Kaptive 2.0: updated capsule and LPS locus typing for the *Klebsiella pneumoniae* species complex

Margaret M. C. Lam[1 *] (0000-0001-9590-6849), Ryan R. Wick[1] (0000-0001-8349-0778), Louise M. Judd[1] (0000-0003-3613-4839), Kathryn E. Holt[1,2] (0000-0003-3949-2471) and Kelly L. Wyres[1 *] (0000-0002-4033-6639)

[1] Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria, 3004, Australia

[2] Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK

Corresponding authors: Margaret.lam@monash.edu; Kelly.wyres@monash.edu

Keywords: *Klebsiella pneumoniae, Klebsiella* capsule K-locus genomic surveillance, K-locus, O-locus

Repositories:

1.1 Repositories:

Genome sequence from which the novel K locus KL182 was defined has been deposited under the accession JAJHNT000000000.

Abstract

The outer polysaccharide capsule and lipopolysaccharide antigens are key targets for novel control strategies targeting *Klebsiella pneumoniae* and related taxa from the *K. pneumoniae* species complex (KpSC), including vaccines, phage and monoclonal antibody therapies. Given the importance and growing interest in these highly diverse surface antigens, we had previously developed Kaptive, a tool for rapidly identifying and typing capsule (K) and outer lipopolysaccharide (O) loci from whole genome sequence data. Here, we report two significant updates, now freely available in Kaptive 2.0 (github.com/katholt/kaptive); i) the addition of 16 novel K locus sequences to the K locus reference database following an extensive search of >17,000 KpSC genomes; and ii) enhanced O locus typing to enable prediction of the clinically relevant O2 antigen (sub)types, for which the genetic determinants have been recently described. We applied Kaptive 2.0 to a curated dataset of >12,000 public KpSC genomes to explore for the first time the distribution of predicted O (sub)types across species,

sampling niches and clones, which highlighted key differences in the distributions that warrant further investigation. As the uptake of genomic surveillance approaches continues to expand globally, the application of Kaptive 2.0 will generate novel insights essential for the design of effective KpSC control strategies.

Significance as a BioResource to the community

Klebsiella pneumoniae is a major cause of bacterial healthcare associated infections globally, with increasing rates of antimicrobial resistance, including strains with resistance to the drugs of last resort. The latter have therefore been flagged as priority pathogens for the development of novel control strategies.

K. pneumoniae produce two key surface antigen sugars (capsular polysaccharide and lipopolysaccharide (LPS)) that are immunogenic and targets for novel controls such as a vaccines and phage therapy. However, there is substantial antigenic diversity in the population and relatively little is understood about the distribution of antigen types geographically and among strains causing different types of infections. Whereas laboratory-based antigen typing is difficult and rarely performed, information about the relevant synthesis loci can be readily extracted from whole genome sequence data. We have previously developed Kaptive, a freely available tool for rapid typing of *Klebsiella* capsule and LPS loci from genome sequences.

Kaptive is now used widely in the global research community and has facilitated new insights into *Klebsiella* capsule and LPS diversity. Here we present an update to Kaptive facilitating i) the identification of 16 additional novel capsule loci, and ii) the prediction of immunologically relevant LPS O2 antigen subtypes. These updates will enable enhanced sero-epidemiological surveillance for *K. pneumoniae*, to inform the design of vaccines and other novel *Klebsiella* control strategies.

Data summary

1. The updated code and reference databases for Kaptive are available at https://github.com/katholt/Kaptive
2. Genome accessions from which reference sequences of novel K loci were defined are listed in Supplementary Table 1, and genomes from which these loci were detected (along with the corresponding Kaptive output) are listed in Supplementary Table 2.
3. Accessions for the genomes screened for O types/subtypes (along with the corresponding Kaptive output) are listed in Supplementary Table 3.

The authors confirm all supporting data, code and protocols have been provided within the article or through supplementary data files.

Introduction

The *Klebsiella pneumoniae* species complex (KpSC) is a group of closely related Gram-negative bacterial taxa including the opportunistic pathogen, *Klebsiella pneumoniae* [1]. The 'K' in the ESKAPE pathogens, *K. pneumoniae* is considered one of the six most important causes of drug resistant healthcare-associated infections [2], and antimicrobial resistant strains contribute significantly to the total burden of communicable disease in high income countries [3]. In low- and middle- income countries *K. pneumoniae* is also recognised as the leading cause of Gram-negative neonatal sepsis, contributing to 10% of total neonatal sepsis deaths [4, 5]. *K. pneumoniae* with resistance to the third-generation cephalosporins and carbapenems are disseminating globally and of particular concern because they cause infections with very limited treatment options. As a consequence, there is increasing interest in developing novel anti-KpSC control strategies such as vaccines, phage and monoclonal antibody therapies [6–9].

The KpSC polysaccharide capsule and lipopolysaccharide (LPS) antigens are pathogenicity factors [10–13] and among the key targets for novel control strategies [7, 14–18]. Of particular interest, there is mounting evidence that capsule and/or LPS immunisation can elicit a protective immune response, and several anti-KpSC capsule and/or LPS vaccines have entered clinical trials [19–23]. However, considerable antigenic diversity exists within the KpSC population (>77 serologically defined capsule types [24–26], >8 LPS O antigen serotypes [16, 27], and many more predicted on the basis of genomic data [28, 29]), and there is a paucity of sero-epidemiological knowledge, which hampers efficient vaccine design. While traditional KpSC serological typing techniques are technically challenging and rarely performed, information about capsule and O antigen serotypes can be rapidly extracted from whole genome sequence (WGS) data by typing the corresponding capsule (K) and O biosynthesis loci [29, 30].

The K locus comprises a ~10-30 kbp region of the chromosome and ~10-30 genes. The conserved capsule synthesis and export machinery genes are located in the 5'- (*galF*, *cpsACP*, *wzi*, *wza, wzb*, *wzc*) and 3'- (*ugd*) most regions, and flank the genes encoding the capsule-sugar-specific synthesis machinery, Wzx capsule-specific flippase, and Wzy repeat unit polymerase [31, 32]. The O locus comprises a ~7-13 kbp region of the chromosome between the K locus and the *hisE* gene [28]. All O loci identified to-date contain the *wzm* and *wzt* genes encoding the membrane transporter complex [33], but the relative positions of these genes is not consistent [28].

Comparison of the K loci of the 77 original capsule serotype reference strains showed that all but two could be distinguished by a unique combination of genes in the centre of the locus, i.e. there was an approximate one-to-one relationship between K locus and serologically-defined K type [32].

105 In light of these findings, we previously leveraged a collection of >2500 KpSC genomes to identify
106 additional K loci, defined on the basis of unique gene content [28, 29]. A total of 59 loci were
107 identified, and labelled KL101-KL159. While the corresponding serotypes remain uncharacterised,
108 we predict that these loci encode distinct capsule types that likely correspond to the majority of the
109 10-70% of strains that have been deemed non-typeable and/or cross-reactive via serological typing
110 techniques [34–36].

111

112 To facilitate the rapid identification of K loci from KpSC whole genome sequences, we developed a
113 tool known as Kaptive [29], which uses a combination of BLASTn and tBLASTn search to identify the
114 best matching locus from the reference database, and provide an indication of the match
115 confidence. We recommend reporting matches of 'Good' confidence or higher. Low or no
116 confidence matches generally result from sequencing and/or assembly problems that cause the K
117 locus to be split across multiple assembled contigs, but can also represent novel loci that were not
118 captured in the original discovery genome set. Indeed, subsequent studies have identified 11
119 additional loci that have been incorporated into the database ([37] and see details at
120 github.com/katholt/kaptive), resulting in a total of 146 loci defined to-date. We anticipate the
121 discovery of additional novel loci, particularly given the availability of greater numbers of KpSC
122 genomes from diverse sources and locations. Notably, our recent analysis of >13,000 publicly
123 available KpSC genomes identified 19% with low or no confidence K locus matches that remain to be
124 explored [38].

125

126 The majority of distinct LPS O antigens are also associated with distinct O loci (also known as *rfb*
127 loci), distinguished on the basis of gene content [39–42]. However, key exceptions are strains
128 expressing the O1 and O2 antigens, both of which can be associated with either of two distinct O
129 loci, O1/O2v1 and O1/O2v2 [15, 41, 43]. Expression of either locus results in the production of an O2
130 antigen (D-galactan I or III), that can be converted to O1 by addition of a D-galactan II repeat unit by
131 WbbYZ (encoded elsewhere in the genome) [44].

132

133 In 2018 we extended Kaptive to support KpSC O locus typing and included discrimination of the O1
134 and O2 types by tBLASTn search for the *wbbYZ* genes, reported as shown in **Table 1** [30]. However,
135 recent experimental evidence has shown that only the WbbY protein is needed to convert O2 to O1
136 [45]. Additionally, the genetic determinants of four distinct O2 antigen subtypes have now been fully
137 elucidated (O2a, O2afg, O2ac, O2aeh) [45, 46]. The O2a antigen comprises alternating repeat units
138 of α-(1-3)–linked galactopyranose (Gal*p*) and β-(1-3)-linked galactofuranose (Gal*f*) residues (D-
139 galactan I) [40]. O2afg comprises O2a with a (1-4)-linked Gal*p* side chain, while O2aeh comprises O2a
140 with a (1-2)-linked Gal*p* side chain [43]. Like O1, O2ac comprises O2a or O2afg with an additional
141 repeat unit covalently linked to the non-reducing terminus (it also seems likely that O2aeh can be
142 modified in this way). While the O1 repeat unit comprises [-3)-α-D-Gal*p*-(1-3)-β-D-Gal*p*-(1-] (also

143 known as D-galactan II) [40], the O2ac repeat unit comprises [-3)-β-D-Glc*p*NAc-(1-5)-β-D-Gal*f*-(1-]
144 [43, 46]. These subtypes are associated with specific combinations of O loci (O1/O2v1, O1/O2v2 and
145 a novel locus identified in the O2aeh type strain, here called O1/O2v3) with/without additional
146 genes located elsewhere in the genome (*wbbY*, *gmlABD* and *wbmVW* [45, 46], see **Table 1**). There is
147 emerging evidence that these subtypes differ in terms of immunogenicity [15, 17, 47] but little is
148 known about their distribution in the KpSC population, which may have implications for the design
149 of vaccines or monoclonal antibody therapies targeting KpSC LPS.

150

151 Here we report; i) an update to the KpSC K locus database to include 16 novel loci identified from a
152 high-throughput screen of >17,000 publicly available genomes; and ii) an updated version of Kaptive
153 that can rapidly distinguish O2 subtypes to support enhanced LPS sero-epidemiological
154 investigations. We apply this update to explore the distribution of O (sub)types between KpSC
155 species and among isolates from different sources and clonal groups (CGs), including the well-known
156 globally distributed multi-drug resistant clones and the hypervirulent clones associated with severe
157 community-acquired disease.

158

159 **Table 1: Genetic determinants of *Klebsiella pneumoniae* species complex O1 and O2 outer**
160 **lipopolysaccharide antigens as reported in Kaptive.**

161

O locus	**Extra genes**	**Kaptive <v2.0**	**Kaptive ≥v2.0**	
		Locus[a]	**Locus[a]**	**Type[b]**
O1/O2v1	none	O2v1	O1/O2v1	O2a
O1/O2v2	none	O2v2	O1/O2v2	O2afg
O1/O2v3	none	N/A	O1/O2v3	O2a
O1/O2v1	*wbbYZ*[c]	O1v1	O1/O2v1	O1
O1/O2v2	*wbbYZ*[c]	O1v2	O1/O2v2	O1
O1/O2v3	*wbbYZ*[c]	N/A	O1/O2v3	O1
O1/O2v1	*wbbY* OR *wbbZ*	O1/O2v1	N/A	N/A
O1/O2v2	*wbbY* OR *wbbZ*	O1/O2v2	N/A	N/A

O1/O2v3	*wbbY* OR *wbbZ*	N/A	N/A	N/A
O1/O2v1	*wbmVW*	N/A	O1/O2v1	O2ac
O1/O2v2	*wbmVW*	N/A	O1/O2v2	O2ac
O1/O2v3	*wbmVW*	N/A	O1/O2v3	O2ac
O1/O2v1	*gmlABD*	N/A	O1/O2v1	O2aeh
O1/O2v2	*gmlABD*	N/A	O1/O2v2	O2aeh
O1/O2v3	*gmlABD*	N/A	O1/O2v3	O2aeh
O1/O2v1	*wbbY* AND *wbmVW*	N/A	O1/O2v1	O1 (O2ac)[d]
O1/O2v2	*wbbY* AND *wbmVW*	N/A	O1/O2v2	O1 (O2ac)[d]
O1/O2v3	*wbbY* AND *wbmVW*	N/A	O1/O2v3	O1 (O2ac)[d]

[a] as reported in the 'Best match locus' column in the Kaptive output

[b] predicted antigenic serotype reported in the 'Best match type' column in the Kaptive output (v2.0 and above)

[c] Kaptive v2.0 and above check only for *wbbY*.

[d] predicted antigenic serotype likely O1 but may also be O2ac (there is currently no corresponding type strain with *wbbY* and *wbmVW*)

N/A - not applicable

Methods

Identification of novel K loci

We have previously reported genotyping information for 13,156 publicly available KpSC genomes (including species, multi-locus sequence types (STs), K loci and K locus confidence calls) [38]. Here we leverage these data in combination with 3,958 additional published KpSC genomes and 638 from our private collection (**Table 2**) for which the corresponding sequence reads were *de novo* assembled using Unicycler v0.4.7. Species and STs were determined using Kleborate v2.0.3 [38] and K loci identified using Kaptive v0.7.3 [29].

Table 2: Sources of *K. pneumoniae* species complex (KpSC) genomes from which novel K loci were identified

180

Collection description	Reference	N Genomes (N low confidence K loci[a])
Curated, publicly available KpSC isolate genomes	[38] (also see refs therein)	13156 (1802)
KpSC isolate genomes from diverse niches in a single city in Italy	[48]	1965 (75)
Blood and urinary tract KpSC isolate genomes from Norwegian hospitals	[49]	868 (358)
Clinical KpSC isolate genomes from an Australian hospital	This study	638 (18)
Community rectal carriage KpSC isolate genomes from Norway	[50]	484 (36)
KpSC neonatal sepsis isolate genomes from south east Asia and Africa	[5]	276 (93)
KpSC isolate genomes from pigs and pig farmers in Thailand	[51]	253 (31)
Klebsiella variicola from various sources	[52]	112 (6)
Total		**17752 (2419)**

181 [a] Genomes harbouring K loci for which the Kaptive confidence call was 'Low' or 'None', and subsequently
182 included in this study.

183

184 Genomes for which the Kaptive K locus confidence call was "Low" or "None" were included for
185 further analysis as follows: genomes for which the Kaptive output did not indicate a fragmented K
186 locus assembly (i.e. the K locus problems column did not contain '?') were subjected to manual
187 inspection using Bandage v0.8.1 [53] to visualise the BLASTn coverage to the best match K locus and
188 assess if the genome truly harboured the best match locus, a variant thereof (insertion sequence (IS)
189 or deletion variant) or a putative novel locus. Genomic regions corresponding to putative novel loci
190 were extracted and clustered using CD-HIT-EST v4.8.1 (default parameters) [54]. A single
191 representative of each cluster was; a) annotated using Prokka v1.14.6 [55] with a reference database
192 of known KpSC K locus genes; b) subjected to BLASTn search for known KpSC K locus genes and
193 those annotated in each of the other putative novel K loci. Inspection of the BLASTn results
194 highlighted putative novel loci with similarity to each other and/or existing loci (i.e. those with

195 BLASTn hits ≥80% identity and ≥80% coverage to multiple capsule-sugar-specific synthesis genes
196 from the same reference locus). The corresponding loci were subsequently compared by BLASTn and
197 visualised with the Artemis Comparison Tool v18.0.2 [56] to clarify if they were distinct or should be
198 considered as IS or deletion variants of the same locus.

199

200 Annotations were manually curated for one representative of each of the final set of distinct novel
201 loci (**Supplementary Table 1**). Where possible, sequences without IS transposase annotations were
202 preferentially selected for curation, as is recommended in order to prevent Kaptive reporting
203 spurious gene matches to transposases that may be present in multiple copies in any location of a
204 query genome. Where no IS transposase-free representative was available (n=5 loci, subsequently
205 assigned KL173, KL176, KL181, KL184 and KL185), full length ISs were identified by BLASTn search of
206 the ISfinder database [57] and manually deleted along with their associated direct repeats and one
207 copy of the associated target site duplicated repeat to obtain a putative IS-free reference sequence.
208 Curated locus annotations were added to the reference database and Kaptive was rerun on all
209 genomes to determine the prevalence of novel loci (**Supplementary Table 2**). Only Kaptive locus
210 calls with confidence 'Good' or better are reported. Visual comparisons of the novel K loci and
211 annotated coding sequences (CDS) were generated with clinker v0.0.21 [58].

212

213

214

215 ***Implementation of O2 subtyping***

216 Kaptive takes as input a query genome assembly and a reference locus database in GenBank format.
217 Loci are identified by a note field within the sequence source information in the format:

218 /note="O locus: O1/O2v1"

219

220 Additional genes relevant to KpSC O typing (i.e those located outside of the O locus) are marked as
221 follows:

222 /note="Extra genes: wbbY"

223

224 In Kaptive 2.0 we have implemented an additional note field for reference loci that indicates the
225 corresponding serotype. E.g. for the O5 locus the corresponding O type is known to be O5 and hence
226 the corresponding note fields read:

227 /note="O locus: O5"

228 /note="O type: O5"

229

230 For loci identified from genome data and defined on the basis of gene content alone, e.g. OL101,
231 there is no known serotype and hence the corresponding note fields read:

232 /note="O locus: OL101"

233 /note="O type: unknown (OL101)"

234

235 For the O1 and O2 loci, the corresponding serotypes/subtypes are distinguished by the
236 presence/absence of additional genes located elsewhere in the genome, as is now indicated by the
237 use of the term 'special logic' in the serotype notes field e.g:

238 /note="O locus: O1/O2v1"

239 /note="O type: special logic"

240

241 After identifying the best match locus from the database (the locus with the highest BLASTn
242 coverage), Kaptive 2.0 will extract the corresponding type information from the GenBank
243 annotation. When the type is indicated as 'special logic' Kaptive will perform tBLASTn search for all
244 coding sequences within database entries marked with the note field 'Extra genes' (*wbbY*, Genbank
245 accession MG458672.1; *wbmVW*, accession MG602074.1; *gmlABD*, accession MG458670.1). tBLASTn
246 hits exceeding the minimum identity and coverage cut-offs (default 80% and 90%, respectively) are
247 interpreted to indicate the presence of the corresponding gene(s). The reported O type is
248 determined by the combination of best match O locus and any extra gene(s) detected, as specified in
249 the "Klebsiella_o_locus_primary_reference.logic" file and shown in **Table 1**. In the event that a
250 combination is not represented in the logic file, Kaptive will report the O type as "unknown."

251

252 ***Updated locus and serotype reporting***

253 In the previous version of Kaptive the distinction between the KpSC O1 and O2 types was indicated
254 in the output table in the 'Best match locus' as shown in **Table 1**. However, we have noted that this
255 has resulted in confusion in the community regarding the distinction between O loci and O types. In
256 order to clarify these differences and ensure that all relevant information is included in the output,
257 Kaptive 2.0 reports the best match locus and the predicted serotype in separate columns in the
258 output (see **Table 1**). The three distinct O loci associated with the O1 and O2 (sub)types are reported
259 simply as O1/O2v1, O1/O2v2 and O1/O2v3 in the 'Best match locus' column, meaning that genomes
260 harbouring the same O locus can be easily identified even when predicted to express a different O
261 type due to variation elsewhere in the genome.

262

263 The same approach for distinguishing 'Best match locus' from 'Best match type' is applied to all
264 databases parsed by Kaptive 2.0 in order to aid interpretation of the output and clarify what is
265 known about the relationships between loci and serotypes. In particular we hope this will clarify
266 where serotypes are or are not known. To enable backwards compatibility for reference locus
267 database that do not contain serotype note fields, the corresponding types are left blank if no type
268 information is available.

269

270 ***Exploring the distribution of O (sub)types***

271 We used Kaptive 2.0 to identify O loci and O types for 10,734 non-redundant KpSC genomes in the
272 curated collection reported previously [38]. Additionally, we supplemented this collection with two
273 systematically collected datasets of KpSC from underrepresented sources: 484 human gut carriage
274 isolate genomes reported by Raffelsburger *et al* [50], and 875 non-human associated isolate
275 genomes reported by Thorpe *et al* [48]. (We excluded human-associated KpSC genomes reported as
276 part of the latter study because these data were biased by a significant overrepresentation of the
277 ST307 and ST512 clones that were circulating in the local hospital and were already represented in
278 high numbers in our curated genome collection.) Only Kaptive calls with confidence 'Good' or better
279 are reported.

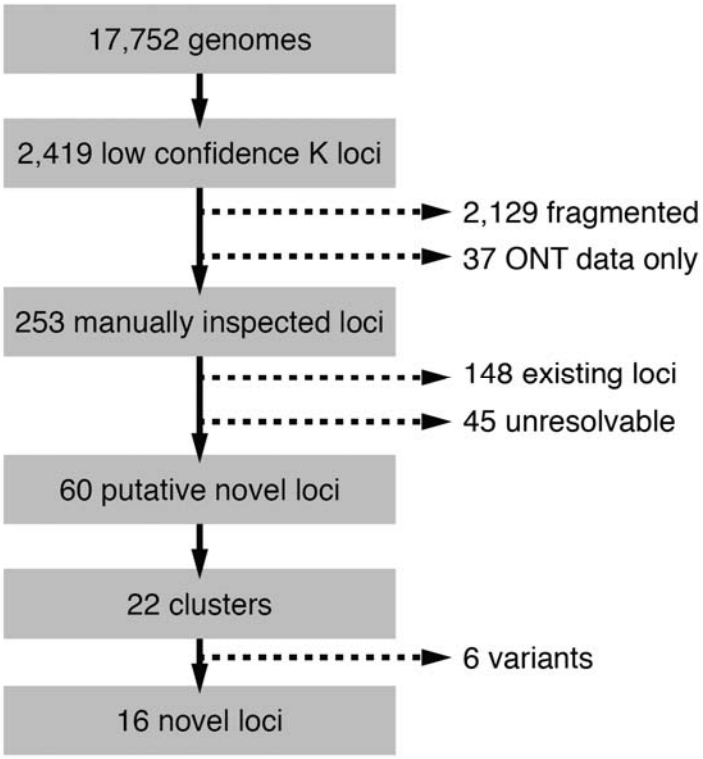
280

281 Results and Discussion

282 ***Identification and characterisation of 16 novel K loci***

283 A total of 2,419 of 17,752 KpSC genomes (13.6%) had low or no confidence Kaptive K locus calls
284 (**Figure 1**), of which the vast majority (2,129; 88%) were indicated as fragmented and 37 were
285 excluded because the K loci were likely to harbour an overrepresentation of homopolymer repeat
286 errors due to the assembly approach comprising Oxford Nanopore Technologies read data only.
287 Following manual inspection, 148 of the 253 remaining genomes were considered to match to
288 known K loci and 45 were unresolvable because the K locus was fragmented across multiple
289 assembly contigs. The former had not been confidently typed by Kaptive for a variety of reasons
290 including IS insertions, small- or large-scale deletions and/or numerous frameshift mutations that
291 were interpreted as missing genes. After exclusion of these genomes, 60 putative novel loci
292 remained and were grouped into 22 sequence clusters at 90% nucleotide identity. Sixteen clusters
293 were determined to represent true distinct and novel K loci, while 6 were identified as additional
294 IS/deletion variants of other putative novel loci or existing loci (**Figure 1,** see **Methods**).

295



296

297 **Figure 1: Process of identification of novel K loci from 17,752 *K. pneumoniae* species complex genomes.**
298 Candidate genomes were identified as those with Kaptive K locus confidence calls 'Low' or 'None,' and were
299 iteratively filtered to remove; i) fragmented or low quality locus sequences including 37 assembled using only
300 Oxford Nanopore Technologies (ONT) data; ii) true matches to existing K loci; iii) insertion sequence and/or
301 deletion variants of existing loci or other putative novel loci (see **Methods** for full details).

302

303 The final set of 16 novel loci were assigned KL171-KL186, and ranged in length from 22,624 to 30,215
304 bp and contained 19 to 25 genes (**Figure 2).** All loci contained the conserved capsule synthesis and
305 export machinery genes; *galF*, *cpsACP*, *wza*, *wzb*, *wzc*, *gnd* and *ugd*; however, only 14 loci harboured
306 *wzi* (present in >98% of all previously reported KpSC K loci [29]). The Wzi protein is not considered
307 essential for capsule production but plays a role in capsule surface attachment such that mutants
308 lacking *wzi* produce lower levels of bound and higher levels of cell-free polysaccharide [59]. In the
309 novel KL175 and KL176 loci, the *wzi* gene appeared to be replaced by four coding sequences
310 predicted to encode two hypothetical proteins, a GfcC protein and a putative lipoprotein YjbH
311 precursor. N=3/4 of these sequences had 76-84% sequence identity to those present in the KL33 and
312 KL40 reference loci, wherein *wzi* is also absent [32].

313

314 GfcC proteins are known to be involved in the export of *Escherichia coli* group 4 capsules where they
315 form a pore that allows the capsule to cross the cell membrane in a role similar to that of the KpSC

316 Wza protein [60]. However, apparently intact copies of *wza* were also present in each of the KL33,
317 KL40, KL175 and KL176 loci, so it is unclear what if any role GfcC plays in the export of these
318 capsules. Similarly, the role of the YjbH precursor protein is unclear. YjbH is indicated as a regulator
319 of exopolysaccharide expression in *E. coli* but it is encoded as part of the *yjbEFGH* [61] operon that
320 was absent from these KpSC K loci.

321

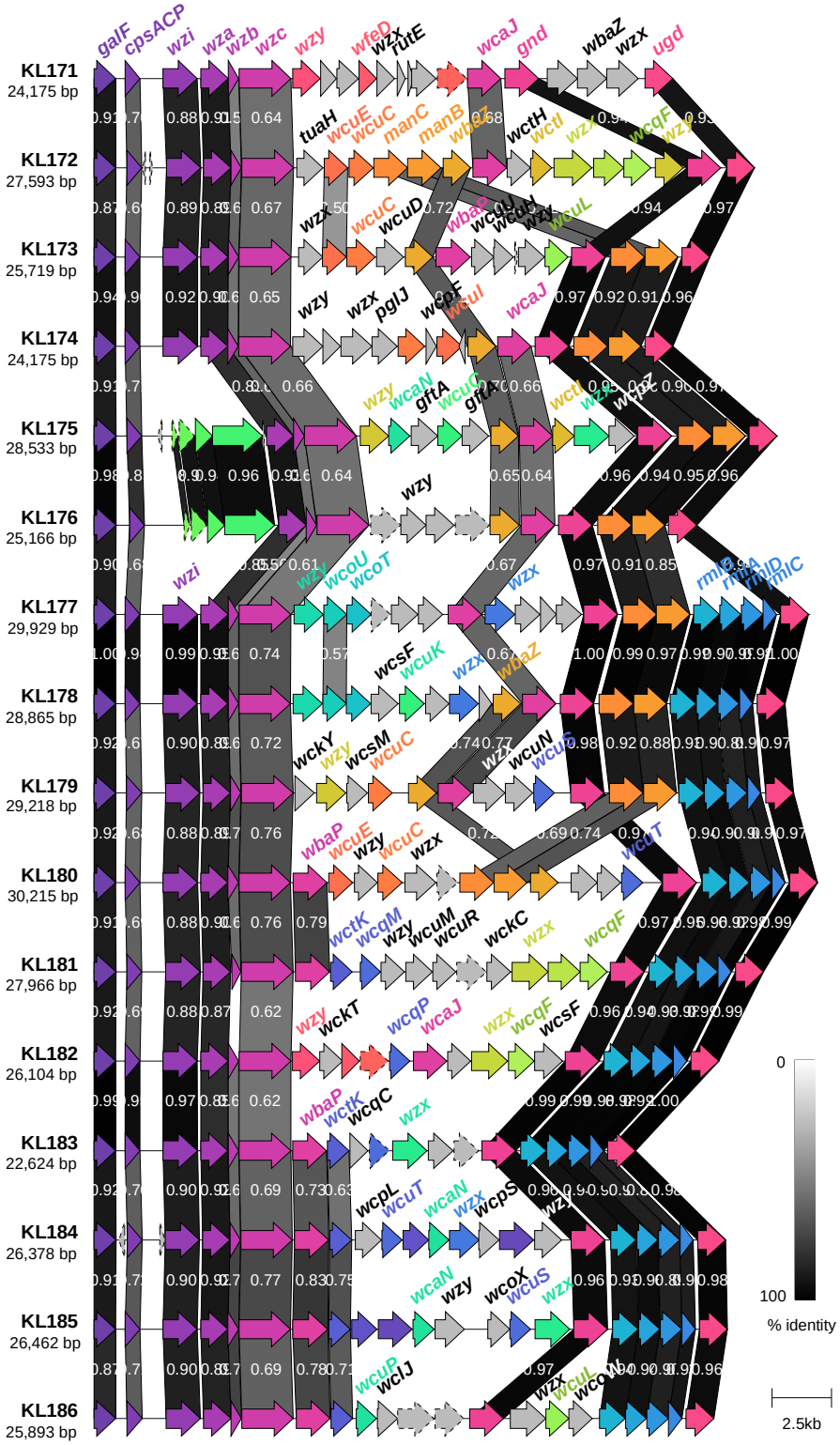
322 As expected, all novel loci harboured a variant of *wzx* and of *wzy*, encoding the capsule flippase and
323 repeat unit polymerase, respectively, as well as either *wcaJ* or *wbaP* (encoding the initiating glycosyl
324 transferases) [31]. Ten loci (63%) harboured the *rmlBADC* capsule-specific sugar synthesis genes
325 responsible for the inclusion of dTDP-L-rhamnose in the capsule and 9 (57%) harboured *manBC*
326 associated with the inclusion of GDP-D-mannose [32]. The remaining capsule-specific sugar synthesis
327 genes were each found in ≤5 loci each.

328

329 We assessed the prevalence of the novel loci among the public genome collections (**Table 2**). Most
330 (n=11; 69%) were rare, identified in ≤5 genomes. However, KL174, KL177, KL180, KL181 and KL183
331 were each detected in ≥7 and up to 33 genomes. Notably, these loci were generally associated with
332 multiple STs and geographies; n=14 KL174 genomes with 9 STs from 11 countries, n=8 KL177
333 genomes with 3 STs (n=6 ST656) from 6 countries, n=7 KL180 genomes with 6 STs from 5 countries,
334 n=9 KL181 genomes with 2 STs from 4 countries, and n=33 KL183 genomes with 22 STs from 14
335 countries. Further, 9 of the novel loci were detected in genomes sequenced from non-human
336 samples including animal (KL172, KL173, KL175, KL176, KL181, KL183, KL186), environmental (KL178,
337 KL181, KL183) and food sources (KL180, KL183). In particular, 89% of KL181 genomes were sourced
338 from non-human sources (migratory birds, bovine, swine, murine or soil).

339

340 While the vast majority of KpSC genomes can be confidently assigned K loci from the existing
341 database, as greater numbers of KpSC genomes become available we anticipate that additional
342 novel K loci will be identified. In particular, it is unlikely that the current database fully captures the
343 diversity present among KpSC circulating in the environment and non-human hosts that remain
344 comparatively underrepresented in current genome collections. Furthermore, *Klebsiella* K loci are
345 thought to undergo frequent reassortment via homologous recombination and IS-mediated
346 horizontal gene transfer both within and outside the *Klebsiella* genus [29, 62], providing a
347 mechanism for the continuous generation of novel loci, and highlighting the need for ongoing
348 population surveillance.


KL171
24,175 bp
KL172
27,593 bp
KL173
25,719 bp
KL174
24,175 bp
KL175
28,533 bp
KL176
25,166 bp
KL177
29,929 bp
KL178
28,865 bp
KL179
29,218 bp
KL180
30,215 bp
KL181
27,966 bp
KL182
26,104 bp
KL183
22,624 bp
KL184
26,378 bp
KL185
26,462 bp
KL186
25,893 bp
0
100
% identity
2.5kb

349

350 **Figure 2: Genetic structure of novel K loci identified in this study.** Coding sequences are represented by
351 arrows, labelled by their gene name where applicable, and coloured by homology (sequence identity ≥30%).
352 Coding sequences predicted to encode hypothetical proteins are represented by arrows with a dashed outline.
353 Shading between the K loci represents regions of similarity between coding sequences as identified by clinker
354 [58], and the level of similarity indicated in the legend.

355 ***O locus and O type distributions***

356 In 2018 we implemented KpSC O locus typing in Kaptive, with the capacity to distinguish the O1 and
357 O2 antigen types [30]. In the latest version of Kaptive we have additionally implemented O2 antigen
358 subtype prediction (see **Methods** and **Table 1**). Subsequently, we used Kaptive to explore the
359 distribution of O loci and predicted O types among 12,093 publicly available KpSC genomes (see
360 **Methods** and **Supplementary Table 3**). A total of 11,571 genomes (95.7%) were assigned O locus
361 calls with 'Good' confidence or better. Among these, the most common O loci were O1/O2v2
362 (n=4,015, 34.7%), O1/O2v1 (n=3,413, 29.5%) and O3b (n=1,101, 9.5%). The O1/O2v3 locus, which
363 was added to the database as part of this work (first described in association with O2aeh, GenBank
364 accession MG280710.1 [46]), was identified in only 23 genomes (0.2%).

365

366 The distribution of O loci differed by species (**Figure 3**), notably the O1/O2 loci were
367 overrepresented among *K. pneumoniae* compared to other species ($p < 2.2x10^{-16}$, OR 64.3, 95% CI
368 48.9-86.1, by Fisher's Exact test), whereas the O3/O3a and O5 loci were underrepresented among *K.*
369 *pneumoniae* ($p < 2.2x10^{-16}$, OR 0.03, 95% CI 0.025-0.035; $p < 2.2x10^{-16}$, OR 0.05, 95% CI 0.042-0.060;
370 respectively by Fisher's Exact test). Within *K. pneumoniae*, 47.6% of the isolates carrying an O1/O2
371 locus were predicted to express an O1 antigen (n=3,521/7,396), while 34.9% (n=2,581) and 16.6%
372 (n=1,224) were predicted to express O2afg and O2a, respectively. No isolates of any species were
373 predicted to express O2aeh and only 75 isolates were predicted to express O2ac (including n=69 *K.*
374 *pneumoniae*). A single *K. pneumoniae* isolate carried the O1/O2v1 locus in combination with the
375 *wbbY* gene (predicted to result in conversion of O2a to O1), and the *wbmVW* genes (predicted to
376 result in conversion of O2a to O2ac), suggesting that this isolate (NCTC8849, from a respiratory tract
377 specimen collected in the United Kingdom in 1951) may be able to express both the O1 and the O2ac
378 antigens. However, there is evidence that expression of O2ac is thermoregulated, specifically
379 downregulated at 37°C, hence it is likely that this strain expressed the O1 antigen within the human
380 host [46].

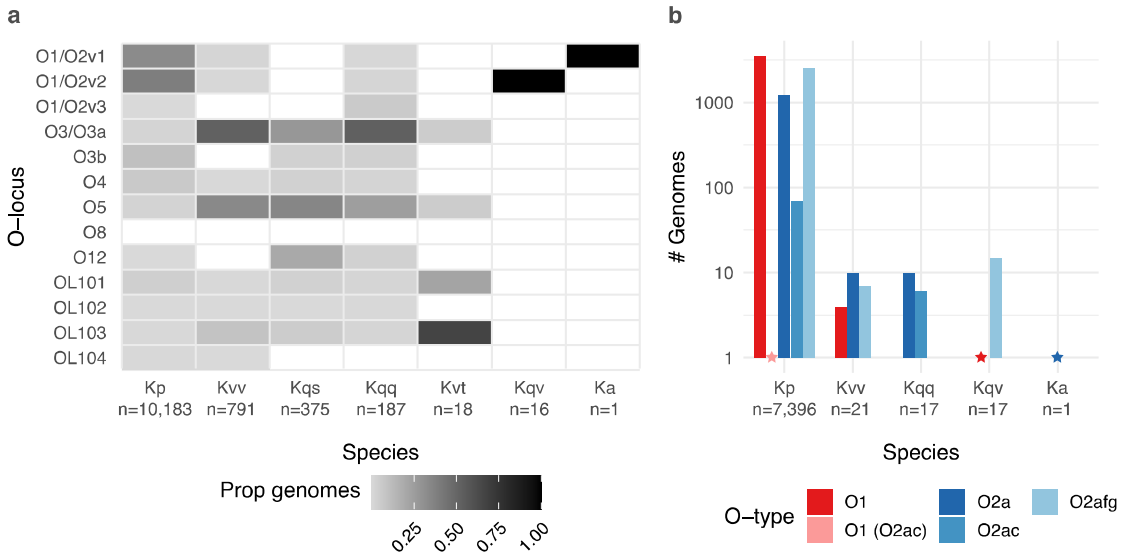
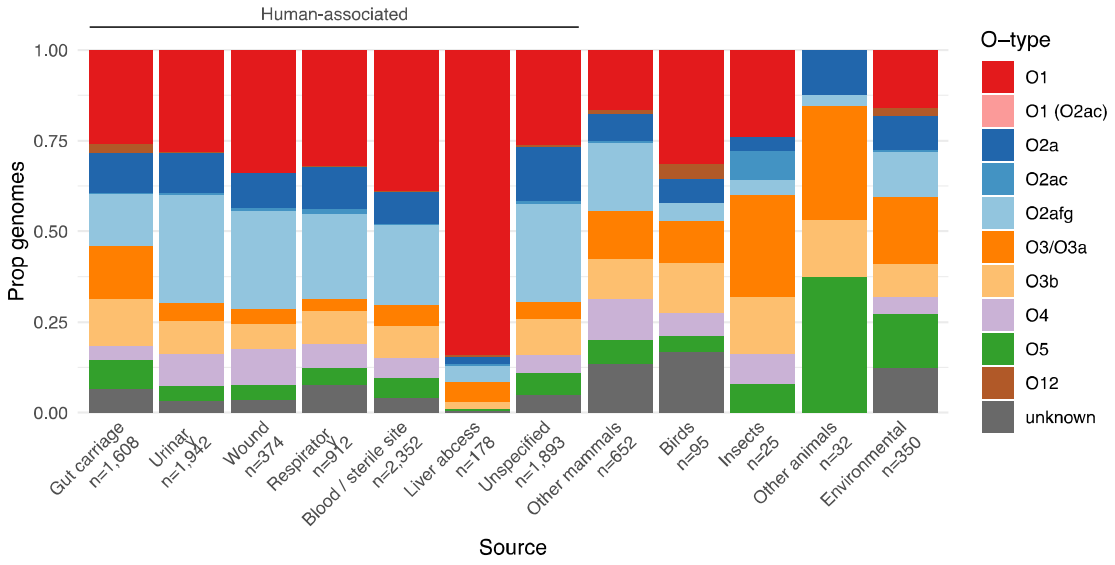


Figure 3: Distribution of O loci and predicted O1/O2 (sub)types by species. a) Heatmap showing the proportion of genomes of each species harbouring each distinct O locus. Total sample sizes for each species are indicated below x-axis labels. **b)** Bar graph showing the number of genomes of each species predicted to express O1 and O2 antigens. Only species for which at least one genome was predicted to express an O1 or O2 antigen are shown (total genomes indicated below x-axis labels). Stars indicate the position of bars of size 1. Kp, *K. pneumoniae*; Kvv, *K. variicola* subsp. *variicola*; Kqs, *K. quasipneumoniae* subsp. *similipneumoniae*; Kqq, *K. quasipneumoniae* subsp. *quasipneumoniae*; Kvt, *K. variicola* subsp. *tropica*; Ka, *K. africana*.

Our data also revealed differences in the distribution of O types by isolate source (**Figure 4**). Isolates from human specimens appeared to be enriched for types O1, O2a and/or O2afg compared to those from other hosts and/or environmental sources, which in part likely reflects the dominance *of K. pneumoniae* among clinical specimens. Interestingly, with the exception of liver abscess isolates, the prevalence of O2a was similar across all human-associated source types (8.5% - 11.7%) whereas clinical isolates were generally enriched for O2afg compared to gut carriage isolates (22.2% - 30.1% vs. 14.2%, $p < 2.2x10^{-16}$, OR 0.49, 95% CI 0.42-0.57 using Fisher's Exact test of gut carriage isolates vs all clinical categories combined) and the prevalence of O1 appeared to be higher among invasive than non-invasive isolates (38.8% among blood/sterile site isolates, 32.0% among respiratory isolates, 28.2% among urinary tract isolates). Most strikingly, the overwhelming majority of liver abscess isolates (84.3%) were predicted to express O1. We speculated that the latter may be driven at least in part by the underlying population structure of *K. pneumoniae* causing liver abscess disease, which is dominated by a small number of so-called 'hypervirulent' clones. Indeed our current and previously reported data [28] showed that the most common of these clones (ST23, ST86, ST65 and ST25) were each associated with very high prevalence of O1 (≥81%, **Figure 5**). The data also indicated that the majority of the well-known globally distributed MDR clones and other common clones were each associated with a single dominant O (sub)type, however there was diversity between clones: ST14, ST15, ST20, ST29 and ST101 were each associated with O1 (≥77%

408 each); ST307, ST258 and ST512 were associated with O2afg (≥98%); ST340 and ST437 were
409 associated with O4 (≥91%); ST147 was associated with O2a (68%). In contrast, ST11 (the putative
410 ancestor of ST258 and its descendent, ST512, as well as ST347 and ST437) and the common clones,
411 ST17 and ST37, were each associated with much greater O (sub)type diversity (**Figure 5**).

412



413

414 **Figure 4: Distribution of predicted O types by isolate source.** Bars show the proportion of isolates for each of
415 12 selected sources of interest that were predicted to encode each O antigen (as indicated in the legend). The
416 total number of isolates representing each source are indicated below the x-axis.

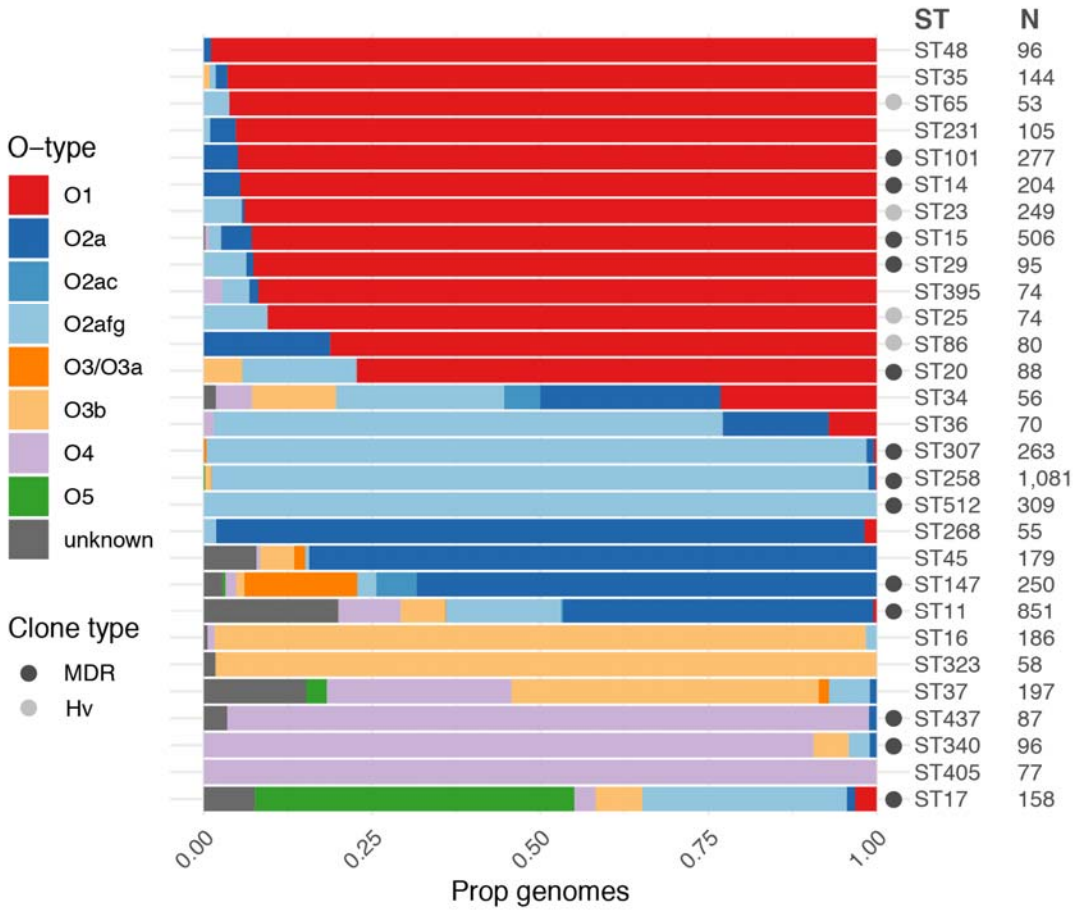


Figure 5: Distribution of predicted O types among common *K. pneumoniae* clones. Bars show the proportion of genomes predicted to express each O (sub)type within each clone (coloured as per legend). Sample sizes are indicated and only clones for which N>50 are shown. Globally distributed multi-drug resistant (MDR) and hypervirulent (Hv) clones are indicated by grey circles as per legend (clones as described previously [1], additionally ST340 and ST437 belong to clonal group 258, ST16 belongs to clonal group 20)).

Previous reports have also indicated a dominance of O1 antigen among human clinical *K. pneumoniae* isolates [27, 28, 63, 64], followed by the O3 group in earlier reports [27, 63] and the O2 group in later reports [28, 64]. While few prior studies have distinguished between O2 subtypes, a recent re-analysis of 573 KpSC genome sequences reported in [28] indicated that 42% of the 387 genomes carrying an O1/O2 O locus were predicted to express O2afg, followed by O1, a small number O2a or O2ac, and no O2aeh [46]. Additionally, the O2afg antigen has been previously associated with the ST258, ST512 and ST307 clones [15, 47]. Analysis of 276 neonatal sepsis *K. pneumoniae* genomes from South East Asia and Africa [5] revealed a dominance of O1 (54%) followed by O2afg (14%), O4 (11%), O2a (7%), and less than 5% prevalence of O3/O3a, O3b, O5 and O12. It has been suggested that the lower immunogenicity of O2afg compared to O2a and O1, may provide a selective advantage that has facilitated the widespread dissemination of these clones, i.e.

435 by enabling immune evasion [15, 17]; however, we note that our data show that several other highly
436 successful and widespread MDR clones are not associated with O2afg but instead are associated
437 with the highly immunogenic O1 (e.g. ST15, ST20, ST101), indicating that low O antigen
438 immunogenicity is not a requirement for widespread dissemination and/or other factors are also
439 playing a role (e.g. the interaction with the polysaccharide capsule).

440

441 ***Conclusions***

442 We present updated KpSC K and O loci databases and an updated version of the Kaptive genotyping
443 tool that allows rapid prediction of O1 and O2 antigen (sub)types from whole genome assemblies.
444 Application of the updated approach to a collection of >12,000 publicly available KpSC genomes
445 indicated key differences in the distribution of O1/O2 antigens by isolate source and strong
446 associations with the underlying KpSC population structure. Further investigation of these trends as
447 well as the broader population sero-epidemiology is warranted to inform the effective design of
448 novel KpSC control strategies. As genomic surveillance of KpSC continues to gain momentum,
449 Kaptive and its accompanying K and O locus databases are poised to play a key role in such
450 investigations.

451 Author statements

452 ***Authors and contributors***

453 KEH and KLW were responsible for conceptualisation, supervision, project administration and
454 acquisition of funding. RRW and KLW developed methodology, developed and/or tested software.
455 LMJ performed DNA extractions and isolate sequencing. MMCL and KLW performed investigation,
456 data curation and writing (original draft preparation). All authors performed writing (review and
457 editing).

458

459 ***Conflicts of interest***

460 The authors declare that there are no conflicts of interest.

461

462 ***Funding information***

463 This work was supported by the Bill and Melinda Gates Foundation (Investment Grant INV023041
464 awarded to KEH and KLW). KLW is supported by the National Health and Medical Research Council of
465 Australia (Investigator Grant APP1176192).

466 References

467 1. **Wyres KL, Lam MMC, Holt KE**. Population genomics of *Klebsiella pneumoniae*. *Nat Rev*

468 *Microbiol* 2020;18:344–359.

469 2. **Pendleton JN, Gorman SP, Gilmore BF**. Clinical relevance of the ESKAPE pathogens. *Expert*
470 *Rev Anti Infect Ther* 2013;11:297–308.

471 3. **Cassini A, Högberg LD, Plachouras D, Quattrocchi A, Hoxha A, *et al.*** Attributable deaths and
472 disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU
473 and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect*
474 *Dis* 2019;19:56–66.

475 4. **Okomo U, Akpalu ENK, Le Doare K, Roca A, Cousens S, *et al.*** Aetiology of invasive bacterial
476 infection and antimicrobial resistance in neonates in sub-Saharan Africa: a systematic review
477 and meta-analysis in line with the STROBE-NI reporting guidelines. *Lancet Infect Dis*
478 2019;19:1219–1234.

479 5. **Sands K, Carvalho MJ, Portal E, Thomson K, Dyer C, *et al.*** Characterization of antimicrobial-
480 resistant Gram-negative bacteria that cause neonatal sepsis in seven low- and middle-income
481 countries. *Nat Microbiol* 2021;6:512–523.

482 6. **World Health Organization**. *Global priority list of antibiotic-resistant bacteria to guide*
483 *research, discovery, and devlopment of new antibiotics.* 2017.

484 7. **Motley MP, Fries BC**. A new take on an old remedy: generating antibodies against bacteria in
485 a postantibiotic world. *mSphere* 2017;2:e00397-17.

486 8. **Aleshkin A V., Ershova ON, Volozhantsev N V., Svetoch EA, Popova A V., *et al.*** Phagebiotics
487 in treatment and prophylaxis of healthcare-associated infections. *Bacteriophage*
488 2016;6:e1251379.

489 9. **Assoni L, Girardello R, Converso TR, Darrieux M**. Current stage in the development of
490 *Klebsiella pneumoniae* vaccines. *Infect Dis Ther* 2021;Epub ahead of print.

491 10. **Cortés G, Borrell N, Astorza D, Gómez C, Sauleda J, *et al.*** Molecular analysis of the
492 contribution of the capsular polysaccharide and the lipopolysaccharide O side chain to the
493 virulence of *Klebsiella pneumoniae* in a murine model of pneumonia. *Infect Immun*
494 2002;70:2583–2590.

495 11. **Lawlor MS, Hsu J, Rick PD, Miller VL**. Identification of *Klebsiella pneumoniae* virulence
496 determinants using an intranasal infection model. *Mol Microbiol* 2005;58:1054–1073.

497 12. **Evrard B, Balestrino D, Dosgilbert A, Bouya-Gachancard JL, Charbonnel N, *et al.*** Roles of
498 capsule and lipopolysaccharide O antigen in interactions of human monocyte-derived
499 dendritic cells and *Klebsiella pneumoniae*. *Infect Immun* 2010;78:210–219.

500 13. **March C, Cano V, Moranta D, Llobet E, Pérez-Gutiérrez C, *et al.*** Role of bacterial surface
501 structures on the interaction of *Klebsiella pneumoniae* with phagocytes. *PLoS One* 2013;8:1–
502 16.

503 14. **D'Andrea MM, Marmo P, Henrici De Angelis L, Palmieri M, Ciacci N, *et al.*** ϕbO1E, a newly
504 discovered lytic bacteriophage targeting carbapenemase-producing Klebsiella pneumoniae of
505 the pandemic clonal group 258 clade II lineage. *Sci Rep* 2017;7:1–8.

506 15. **Szijártó V, Guachalla LM, Hartl K, Varga C, Banerjee P, *et al.*** Both clades of the epidemic
507 KPC-producing *Klebsiella pneumoniae* clone ST258 share a modified galactan O-antigen type.
508 *Int J Med Microbiol* 2016;306:89–98.

509 16. **Guachalla LM, Stojkovic K, Hartl K, Kaszowska M, Kumar Y, *et al.*** Discovery of monoclonal
510 antibodies cross-reactive to novel subserotypes of *K. pneumoniae* O3. *Sci Rep* 2017;7:6635.

511 17. **Pennini ME, De Marco A, Pelletier M, Bonnell J, Cvitkovic R, *et al.*** Immune stealth-driven O2
512 serotype prevalence and potential for therapeutic antibodies against multidrug resistant
513 *Klebsiella pneumoniae*. *Nat Comms* 2017;8:1991.

514 18. **Seeberger PH, Pereira CL, Khan N, Xiao G, Diago-Navarro E, *et al.*** A semi-synthetic
515 glycoconjugate vaccine candidate for carbapenem-resistant Klebsiella pneumoniae. *Angew*
516 *Chem Int Ed Engl* 2017;56:13973–13978.

517 19. **Ravinder M, Liao KS, Cheng YY, Pawar S, Lin TL, *et al.*** A synthetic carbohydrate-protein
518 conjugate vaccine candidate against *Klebsiella pneumoniae* serotype K2. *J Org Chem*
519 2020;85:15964–15997.

520 20. **Feldman MF, Mayer Bridwell AE, Scott NE, Vinogradov E, McKee SR, *et al.*** A promising
521 bioconjugate vaccine against hypervirulent Klebsiella pneumoniae. *Proc Natl Acad Sci U S A*
522 2019;116:18655–18663.

523 21. **Campbell WN, Hendrix E, Cryz S, Cross AS**. Immunogenicity of a 24-valent Klebsiella capsular
524 polysaccharide vaccine and an eight-valent Pseudomonas O-polysaccharide conjugate vaccine
525 administered to victims of acute trauma. *Clin Infect Dis* 1996;23:179–181.

526 22. **Hegerle N, Choi M, Sinclair J, Amin MN, Ollivault-Shiflett M, *et al.*** Development of a broad
527 spectrum glycoconjugate vaccine to prevent wound and disseminated infections with
528 Klebsiella pneumoniae and Pseudomonas aeruginosa. *PLoS One* 2018;13:e0203143.

529 23. **ClinicalTrials.gov [Internet]**. Identifier: NCT04959344 Safety and Immunogenicity of a
530 *Klebsiella pneumoniae* tetravalent bioconjugate vaccine (Kleb4V).

531 24. **Edmunds PN**. Further *Klebsiella* capsule types. *J Infect Dis* 1954;94:65–71.

532 25. **Edwards PR, Fife MA**. Capsule types of *Klebsiella*. *J Infect Dis* 1952;91:92–104.

533 26. **Ørskov IDA, Fife-Asbury MA**. New *Klebsiella* capsular antigen, K82, and the deletion of five of
534 those previously assigned. *Int J Syst Bacteriol* 1977;27:386–387.

535 27. **Trautmann M, Ruhnke M, Rukavina T, Held TK, Cross AS, *et al.*** O-antigen seroepidemiology
536 of *Klebsiella* clinical isolates and implications for immunoprophylaxis of *Klebsiella* infections.

537 *Clin Diag Lab Immun* 1997;4:550–555.

538 28. **Follador R, Heinz E, Wyres KL, Ellington MJ, Kowarik M, *et al.*** The diversity of Klebsiella
539 pneumoniae surface polysaccharides. *Microb Genomics*;2. Epub ahead of print 2016. DOI:
540 10.1099/mgen.0.000073.

541 29. **Wyres KL, Wick RR, Gorrie C, Jenney A, Follador R, *et al.*** Identification of *Klebsiella* capsule
542 synthesis loci from whole genome data. *Microb Genomics*;2. Epub ahead of print 2016. DOI:
543 10.1099/mgen.0.000102.

544 30. **Wick RR, Heinz E, Holt KE, Wyres KL**. Kaptive Web: user-friendly capsule and
545 lipopolysaccharide serotype prediction for *Klebsiella* genomes. *J Clin Microbiol*
546 2018;56:e00197-18.

547 31. **Whitfield C**. Biosynthesis and assembly of capsular polysaccharides in *Escherichia coli. Annu*
548 *Rev Biochem* 2006;75:39–68.

549 32. **Pan Y-J, Lin T-L, Chen C-T, Chen Y-Y, Hsieh P-F, *et al.*** Genetic analysis of capsular
550 polysaccharide synthesis gene clusters in 79 capsular types of *Klebsiella* spp. *Nat Sci Rep*
551 2015;5:15573.

552 33. **Cuthbertson L, Kimber MS, Whitfield C**. Substrate binding by a bacterial ABC transporter
553 involved in polysaccharide export. *Proc Natl Acad Sci U S A* 2007;104:19529–19534.

554 34. **Cryz SJ, Mortimer PM, Mansfield V, Germanier R**. Seroepidemiology of *Klebsiella* bacteremic
555 isolates and implications for vaccine development. *J Clin Microbiol* 1986;23:687–690.

556 35. **Tsay R-W, Siu LK, Fung C-P, Chang F-Y**. Characteristics of bacteremia between community-
557 acquired and nosocomial *Klebsiella pneumoniae* infection. *Arch Intern Med* 2002;162:1021–
558 1027.

559 36. **Jenney AW, Clements A, Farn JL, Wijburg OL, McGlinchey A, *et al.*** Seroepidemiology of
560 *Klebsiella pneumoniae* in an Australian tertiary hospital and its implications for vaccine
561 development. *J Clin Microbiol* 2006;44:102–107.

562 37. **Wyres KL, Nguyen TNT, Lam MMC, Judd LM, van Vinh Chau N, *et al.*** Genomic surveillance
563 for hypervirulence and multi-drug resistance in invasive *Klebsiella pneumoniae* from South
564 and Southeast Asia. *Genome Med* 2020;12:11.

565 38. **Lam MM., Wick R., Watts SC, Cerdeira LT, Wyres KL, *et al.*** A genomic surveillance framework
566 and genotyping tool for *Klebsiella pneumoniae* and its related species complex. *Nat Commun*
567 2021;12:4188.

568 39. **Fang C, Shih Y, Cheong C, Yi W**. Rapid and accurate determination of lipopolysaccharide O-
569 antigen types in Klebsiella pneumoniae with a novel PCR-based O-genotyping method.
570 2015;54:666–675.

571 40. **Whitfield C, Richards JC, Perry MB, Clarke BR, Maclean LL**. Expression of two structurally
572 distinct D-galactan O-antigens in the lipopolysaccharide of Klebsiella pneumoniae serotype
573 O1. *J Bacteriol* 1991;173:1420–1431.

574 41. **Clarke BR, Whitfield C**. Molecular cloning of the rfb region of Klebsiella pneumoniae serotype
575 O1:K20: The rfb gene cluster is responsible for synthesis of the D-galactan I O polysaccharide.
576 *J Bacteriol* 1992;174:4614–4621.

577 42. **Sugiyama T, Kido N, Kato Y, Koide N, Yoshida T, *et al.*** Evolutionary relationship among *rfb*
578 gene clusters synthesizing mannose homopolymer as O-specific polysaccharides in
579 *Escherichia coli* and *Klebsiella pneumoniae*. *Gene* 1997;198:111–113.

580 43. **Kelly RF, Perry MB, MacLean LL, Whitfield C**. Structures of the O-antigens of *Klebsiella*
581 serotypes 02 (2a,2e), 02 (2a,2e,2h), and 02 (2a,2f,2g), members of a family of related D-
582 galactan O-antigens in *Klebsiella* spp. *J Endotoxin Res* 1995;2:131–140.

583 44. **Hsieh PF, Wu MC, Yang FL, Chen CT, Lou TC, *et al.*** D-galactan II is an immunodominant
584 antigen in O1 lipopolysaccharide and affects virulence in *Klebsiella pneumoniae*: Implication
585 in vaccine design. *Front Microbiol* 2014;5:608.

586 45. **Kelly SD, Clarke BR, Ovchinnikova OG, Sweeney RP, Williamson ML, *et al.*** *Klebsiella*
587 *pneumoniae* O1 and O2ac antigens provide prototypes for an unusual strategy for
588 polysaccharide antigen diversification. *J Biol Chem* 2019;294:10863–10876.

589 46. **Clarke BR, Ovchinnikova OG, Kelly SD, Williamson ML, Jennifer E, *et al.*** Molecular basis for
590 the structural diversity in serogroup O2-antigen polysaccharides in *Klebsiella pneumoniae*. *J*
591 *Biol Chem* 2018;293:4666–4679.

592 47. **Bulati M, Busà R, Carcione C, Iannolo G, Di Mento G, *et al.*** *Klebsiella pneumoniae*
593 lipopolysaccharides serotype O2afg induce poor inflammatory immune responses *ex vivo*.
594 *Microorganisms* 2021;9:1–15.

595 48. **Thorpe H, Booton R, Kallonen T, Gibbon MJ, Couto N, *et al.*** One Health or Three?
596 Transmission modelling of Klebsiella isolates reveals ecological barriers to transmission
597 between humans, animals and the environment. *bioRxiv* 2021;2021.08.05.455249.

598 49. **Fostervold A, Hetland MAK, Bakksjø R, Bernhoff E, Holt KE, *et al.*** A nationwide genomic
599 study of clinical Klebsiella pneumoniae in Norway 2001-2015: Introduction and spread of
600 ESBL facilitated by CG15 and CG307. *bioRxiv* 2021;2021.07.16.452602.

601 50. **Raffelsberger N, Hetland MAK, Svendsen K, Småbrekke L, Löhr IH, *et al.*** Gastrointestinal
602 carriage of Klebsiella pneumoniae in a general adult population in Norway: a cross-sectional
603 study of risk factors and bacterial genomic diversity. *Gut Microbes* 2021;13:1939599.

604 51. **Leangapichart T, Lunha K, Jiwakanon J, Angkititrakul S, Järhult JD, *et al.*** Characterization of
605 Klebsiella pneumoniae complex isolates from pigs and humans in farms in Thailand:

606 Population genomic structure, antibiotic resistance and virulence genes. *J Antimicrob*
607 *Chemother* 2021;76:2012–2016.

608 52. **Potter RF, Lainhart W, Twentyman J, Wallace MA, Wang B, *et al.*** Population structure,
609 antibiotic resistance, and Uropathogenicity of *Klebsiella variicola*. *MBio* 2018;9:e02481-18.

610 53. **Wick RR, Schultz MB, Zobel J, Holt KE**. Bandage: interactive visualization of *de novo* genome
611 assemblies. *Bioinformatics* 2015;31:3350–3352.

612 54. **Fu L, Niu B, Zhu Z, Wu S, Li W**. CD-HIT: Accelerated for clustering the next-generation
613 sequencing data. *Bioinformatics* 2012;28:3150–3152.

614 55. **Seemann T**. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–
615 2069.

616 56. **Carver TJ, Rutherford KM, Berriman M, MA R, Barrell BG, *et al.*** ACT: the Artemis comparison
617 tool. *Bioinformatics* 2005;21:3422–3423.

618 57. **Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M**. ISfinder: the reference centre for
619 bacterial insertion sequences. *Nucleic Acids Res* 2006;34:D32–D36.

620 58. **Gilchrist CLM, Chooi Y-H**. clinker & clustermap.js: automatic generation of gene cluster
621 comparison figures. *Bioinformatics* 2021;Epub ahead of print.

622 59. **Rahn A, Beis K, Naismith JH, Whitfield C**. A novel outer membrane protein, Wzi, is involved in
623 surface assembly of the Escherichia coli K30 group 1 capsule. *J Bacteriol* 2003;185:5882–90.

624 60. **Sathiyamoorthy K, Mills E, Franzmann TM, Rosenshine I, Saper MA**. The crystal structure of
625 Escherichia coli group 4 capsule protein GfcC reveals a domain organization resembling that
626 of Wza. *Biochemistry* 2011;50:5465–5476.

627 61. **Ferrières L, Aslam SN, Cooper RM, Clarke DJ**. The yjbEFGH locus in Escherichia coli K-12 is an
628 operon encoding proteins involved in exopolysaccharide production. *Microbiology*
629 2007;153:1070–1080.

630 62. **Holt KE, Lassalle F, Wyres KL, Wick R, Mostowy RJ**. Diversity and evolution of surface
631 polysaccharide synthesis loci in Enterobacteriales. *ISME J* 2020;14:1713–1730.

632 63. **Hansen DS, Mestre F, Dome A, Toma JM**. Klebsiella pneumoniae lipopolysaccharide O typing:
633 Revision of prototype strains and O-group distribution among clinical isolates from different
634 sources and countries. *J Clin Microbiol* 1999;37:56–62.

635 64. **Choi M, Hegerle N, Nkeze J, Sen S, Jamindar S, *et al.*** The diversity of lipopolysaccharide (O)
636 and capsular polysaccharide (K) antigens of invasive Klebsiella pneumoniae in a multi-country
637 collection. *Front Microbiol* 2020;11:1249.

638