Development of *Kaptive* databases for *Vibrio parahaemolyticus* O- and K-antigen serotyping

Authors

Linda van der Graaf – van Bloois[1,2], Hongyou Chen[3], Jaap A. Wagenaar[1,2,4] and Aldert L. Zomer[1,2]

[1] Department of Biomolecular Health Sciences, Faculty of Veterinary Medicine, Utrecht University, the Netherlands

[2] WHO Collaborating Centre for Reference and Research on *Campylobacter* and Antimicrobial Resistance from a One Health Perspective / OIE Reference Laboratory for Campylobacteriosis, Utrecht, the Netherlands

[3] Shanghai Municipal Center for Disease Control and Prevention, No. 1380, Zhong Shan Xi Rd., Shanghai 200336, PR China

[4] Wageningen Bioveterinary Research, Lelystad, the Netherlands

Running title: Development of *V. parahaemolyticus* O- and K-antigen Kaptive databases

Corresponding author: Aldert Zomer, a.l.zomer@uu.nl

Keywords: Vibrio parahaemolyticus, serotyping, whole genome sequencing, O-locus, K-locus

Abbreviations: KL, K-locus; KUT, K-Untypable; KLUT, K-locus Untypable; LPS/CPSgc, LPS and CPS gene clusters; MLST, Multi Locus Sequence Typing; OL, O-locus; *V. parahaemolyticus*, Vibrio parahaemolyticus; WGS, whole genome sequencing;

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

Data summary: The sequence data are deposed at the European Nucleotide Archive (ENA) under projects PRJEB39490 and PRJNA483379, and accession numbers are listed in Supplemental Table 1.

Abstract

Vibrio parahaemolyticus is an important food-borne human pathogen and is divided in 16 O-serotypes and 71 K-serotypes. Agglutination tests are still the gold standard for serotyping, but many *V. parahaemolyticus* isolates are not typable by agglutination. An alternative for agglutination tests is serotyping using whole genome sequencing data. In this study, *V. parahaemolyticus* isolates are serotyped and sequenced, and all known and several novel O- and K-loci are identified. We developed *Kaptive* databases for all O- and K-loci after manual curation of the loci. These *Kaptive* databases with the identified *V. parahaemolyticus* O- and K -loci can be used to identify the O- and K-serotypes of *V. parahaemolyticus* isolates from genome sequences.

42 **Introduction**

43

44 *Vibrio parahaemolyticus* (*V. parahaemolyticus*) is an important food-borne human pathogen that
45 naturally inhabits marine environments worldwide, and can cause acute gastroenteritis and
46 septicemia in human (1,2). *V. parahaemolyticus* is typically serotyped on the basis of its heat-stable
47 somatic antigen (O) and its capsular antigen (K) and classified into 16 O-serotypes and 71 K-
48 serotypes.

49

50 Serotyping of *V. parahaemolyticus* is important for pathogen detection and epidemiological
51 surveillance. Many *V. parahaemolyticus* serotypes have been identified as pandemic clones, and
52 certain serotypes, for example O3:K6, O1:KUT and O4:K68, are generally considered to be more
53 virulent than others (3,4). Agglutination tests are the gold standard for *V. parahaemolyticus*
54 serotyping but frequently, *V. parahaemolyticus* strains are non-typable with agglutination tests in
55 routine screening (3,5–7). For several pathogenic bacterial species, e.g. several pathogenic *E. coli*
56 serogroups (8) and *Vibrio cholerae* serogroup O1 and O139 (9,10), PCR methods are developed to
57 detect the specific serogroups, but no PCR methods are available for *V. parahaemolyticus*
58 serotyping.

59

60 The software tool *Kaptive* was developed for rapid O- and K-loci typing of *Klebsiella* strains from
61 whole genome sequences (11). This tool includes an option to use other, self-created databases. The
62 genomic regions associated with the somatic synthesis locus (O-locus) and capsule synthesis locus
63 (K-locus) for *V. parahaemolyticus* are generally found on the same genomic location (2,12,13), which
64 makes it possible to locate and extract the O- and K-loci nucleotide sequences from whole genome
65 sequencing data of *V. parahaemolyticus* strains. Recently, the algorithm VPsero was developed,
66 which identified serogroup-specific genes as markers *in silico*, but only marker genes for 12 O- and
67 43 K-serotypes were covered (14).

68

69 The aim of this study was to identify all 16 O- and 71 K-loci of *V. parahaemolyticus* by serotyping,
70 sequencing and analyzing in-house isolates and investigating public data of serotyped strains and
71 manually curate reference databases for O- and K-loci, which can be used to determine the O- and K-
72 serotype of *V. parahaemolyticus* strains from whole genome sequencing data.
73

74 **Materials and Method**

75

76 *Datasets of* V. parahaemolyticus

77 In this study, 56 *V. parahaemolyticus* isolates were sequenced and serotyped. For whole genome
78 sequencing (WGS), DNA was extracted with TIANamp Bacteria DNA Kit (Tiagen Biotech (Beijing) CO.,
79 LTD, Beijing, China) according to the manufacturer's recommendation and DNA quality control was
80 performed using agarose gel electrophoresis and the Qubit dsDNA HS Assay Kit. DNA libraries and
81 DNA nanoballs (DNB) were constructed on BGISP-100 platform (WuHan MGI Tech Co.,Ltd, Wuhan,
82 China) with input of 150ng DNA. 50 bp single-end reads were generated with BGISEQ-50 sequencer
83 (WuHan MGI Tech Co.,Ltd, Wuhan, China) and assembled with SPAdes v3.12 (15). All sequence reads
84 of strains sequenced in this study are available under accession PRJEB39490 and PRJNA483379 of
85 the ENA short read archive.

86 Two datasets of public available *V. parahaemolyticus* genomes were included, consisting of 776
87 genomes from NCBI Genbank (Accession date: 14 December 2017) and 1498 genomes from PATRIC
88 database (16) (Accession date: 23 July 2020) were downloaded.

89 Taxonomic classification of all genomes was performed by *in silico* 16S DNA analysis with SPINGO
90 v1.3 (17) and genomes with ambiguous 16S results were classified as *Vibrio parahaemolyticus* using
91 KmerFinder v3.2 (18).

92

93 *Serotyping of* V. parahaemolyticus *strains*

94 Serotyping of the 56 in this study sequenced strains was performed by slide agglutination according
95 to the manufacturer's instructions, using the 65 type-K serum set and 11 type-O serum set (Denka
96 Seiken CO., LTD, Tokyo, Japan).

97

98 *Development* Kaptive *databases*

99 I. Identification of O-loci

100 *V. parahaemolyticus* is divided into 16 O-serotypes (1,2). The O-locus has been defined located
101 between *dgkA* and *gmhD* genes, which encode a diacylglylcerol-kinase and an epimerase
102 respectively (2). With BlastN (minimum coverage of 80% and minimum identity of 80%), the flanking
103 genes of the O-region, *dgkA* and *gmhD*, were searched in all genome sequences. The O-loci
104 sequences were extracted from these sequences manually.
105
106 II. Identification of K-loci

107 *V. parahaemolyticus* is divided in 71 K-serotypes (2). The K-locus has been defined located between
108 gene *gmhD* which encodes an epimerase at one side, and at the other side gene *rjg* encoding a
109 metallo-hydrolase (13). With BlastN (minimum coverage of 80% and minimum identity of 80%), the
110 flanking genes of the K-region, *gmhD* and *rjg,* were located in all 1601 genome sequences. In some
111 cases where the *rjg* gene was not found, gene *ugd* or gene *gtaB* was selected as flanking gene. The
112 K-loci sequences were extracted from these sequences manually.
113
114 III. Development of the *Kaptive* databases

115 The extracted O- and K-loci sequences were annotated using Prokka v1.13 (19). For annotating the
116 loci, a custom database was used which was built with the annotations of the previous described loci
117 of *V. parahaemolyticus* (1,2). Genes of the O- and K-loci respectively were clustered using Roary with
118 an 80% cutoff on amino acid identity (20). The Roary gene presence absence table was used to
119 create gene presence-absence clusters of the O- and K-loci respectively and the unique gene
120 presence-absence loci were linked to the known serotypes of the genomes.
121 For the identified O- and K-loci, the nomenclature of the *Klebsiella* capsule synthesis loci (11) was
122 used; each distinct O- and K-locus was designated as OL (O-locus) and KL (K-locus), followed by an
123 unique number. The O- and K-loci of known serotypes were assigned with the same number as the
124 corresponding O- and K-serotypes, e.g. O-serotype 1 is encoded by the OL1 locus etcetera. O- and K-

125 loci with unknown serotypes were assigned with numbers starting from 101 (e.g. OL101 and KL101).
126 Variants of loci with maximal two genes different by insertion or deletion were given the suffix -1.
127 Eight gene patterns of K-Untypable (KUT) strains were found and these patterns are assigned in the
128 k-loci database as KLUT (K-locus UnTypable) followed by subsequent numbers 1-8
129 Of each loci, one genome was selected as reference and the nucleotide sequences of the O- and K-
130 loci of the selected genomes were added to the *Kaptive* database file and curated manually.
131 The *V. parahaemolyticus* O-locus and K-locus *Kaptive* databases are available on the following
132 Github page: https://github.com/aldertzomer/vibrio_parahaemolyticus_genomoserotyping and will
133 be added to Kaptive web: https://kaptive-web.erc.monash.edu/
134
135 *Comparison of genes in O- and K-loci*
136 The reference genomes of the identified O- and K-loci were aligned using Parsnp v1.2 (21),
137 recombination regions were filtered using Gubbins v2.3.4 (22) and the tree was built with FastTree
138 v2.1.8 (23). The tree was visualized with iTol v6.5.4 (24) and Multi Locus Sequence Typing (MLST)
139 was performed in silico with the MLST software (Seemann) using PubMLST typing schemes (26).
140 Gene cluster comparisons of the O- and K-loci are made with Clinker v0.0.37 (27).
141
142 *Performance of* Kaptive *databases compared to VPSero*
143 Of the total set of 831 selected genomes from Genbank (n=775) and in-house sequenced genomes
144 (n=56), 436 genomes had a known O-serotype and 294 genomes had a known K-serotype, whereas
145 the PATRIC genomes had all unknown O- and K-loci serotypes. The O- and K-loci of the genomes with
146 the known serotypes were all serotyped by the newly developed *Kaptive* databases and VPSero (14).
147 Furthermore, performance of the *Kaptive* databases was tested and compared with VPSero by using
148 the VPSero sequence data collection, deposited into CNGB Sequence Archive project CNP0000343
149 described by Bian *et al* (14).

150 **Results and Discussion**

151

152 *Development of* Kaptive *databases for V. parahaemolyticus O- and K-loci*

153 The study started with a set of 2330 *V. parahaemolyticus* genomes, consisting of 56 in this study
154 sequenced genomes, 776 genomes downloaded from NCBI GenBank and 1498 from the PATRIC
155 database. The genomes that were *in silico* identified as *V. parahaemolyticus* with 16S and
156 KmerFinder analysis were included. For the genomes downloaded from NCBI GenBank and PATRIC
157 database, only genomes that contained the complete loci on one contig were included. This
158 selection ended-up with a total set of 1601 genomes, consisting of 56 in-house sequenced strains,
159 775 from Genbank and 770 genomes from PATRIC database. For the development of the O- and K-
160 loci databases, only genomes that contained both flanking genes of the loci on one contig were
161 used. No isolate or genome were available for K-serotypes 26 and 71, and these serotypes were
162 therefore not included in the database.

163 The selected genomes that were used as reference for each O- and K-loci are shown in Supplemental
164 Table 1. For each O-serotype, unique loci were identified, except for serotype O3 and O13 genomes,
165 since the loci of these serotypes contained the same genes (1). Therefore, loci of serotype O3 and
166 O13 could not be distinguished based on gene presence and are assigned in the *Kaptive* O-database
167 as "O3_or_O13". It is possible that a second cluster modifies the O-antigen, similar to what has been
168 described for *Shigella* (28), however an O-modifying locus has not been described in literature for
169 O13.

170 Eight gene patterns of K-UnTypable (KUT) strains previously described were found and these
171 patterns are assigned in the *Kaptive* database as K-Locus UnTypable (KLUT) followed by subsequent
172 numbers 1-8 (Supplemental Table 1).

173 For both O- and K-loci, not all gene patterns could be assigned to specific serotypes, because these
174 gene patterns were found in genomes with unknown serotypes from NCBI and PATRIC databases.
175 These unknown gene patterns were assigned for O-loci as OL101 and for K-loci as KL101-KL157
176 (Supplemental Table 1).

177 For some O- and K-serotypes, variants in the loci gene patterns by insertion of deletion of maximal
178 two genes were found, and therefore, these serotypes have two gene patterns included in the
179 database, distinguished with the addition of "-1" to the locus name (OL1, OL4, OL7 and KL20, KL30,
180 KL68, KLUT4) (Supplemental Figures 1 and 2).
181 Examination of the phylogenetic tree of the reference isolates showed the O- and K-loci are not
182 associated with MLST sequence types, e.g. six different K-loci and four different O-loci are found in
183 ST3 reference isolates (Figure 1).
184
185 *Performance of* Kaptive *databases*
186 - Genome collection of this study
187 To investigate the performance of the newly developed *Kaptive* databases, 436 genomes with a
188 known O-serotype and 295 genomes with a known K-serotype were used from our genome
189 collection. Of the 436 known O-serotypes, 398 O-serotypes were correctly identified (Figure 2A,
190 Supplemental Table 1) and 265 of the 295 known K-serotypes were correctly identified (Figure 2B,
191 Supplemental Table 1).
192 The tool VPSero was developed for 43 K- and 12 O-serotypes (14), and covers part of the known 16
193 O-serotypes and 71 K-serotypes. The performance of VPSero was tested with a set of 420 genomes
194 for O-serotyping and 274 genomes for K-serotyping from our genome collection, consisting of
195 genomes with known serotypes of the 12 and 46 O- and K-serotypes respectively which are included
196 in VPSero. Of this set, 343 out of 420 genomes with known O-serotypes were correctly identified
197 (Supplemental Figure 2A, Supplemental Table 1) and 183 out of 274 genomes with known K-
198 serotypes (Supplemental Figure 2B, Supplemental Table 1). The newly developed *Kaptive* databases
199 outcompete VPSero with 395/420 correctly identified O-serotype genomes and 254/274 correctly
200 identified K-serotype genomes (Supplemental Figure 2AB, Supplemental Table 1).
201 - Sequence collection of VPsero

202 In the publication of VPsero, intact LPS and CPS gene clusters (LPS/CPSgc) were identified and
203 extracted, and deposited in CNGB Sequence Archive project CNP0000343. These gene clusters did
204 not contain the full O-serotype loci of the newly developed O-serotype *Kaptive* database, and
205 therefore, these sequence data collection could not be used to test the performance of the O-
206 serotype *Kaptive* database, the K-serotype loci were available however. The LPS/CPSgcs were tested
207 with both *Kaptive* K-database and VPSero, and results are listed in Supplemental Table 2. One K58
208 gene cluster was identified as K6 by both tested tools, possibly containing a wrong serotype in the
209 CNGB database. Of the remaining 56 gene clusters with known K-serotypes, seven were not typable
210 by VPSero and two were mis-identified by *Kaptive* databases.
211
212 The mis-identified O- and K-serotypes from our collection were all genomes downloaded from NCBI.
213 Since we only have the genomes and not the strains, we cannot check if these strains were
214 serotyped correctly. O-serotype O14 has been recently described (2) and only one genome with this
215 O-serotype is available. In our dataset of 831 genomes, 14 NCBI genomes with serotype O5 are
216 identified with the *Kaptive* database as the new serotype O14. It is very likely that these genomes
217 are mis-serotyped, because the serum for serotype O14 was not available when these genomes
218 were serotyped.
219
220 For several K-serotypes, only one genome sequence was available and we could not determine if
221 there is variation in these K-loci gene content. Furthermore, several of the genomes sequenced in
222 this study were sequenced with a 50-bp single-end BGI sequencer, resulting in a higher number of
223 short contigs, therefore for several of these genomes, the K-locus was not assembled on one single
224 contig. For these genomes, the contigs with flanking genes were selected and concatenated
225 manually. It is possible that some genes are missing in these K-locus reference sequences, although
226 these likely represent repeat sequences. If more *V. parahaemolyticus* genome sequences with closed

227 K-loci become available, the *Kaptive* K-serotype database will be updated with the closed K-locus
228 sequence of these K-serotypes.
229
230 **Conclusion**
231 The in this study developed *Kaptive* databases with the identified 16 O- and 71 K -loci can be used to
232 identify the O- and K-serotypes of *V. parahaemolyticus* isolates from whole genome sequencing
233 data. The variation of K-antigen loci is much higher than expected as we identified 57 new K-locus
234 variants.
235
236 **Author Statements**
237 ***Author contributions***
238 Designed the study; L.G., J.W., A.Z. Performed the experiments; L.G., H.C., A.Z. Analyzed the data;
239 L.G., A.Z. Writing the paper; all authors.
240
241 ***Conflicts of interest***
242 The authors declare that they have no competing interests.
243
244 ***Funding information***
245 Sequencing of isolates was supported by the National Science and Technology Major Project of
246 China (No. 2018ZX10305409-003).
247
248 ***Acknowledgements***
249 We thank the Kaptive team (Thomas Stanton, Kelly Wyres, Ryan Wick and Kathryn Holt) for the
250 undertaking to include the *Vibio parahaemolyticus* databases to Kaptive web.
251

252 **References**

253 1. Chen M, Guo D, Wong HC, Zhang X, Liu F, Chen H, et al. Development of O-serogroup specific
254 PCR assay for detection and identification of Vibrio parahaemolyticus. Int J Food Microbiol.
255 2012;159(2):122–9.
256 2. Guo X, Liu B, Chen M, Wang Y, Wang L, Chen H, et al. Genetic and serological identification of
257 three Vibrio parahaemolyticus strains as candidates for novel provisional O serotypes. Int J
258 Food Microbiol. 2017;245:53–8.
259 3. Jones JL, Lüdeke CHM, Bowers JC, Garrett N, Fischer M, Parsons MB, et al. Biochemical,
260 serological, and virulence characterization of clinical and oyster Vibrio parahaemolyticus
261 isolates. J Clin Microbiol. 2012;50(7):2343–52.
262 4. Nair GB, Ramamurthy T, Bhattacharya SK, Dutta B, Takeda Y, Sack DA. Global dissemination of
263 Vibrio parahaemolyticus serotype O3:K6 and its serovariants. Clin Microbiol Rev.
264 2007;20(1):39–48.
265 5. Gavilan RG, Zamudio ML, Martinez-Urtaza J. Molecular Epidemiology and Genetic Variation of
266 Pathogenic Vibrio parahaemolyticus in Peru. PLoS Negl Trop Dis. 2013;7(5).
267 6. Gil AI, Miranda H, Lanata CF, Prada A, Hall ER, Barreno CM, et al. O3:K6 Serotype of Vibrio
268 parahaemolyticus identical to the global pandemic clone associated with diarrhea in Peru. Int
269 J Infect Dis. 2007;11(4):324–8.
270 7. Hashii N, Kondo S, Iguchi T, Nishibuchi M, Hisatsune K. Chemical and serological properties of
271 lipopolysaccharides from Vibrio parahaemolyticus O-untypeable strains isolated from
272 patients. Microbiol Immunol. 2000;44(4):229–34.
273 8. Wang Q, Ruan X, Wei D, Hu Z, Wu L, Yu T, et al. Development of a serogroup-specific
274 multiplex PCR assay to detect a set of Escherichia coli serogroups based on the identification
275 of their O-antigen gene clusters. Mol Cell Probes. 2010;24(5):286–90.
276 9. Albert MJ, Islam D, Nahar S, Qadri F, Falklind S, Weintraub A. Rapid detection of Vibrio
277 cholerae O139 bengal from stool specimens by PCR. J Clin Microbiol. 1997;35(6):1633–5.
278 10. Hoshino K, Yamasaki S, Mukhopadhyay AK, Chakraborty S, Basu A, Bhattacharya SK, et al.
279 Development and evaluation of a multiplex PCR assay for rapid detection of toxigenic Vibrio
280 cholerae O1 and O139. FEMS Immunol Med Microbiol. 1998;20(3):201–7.
281 11. Wyres KL, Wick RR, Gorrie C, Jenney A, Follador R, Thomson NR, et al. Identification of
282 Klebsiella capsule synthesis loci from whole genome data. Microb genomics.
283 2016;2(12):e000102.
284 12. Okura M, Osawa R, Tokunaga A, Morita M, Arakawa E, Watanabe H. Genetic analyses of the
285 putative O and K antigen gene clusters of pandemic Vibrio parahaemolyticus. Microbiol
286 Immunol. 2008;52(5):251–64.
287 13. Chen Y, Dai J, Morris JG, Johnson JA. Genetic analysis of the capsule polysaccharide (K
288 antigen) and exopolysaccharide genes in pandemic Vibrio parahaemolyticus O3:K6. BMC
289 Microbiol. 2010;10(1):274.
290 14. Bian S, Jia Y, Zhan Q, Wong NK, Hu Q, Zhang W, et al. VPsero: Rapid Serotyping of Vibrio
291 parahaemolyticus Using Serogroup-Specific Genes Based on Whole-Genome Sequencing
292 Data. Front Microbiol. 2021;12:2446.
293 15. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new
294 genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol.
295 2012;19(5):455–77.
296 16. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial
297 bioinformatics database and analysis resource. Nucleic Acids Res. 2014;42(D1):581–91.
298 17. Allard G, Ryan FJ, Jeffery IB, Claesson MJ. SPINGO: A rapid species-classifier for microbial
299 amplicon sequences. BMC Bioinformatics. 2015;16(1):1–8.
300 18. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Moller N, et al. Rapid
301 whole-genome sequencing for detection and characterization of microorganisms directly
302 from clinical samples. J Clin Microbiol [Internet]. 2014;52(1):139–46.
303 19. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics.

2014;30(14):2068–9.
20. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691–3.
21. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. 2014;15(11):1–15.
22. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43(3):e15.
23. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3).
24. Letunic I, Bork P. Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. Bioinformatics. 2007;23(1):127–8.
25. Seemann T. MLST. Available from: https://github.com/tseemann/mlst.
26. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010;11(1):1–11.
27. Gilchrist CLM, Chooi Y-H. clinker & clustermap.js: Automatic generation of gene cluster comparison figures. Bioinformatics. 2021;37(16); 2473-2475.
28. Lehane AM, Korres H, Verma NK. Bacteriophage-encoded glucosyltransferase GtrII of Shigella flexneri: Membrane topology and identification of critical residues. Biochem J. 2005;389(1):137–43.

328 **Figures and Tables**

329

330 **Figure 1.** Phylogenetic tree based on core genome analysis of O- and K-representative isolates,
331 including the O- and K-loci, and MLST STs.

332

333 **Figure 2. Performance of the *Kaptive* database and VPSero on genomes with known (A) O- and (B)**
334 **K-serotypes.** In green the percentages of correctly typed genomes with *Kaptive* databases, in orange
335 the percentages of incorrectly typed genomes with *Kaptive* databases, in blue the percentages of
336 correctly typed genomes with VPSero, in red the percentages of incorrectly typed genomes with
337 VPSero, in grey the percentages of the with VPSero not-typable genomes. The numbers in the bars
338 represent the number of genomes.

339

340 **Supplemental Figure 1.** Gene cluster comparisons of the identified O-loci. All similar gene homologs
341 have been assigned an unique color and are linked with similar color boxes. Hypothetical proteins
342 are assigned with an "1".

343

344 **Supplemental Figure 2.** Gene cluster comparisons of the identified K-loci. All similar gene homologs
345 have been assigned an unique color and are linked with similar color boxes. Hypothetical proteins
346 are assigned with an "1".

347

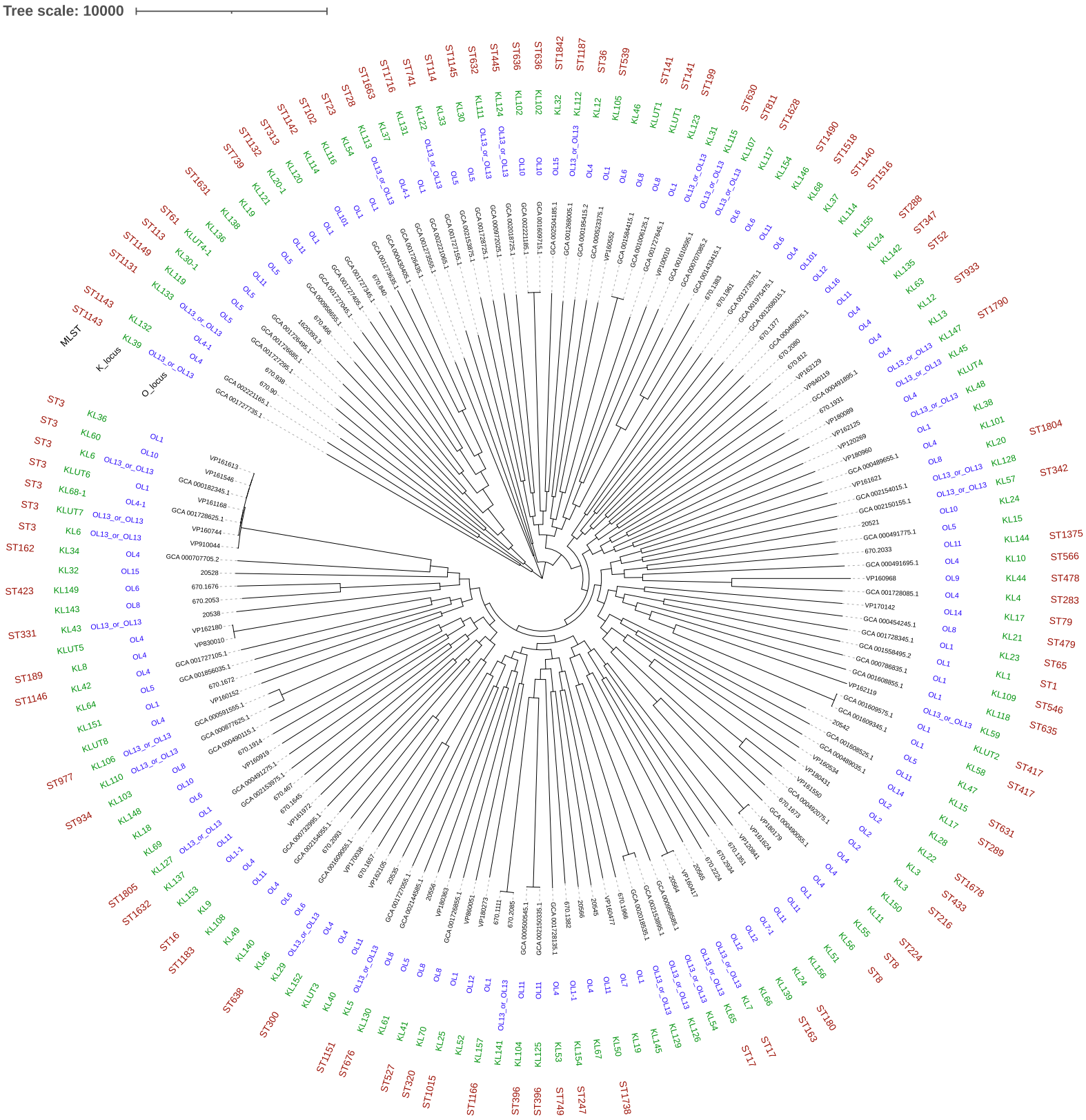
348 **Supplemental Table 1.** Selected reference genomes used for the O- and K-loci *Kaptive* databases.
349 Included are the numbers of genomes with known O- and K-serotypes and numbers of correctly
350 identified genomes with *Kaptive* databases and VPSero.

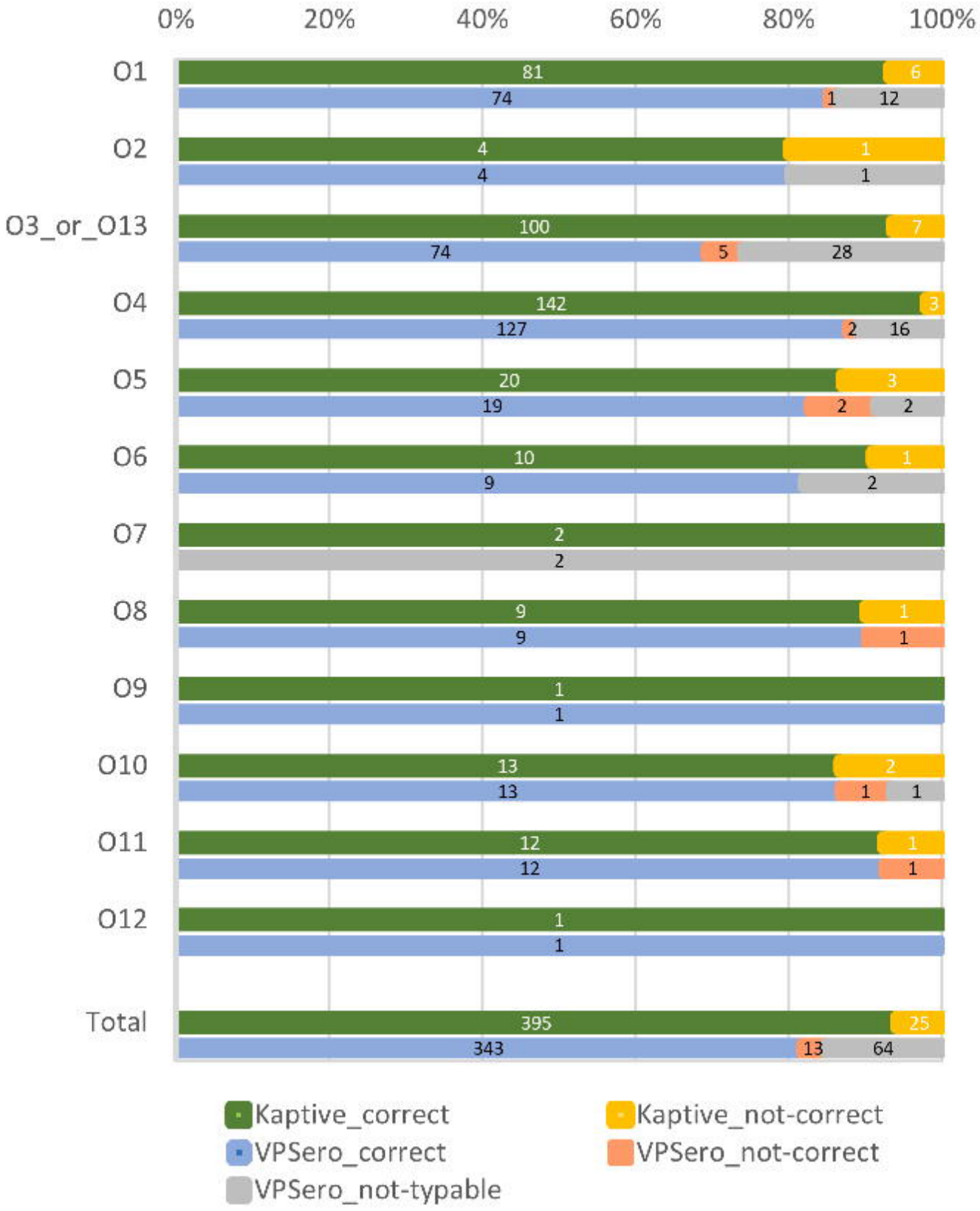
351

352 **Supplemental Table 2.** Performance of *Kaptive* databases and VPSero on sequence collection of
353 VPSero.

354

355


Tree scale: 10000
MLST
K_locus
O_locus


0%
20%
40%
60%
80%
100%
O1
81
6
74
1
12
O2
4
1
4
1
O3_or_O13
100
7
74
5
28
O4
142
3
127
2
16
O5
20
3
19
2
2
O6
10
1
9
2
O7
2
2
O8
9
1
9
1
O9
1
1
O10
13
2
13
1
1
O11
12
1
12
1
O12
1
1
Total
395
25
343
13
64
Kaptive_correct
Kaptive_not-correct
VPSero_correct
VPSero_not-correct
VPSero_not-typable

0%
20%
40%
60%
80%
100%
K1
K3
K4
K5
K6
K8
K9
K10
K11
K12
K13
K15
K17
K18
K19
K20
K21
K23
K25
K28
K29
K30
K31
K32
K33
K34
K36
K37
K38
K39
K41
K42
K44
K47
K48
K49
K53
K55
K56
K58
K60
K63
K68
K69
K70
K71
Total
Kaptive_correct
Kaptive_not-correct
VPSero_correct
VPSero_not-correct
VPSero_not-typable