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

Authors

Linda van der Graaf – van Bloois^{1,2}, Hongyou Chen³, Jaap A. Wagenaar^{1,2,4} and Aldert L. Zomer^{1,2}

¹ Department of Infectious Diseases and Immunology, Faculty of Veterinary Medicine, Utrecht University, the Netherlands

² WHO Collaborating Centre for Campylobacter / OIE Reference Laboratory for Campylobacteriosis

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, Department of Infectious Diseases and Immunology, Utrecht University, Utrecht, The Netherlands, +31 (0) 30 253 36 03, a.l.zomer@uu.nl

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 genome sequence data. In this study, we manually identified all known O- and K-loci from *V. parahaemolyticus* isolates which we serotyped and sequenced, and extracted additional O- and K-loci from publicly available genomes. 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.

Introduction

Vibrio parahaemolyticus (*V. parahaemolyticus*) is an important food-borne human pathogen that naturally inhabits marine environments worldwide, and can cause acute gastroenteritis and septicemia in human (M. Chen et al. 2012; Guo et al. 2017). *V. parahaemolyticus* is typically serotyped on the basis of its heat-stable somatic antigen (O) and its capsular antigen (K) and classified into 16 O-serotypes and 71 K-serotypes.

Serotyping of *V. parahaemolyticus* is important for pathogen detection and epidemiological surveillance. Many *V. parahaemolyticus* serotypes have been identified as pandemic clones, and certain serotypes, for example O3:K6, O1:KUT and O4:K68, are generally considered to be more virulent than others (Jones et al. 2012; Nair et al. 2007). Agglutination tests are the gold standard for *V. parahaemolyticus* serotyping but frequently, *V. parahaemolyticus* strains are non-typable with agglutination tests in routine screening (Gavilan, Zamudio, and Martinez-Urtaza 2013; Gil et al. 2007; Hashii et al. 2000; Jones et al. 2012). For several pathogenic bacterial species, e.g. several pathogenic *E. coli* serogroups (Wang et al. 2010) and *Vibrio cholerae* serogroup O1 and O139 (Albert et al. 1997; Hoshino et al. 1998), PCR methods are developed to detect the specific serogroups, but no PCR methods are available for *V. parahaemolyticus* serotyping.

Recently, the software tool *Kaptive* was developed for rapid O- and K-loci typing of *Klebsiella* strains from whole genome sequences (Wyres et al. 2016). This tool includes an option to use other, self-created databases. The genomic regions associated with the somatic synthesis locus (O-locus) and capsule synthesis locus (K-locus) for *V. parahaemolyticus* are generally found on the same genomic location (Okura et al. 2008; Y. Chen et al. 2010; Guo et al. 2017), which makes it possible to locate and extract the O- and K-loci nucleotide sequences from whole genome sequencing data of *V. parahaemolyticus* strains.

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

60

Materials and Method

Sequences of V. parahaemolyticus

A total of 1602 *V. parahaemolyticus* genomes were used in this study, of which 775 genomes were publicly available and downloaded from NCBI Genbank, 770 additional genomes were downloaded from PATRIC (Wattam et al. 2014) and 57 isolates were in-house sequenced and serotyped. For strains sequenced in this study, DNA was extracted with TIANamp Bacteria DNA Kit (Tiagen Biotech (Beijing) CO., LTD, Beijing, China) according to the manufacturer's recommendation and DNA quality control was performed using agarose gel electrophoresis and the Qubit dsDNA HS Assay Kit. DNA libraries and DNA nanoballs (DNB) were constructed on BGISP-100 platform (WuHan MGI Tech Co.,Ltd, Wuhan, China) with input of 150ng DNA. 50 bp single-end reads were generated with BGISEQ-50 sequencer (WuHan MGI Tech Co.,Ltd, Wuhan, China) and assembled with SPAdes v3.12 (Bankevich et al. 2012). All sequence reads are available under accession PRJEB39490 and PRJNA483379 of the ENA short read archive.

Serotyping of V. parahaemolyticus strains

Serotyping of the 57 in-house sequenced strains was performed by slide agglutination according to the manufacturer's instructions, using the 65 type-K serum set and 11 type-O serum set (Denka Seiken CO., LTD, Tokyo, Japan).

Development Kaptive database

I. Identification of O-loci

V. parahaemolyticus is divided into 16 O-serotypes (M. Chen et al. 2012; Guo et al. 2017). The O-locus has been defined between *dgkA* and *gmhD* genes, which encode a diacylglycerol-kinase and an epimerase respectively (Guo et al. 2017). With BlastN (minimum coverage of 80% and minimum

identity of 80%), the flanking genes of the O-region, *dgkA* and *gmhD*, were searched in all 1602 genome sequences and the O-loci sequences were extracted from the genomes manually.

II. Identification of K-loci

V. parahaemolyticus is divided in 71 K-serotypes (Guo et al. 2017). The K-locus has been defined between *gmhD* and *rjg* or *ugd* or *gtab* genes, which encode an epimerase and metallo-hydrolase, respectively (Y. Chen et al. 2010). With BlastN (minimum coverage of 80% and minimum identity of 80%), the flanking genes of the K-region, *gmhD* and *rjg*, were searched in all 1602 genome sequences. In some cases where the *rjg* gene was not found, gene *ugd* or gene *gtab* was selected as flanking gene. The K-loci sequences were extracted from the whole genome sequences manually.

III. Development of the *Kaptive* databases

For the development of the *Kaptive* databases, the set of 775 genomes from Genbank and 57 in-house sequenced genomes were used, and PATRIC genomes were excluded, since the PATRIC database did not contain metadata with the serotypes of the genomes. The extracted O- and K-loci sequences were annotated using Prokka (Seemann 2014). For annotating of the O-loci, a custom database was used which was built with the annotations of the previous described O1-O16 loci of *V. parahaemolyticus* (M. Chen et al. 2012; Guo et al. 2017), and K-loci sequences were annotated with the default Prokka database. Genes of the O- and K-loci respectively were clustered using Roary with an 80% cutoff on amino acid identity (Page et al. 2015). The Roary gene presence absence table was used to create gene presence-absence clusters of the O- and K-loci respectively. Of the total set of 832 genomes from Genbank (n=775) and in-house sequenced genomes (n=57), 449 genomes had a known O-serotype and 291 genomes had a known K-serotype. The known serotypes were linked to unique gene presence-absence clusters, and the clusters that contained only genomes with the same specific serotypes were selected. Of each unique cluster, one genome was selected as reference and the

nucleotide sequences of the O- or K-locus of the selected genomes were added to the *Kaptive* database file and curated manually. The selected genomes that were used as reference for each O- and K-serotype are shown in Supplemental Table 1. The *Kaptive* databases were tested with all 832 *V. parahaemolyticus* genomes with default settings and with a minimum gene coverage of and a minimum gene identity of 90%.

After the development of the *Kaptive* databases with the genomes with the known serotypes, the set with 770 PATRIC genomes were analyzed with the databases, and genomes with a novel K-locus were annotated and added to the *Kaptive* database as KUK (K-UnKnown).

The *V. parahaemolyticus* O-locus and K-locus *Kaptive* databases are available on the following Github page: https://github.com/aldertzomer/vibrio_parahaemolyticus_genomoserotyping.

Comparison of genes in O- and K-loci

Genes of the selected O- and K-loci of the *Kaptive* databases were clustered using Roary with 80% identity (Page et al. 2015). Phylogenetic trees were built from the Roary gene presence absence table using RAXML with a BINCAT model, doing 100 bootstrap searches, 20 ML searches and returning the best scoring ML tree. Schematic representations of the O- and K-loci were made with Geneious (Biomatters, Auckland, New Zealand). Gene cluster comparisons are made with Clinker (Gilchrist and Chooi 2020).

Results and Discussion

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

Gene patterns of each O-locus could be linked to a specific O-serotype for O1 to O16, except O3 and O13, since they share the same genes (M. Chen et al. 2012). Therefore O3 and O13 could not be distinguished based on gene presence and are assigned in the *Kaptive* O-database as “O3 or O13”. It is possible that a second cluster modifies the O-antigen, similar to what has been described for *Shigella* (Lehane, Korres, and Verma 2005). For O-serotype 4, two different gene patterns are found, and therefore, serotype O4 has two gene patterns in the database, assigned as “O4” and “O4-a”.

For the K-locus, several gene patterns could not be assigned to a specific K-serotype, because not all K-serotypes of the genomes were known. We nearly doubled the number of observed unique K-antigen encoding loci in the genomes. These unknown K-locus gene patterns (n=67) are assigned as KUK (**K**-serotype **UnKnown**) in the *Kaptive* database. Eight gene patterns of K-Untypable (KUT) strains previously described were found and these patterns are assigned in the *Kaptive* database as KUT followed by subsequent numbers 1-8. No isolate or genome was available for K-serotype 71, and this serotype is therefore not included in the database (or possibly assigned as KUK). For K-serotypes 20, 29, and 68, two different gene patterns were found, and therefore, these K-serotypes have two gene patterns included in the database, distinguished with the addition of “-a” to the K-serotype. For K-serotype 12, serotype K12-a is included in the database, which contain a similar gene pattern, but with different sequence identity.

Examination of the phylogenetic tree of the isolates from which the loci were extracts shows that the K-locus clearly clusters with bacterial clone and MLST type (Supplemental Figure 1), except for MLST ST3 strains, as they appear to be quite comparable but have different K- and O-loci.

Gene composition of O- and K-loci

Phylogenetic trees of the O-loci and K-loci based on gene presence/absence are shown in Figure 1, the representations of the gene content of each locus is shown in Supplemental Figure 2 and gene cluster comparisons are shown in Supplemental Figure 3.

For the O-loci, two main clusters were found, with O14 having an unique organization of genes in the O-locus compared to the other O-loci. The phylogenetic tree of the gene presence absences of the K-loci showed moderate variability of the gene content of the identified K-loci, with many genes in appearing in different organizations in the loci (Supplemental Figure 2), however, loci K3, K22 and K28 clustered separately from the other K-loci (Figure 1). In these three K-loci, a region with genes involved in sulfate assimilation (genes *cysN*, *cysD* and *cysC*) as well as genes encoding a type II secretion system (genes *xcpT*, *xpsE*, *epsF*, *gspK*, *xspD* and *bdbD*) are inserted between two *wecA* genes (Supplemental Figure 2). Interestingly, the loci for K15 and K17 appear to be nearly identical (Figure 1, Supplemental Figure 2).

Testing Kaptive databases on NCBI and in house sequences genomes

To develop the *Kaptive* databases, Roary gene clusters consisting of genomes with the same specific serotypes were selected to be included in the databases, and genomes in clusters consisting of different serotypes were excluded. To test the *Kaptive* databases, all 449 genomes with a known O-serotype and 291 genomes with a known K-serotype were used. Of the 449 known O-serotypes, 403 O-serotypes were correctly identified (Figure 2A, Supplemental Table 1) and 257 of the 291 known K-serotypes were correctly identified (Figure 2B, Supplemental Table 1). The mis-identified O- and K-serotype genomes were all genomes downloaded from NCBI. Since we only have the genomes and not the strains, we cannot check if these strains were serotyped correctly. The O-serotype O14 has been recently described (Guo et al. 2017) and only one genome with this O-serotype is available. In our dataset of 832 genomes, 14 NCBI genomes with serotype O5 are identified with the *Kaptive* database as the new serotype O14. It is very likely that these genomes are mis-serotyped, because the

serum for serotype O14 was not available when these genomes were serotyped. Low performance is also noted for K15 and K17, however it appears that these have identical gene content (Figure 2B, Supplemental Figure 2B) and they may be misclassified.

For several K-serotypes, only one genome sequence was available and we could not determine if there is variation in these K-loci gene content. Furthermore, several of the genomes sequenced in this study were sequenced with a 50-bp single-end BGI sequencer, resulting in a higher number of short contigs, therefore for several of these genomes, the K-locus was not assembled on one single contig. For these genomes, the contigs with flanking genes were selected and concatenated manually. It is possible that some genes are missing in these K-locus reference sequences, although these likely represent repeat sequences. If more *V. parahaemolyticus* genome sequences with closed K-locus become available, the *Kaptive* K-serotype database will be updated with the closed K-locus sequence of these K-serotypes.

Conclusions

The in this study developed *Kaptive* databases with the identified 16 O- and 70 K -loci can be used to identify the O- and K-serotypes of *V. parahaemolyticus* isolates from whole genome sequencing data. The variation of K-antigen loci is much higher than expected as we identified 67 new K-locus variants.

Funding

Sequencing of isolates was supported by the National Science and Technology Major Project of China (No. 2018ZX10305409-003).

References

- Albert, M. John, Dilara Islam, Shamsun Nahar, Firdausi Qadri, Susanna Falkind, and Andrej Weintraub. 1997. "Rapid Detection of *Vibrio Cholerae* O139 Bengal from Stool Specimens by PCR." *Journal of Clinical Microbiology* 35 (6): 1633–35.
- Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." *Journal of Computational Biology* 19 (5): 455–77. <https://doi.org/10.1089/cmb.2012.0021>.
- Chen, Min, Dan Guo, Hin Chung Wong, Xi Zhang, Fenxia Liu, Hongyou Chen, Miao Chen, et al. 2012. "Development of O-Serogroup Specific PCR Assay for Detection and Identification of *Vibrio Parahaemolyticus*." *International Journal of Food Microbiology*. <https://doi.org/10.1016/j.ijfoodmicro.2012.08.012>.
- Chen, Yuansha, Jianli Dai, J. Glenn Morris, and Judith A. Johnson. 2010. "Genetic Analysis of the Capsule Polysaccharide (K Antigen) and Exopolysaccharide Genes in Pandemic *Vibrio Parahaemolyticus* O3:K6." *BMC Microbiology* 10 (1): 274. <https://doi.org/10.1186/1471-2180-10-274>.
- Gavilan, Ronnie G., Maria L. Zamudio, and Jaime Martinez-Urtaza. 2013. "Molecular Epidemiology and Genetic Variation of Pathogenic *Vibrio Parahaemolyticus* in Peru." *PLoS Neglected Tropical Diseases* 7 (5). <https://doi.org/10.1371/journal.pntd.0002210>.
- Gil, Ana I., Hernan Miranda, Claudio F. Lanata, Ana Prada, Eric R. Hall, Carmen M. Barreno, Suraia Nusrin, Nurul A. Bhuiyan, David A. Sack, and Gopinath Balakrish Nair. 2007. "O3:K6 Serotype of *Vibrio Parahaemolyticus* Identical to the Global Pandemic Clone Associated with Diarrhea in Peru." *International Journal of Infectious Diseases* 11 (4): 324–28. <https://doi.org/10.1016/j.ijid.2006.08.003>.
- Gilchrist, Cameron L M, and Yit-Heng Chooi. 2020. "Clinker & Clustermap.js: Automatic Generation of Gene Cluster Comparison Figures." *BioRxiv*, 2020.11.08.370650. <https://doi.org/10.1101/2020.11.08.370650>.
- Guo, Xi, Bin Liu, Min Chen, Yuanyuan Wang, Lu Wang, Hongyou Chen, Yao Wang, Lihong Tu, Xi Zhang, and Lu Feng. 2017. "Genetic and Serological Identification of Three *Vibrio Parahaemolyticus* Strains as Candidates for Novel Provisional O Serotypes." *International Journal of Food Microbiology*. <https://doi.org/10.1016/j.ijfoodmicro.2017.01.010>.
- Hashii, Noritaka, Seiichi Kondo, Takehiro Iguchi, Mitsuaki Nishibuchi, and Kazuhito Hisatsune. 2000. "Chemical and Serological Properties of Lipopolysaccharides from *Vibrio Parahaemolyticus* O-Untypeable Strains Isolated from Patients." *Microbiology and Immunology* 44 (4): 229–34. <https://doi.org/10.1111/j.1348-0421.2000.tb02488.x>.
- Hoshino, Katsuaki, Shinji Yamasaki, Asish K. Mukhopadhyay, Soumen Chakraborty, Arnab Basu, Sujit K. Bhattacharya, G. Balakrish Nair, Toshio Shimada, and Yoshifumi Takeda. 1998. "Development and Evaluation of a Multiplex PCR Assay for Rapid Detection of Toxigenic *Vibrio Cholerae* O1 and O139." *FEMS Immunology and Medical Microbiology* 20 (3): 201–7. [https://doi.org/10.1016/S0928-8244\(98\)00014-5](https://doi.org/10.1016/S0928-8244(98)00014-5).
- Jones, Jessica L., Catharina H.M. Lüdeke, John C. Bowers, Nancy Garrett, Markus Fischer, Michele B. Parsons, Cheryl A. Bopp, and Angelo DePaola. 2012. "Biochemical, Serological, and Virulence Characterization of Clinical and Oyster *Vibrio Parahaemolyticus* Isolates." *Journal of Clinical Microbiology* 50 (7): 2343–52. <https://doi.org/10.1128/JCM.00196-12>.
- Lehane, Adele M., Haralambos Korres, and Naresh K. Verma. 2005. "Bacteriophage-Encoded Glucosyltransferase GtrII of *Shigella Flexneri*: Membrane Topology and Identification of Critical Residues." *Biochemical Journal* 389 (1): 137–43. <https://doi.org/10.1042/BJ20050102>.
- Nair, G. Balakrish, Thandavarayan Ramamurthy, Sujit K. Bhattacharya, Basabjit Dutta, Yoshifumi Takeda, and David A. Sack. 2007. "Global Dissemination of *Vibrio Parahaemolyticus* Serotype O3:K6 and Its Serovariants." *Clinical Microbiology Reviews* 20 (1): 39–48. <https://doi.org/10.1128/CMR.00025-06>.
- Okura, Masatoshi, Ro Osawa, Akihiko Tokunaga, Masatomo Morita, Eiji Arakawa, and Haruo

- Watanabe. 2008. "Genetic Analyses of the Putative O and K Antigen Gene Clusters of Pandemic *Vibrio Parahaemolyticus*." *Microbiology and Immunology* 52 (5): 251–64.
<https://doi.org/10.1111/j.1348-0421.2008.00027.x>.
- Page, Andrew J., Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T.G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. 2015. "Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis." *Bioinformatics* 31 (22): 3691–93.
<https://doi.org/10.1093/bioinformatics/btv421>.
- Seemann, Torsten. 2014. "Prokka: Rapid Prokaryotic Genome Annotation." *Bioinformatics* 30 (14): 2068–69. <https://doi.org/10.1093/bioinformatics/btu153>.
- Wang, Quan, Xiaojuan Ruan, Dongmei Wei, Zhidong Hu, Lixia Wu, Ting Yu, Lu Feng, and Lei Wang. 2010. "Development of a Serogroup-Specific Multiplex PCR Assay to Detect a Set of *Escherichia Coli* Serogroups Based on the Identification of Their O-Antigen Gene Clusters." *Molecular and Cellular Probes* 24 (5): 286–90. <https://doi.org/10.1016/j.mcp.2010.06.002>.
- Wattam, Alice R., David Abraham, Oral Dalay, Terry L. Disz, Timothy Driscoll, Joseph L. Gabbard, Joseph J. Gillespie, et al. 2014. "PATRIC, the Bacterial Bioinformatics Database and Analysis Resource." *Nucleic Acids Research* 42 (D1): 581–91. <https://doi.org/10.1093/nar/gkt1099>.
- Wyres, Kelly L., Ryan R. Wick, Claire Gorrie, Adam Jenney, Rainer Follador, Nicholas R. Thomson, and Kathryn E. Holt. 2016. "Identification of *Klebsiella* Capsule Synthesis Loci from Whole Genome Data." *Microbial Genomics* 2 (12): e000102. <https://doi.org/10.1099/mgen.0.000102>.

Figures and Tables

Figure 1. Phylogenetic trees from the Roary gene presence absence table for (A) O-serotypes and (B) K-serotypes

Figure 2. Performance of the *Kaptive* database on genomes with known (A) O- and (B) K-serotypes.

In blue the percentage of correctly typed genomes, in red the percentage incorrectly typed genomes. The number in the bars represent the number of genomes. *: These are likely O14 serotype genomes incorrectly agglutination serotyped because O14 has recently been described. #: The genetic organization of K15 and K17 is identical. Likely these have been misclassified.

Supplemental Figure 1. Phylogenetic tree based on core genome analysis of O- and K-representative isolates. The O and K loci and MLST STs are given.

Supplemental Figure 2. Schematic representations of gene organization of the (A) O-loci and (B) K-loci.

Supplemental Figure 3. Gene cluster comparisons of the (A) O-loci and (B) K-loci.

Supplemental Table 1. Selected genomes used for the O- and K-serotype *Kaptive* databases.