot: Decorate a 'ggplot' with Associated Information*. 2020.
92. Zeileis, A., et al., *colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes.* Journal of Statistical Software, 2020. **96**(1): p. 1 - 49.
93. Yu, G., et al., *ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data.* Methods in Ecology and Evolution, 2017. **8**(1): p. 28-36.
94. Kassambara, A., *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. 2023.
95. Campitelli, E., *ggnewscale: Multiple Fill and Colour Scales in 'ggplot2'*. 2021.


A
K. pneumoniae Species Complex (KpSC)
K. quasivariicola
K. africana
0.005
Kv ssp. variicola
K. pneumoniae
Kv ssp. tropica
Kq ssp. quasipneumoniae
Kq ssp. similipneumoniae
n = 7835 genomes
Pangenomics + KEGG annotation
Metabolic modelling + substrate usage predictions
B
Gene count (1000s)
Rare
Intermediate
Core
KEGG Ortholog count
Proportion of genomes
Annotation
Transport
Metabolic
Hypothetical
Non-metabolic
C
Aerobic
Anaerobic
Substrate count
Absent
Rare
Int
Core
Substrate usage among KpSC
Substrate
Carbon
Nitrogen
Phosphorus
Sulfur


A
0.005
Kvt
Kvv
Kqq
Kqs
Kp
Ka
Kqv
Aerobic substrate usage (n=134)
N genomes
5
672
201
285
6652
7
13
Frequency within species
Absent
Accessory
Core
B
Amino acid transport and metabolism
Carbohydrate transport and metabolism
Lipid transport and metabolism
Nucleotide transport and metabolism
COGs per genome (%)
ns
****
Kp
Kqq
Kqs
Kvv
Species


A
K. pneumoniae
R2 = 0.33
n = 6,652
K. quasipneumoniae subsp. quasipneumoniae
R2 = 0.48
n = 201
K. quasipneumoniae subsp. similipneumoniae
R2 = 0.04
n = 285
K. variicola subsp. variicola
R2 = 0.26
n = 672
Pairwise metabolic ortholog Jaccard Similarity
Pairwise Average Nucleotide Identity (ANI)
Point density
100
50
B
Aerobic
Anerobic
Pairwise substrate usage Jaccard Similarity
Within sub-lineage
Between sub-lineage


3e-04
Majority SL-specific core
Common SL-specific core
Rare SL-specific core
SL-specific accessory
Isolates per SL
SL15
SL14
SL6
SL13
SL461
SL35
SL133
SL23
SL231
SL383
SL607
SL290
SL101
SL405
SL252
SL321
SL111
SL152
SL34
SL48
SL17
SL22
SL2004
SL29
SL76
SL3010
SL323
SL258
SL395
SL12
SL37
SL268
SL107
SL147
SL661
SL54
SL45
SL113
SL307
SL1
SL70
SL86
SL200
SL39
SL65
SL25
SL611
SL10010
283
119
28
40
37
120
23
92
40
30
15
33
139
56
17
15
51
42
64
44
341
15
30
79
53
77
48
1295
24
15
207
103
91
132
72
18
207
18
250
15
18
27
55
28
15
66
19
15
Hypoxanthine;C;O2
Hypoxanthine;N;O2
Ribitol;C;O2
4-Hydroxybenzoate;C;O2
Aminoimidazole riboside;N;O2
DIMP;N;O2
Deoxyinosine;C;O2
D-Fructuronate;C;O2
IMP;N;O2
Inosine;N;O2
D-Arabinose;C;O2
1,4-alpha-D-glucan;C;O2
Maltose;C;O2
Maltohexaose;C;O2
Maltopentaose;C;O2
Maltotriose;C;O2
Maltotetraose;C;O2
4-Hydroxyphenylacetaldehyde;C;O2
Agmatine;C;O2
Agmatine;N;O2
Tyramine;C;O2
Benzoate;C;O2
2-fluorobenzoate;C;O2
2-Dehydro-3-deoxy-D-gluconate;C;O2
L-tartrate;C;O2
Fructoselysine;C;O2
Fructoselysine;N;O2
L-sorbose;C;O2
Galactitol;C;O2
D-tagatose;C;O2
L-Galactonate;C;O2
Xylitol;C;O2
Butyrate (n-C4:0);C;O2
Acetoacetate;C;O2
L-hydroxyproline;C;O2
L-hydroxyproline;N;O2
Methanesulfonate;S;O2
Methanol;C;O2
Benzonitrile;N;O2
Allantoin;C;O2
Allantoin;N;O2
Fe(III)dicitrate;C;O2
Formamide;C;O2
Formamide;N;O2
5-oxoprolinate;C;O2
5-oxoprolinate;N;O2
Benzonitrile;C;O2
Creatine;C;O2
Creatine;N;O2
N-Acetyl-D-galactosamine;C;O2
N-Acetyl-D-galactosamine;N;O2
Cyanate;N;O2
Phenylpropanoate;C;O2
Alpha-Cyclodextrin;C;O2
cis-Aconitate;C;O2
Acrylonitrile;N;O2
L-Ectoine;C;O2
L-Ectoine;N;O2
L-Galactono-1,4-lactone;C;O2
D-Galactosamine;C;O2
D-Galactosamine;N;O2
Isocitrate;C;O2
Levoglucosan;C;O2
Nitrogen;N;O2
Tetrathionate;S;O2
Trithionate;S;O2
Urate;C;O2
Urate;N;O2
1
100
10000
Growth condition
Global clones
a Hypervirulent
a Multi-drug resistant
Frequency within sub-lineage
Absent
Rare
Intermediate
Core


A
OD600
Time
A
B
C
Log2 fold-change of OD600 in co- vs single-culture
ns
*
***
**
**
Hours
Galactitol: SL29 (auxo) + SL17 (proto)
Galactitol: SL29 (auxo) + SL37 (proto)
Galactitol: SL86 (auxo) + SL23 (proto)
INF122
INF250
INF354
INF171
INF313
INF354
INF191
INF225
INF331
L-Hydroxyproline: SL37 (auxo) + SL29 (proto)
L-sorbose: SL23 (auxo) + SL86 (proto)
L-tartrate: SL17 (auxo) + SL29 (proto)
INF077
INF120
INF265
AJ155
INF079
SA12
INF224
INF300
INF359
Bacterial isolates
Auxotroph
Prototroph
Carbon substrate
Prototroph-specific
Prototroph byproduct/s
Culture condition
Single culture (auxotroph)
Co-culture (auxotroph)