

## Genomic characterization, phylogenetic analysis, and identification of virulence factors in *Aerococcus sanguinicola* and *Aerococcus urinae* strains isolated from infection episodes

Carkaci, Derya; Højholt, Katrine ; Nielsen, Xiaohui Chen ; Dargis, Rimtas; Rasmussen, Simon; Skovgaard, Ole; Fuursted, Kurt; Andersen, Paal Skytt; Stegger, Marc; Christensen, Jens Jørgen

*Published in:*  
Microbial Pathogenesis

*DOI:*  
[10.1016/j.micpath.2017.09.042](https://doi.org/10.1016/j.micpath.2017.09.042)

*Publication date:*  
2017

*Document Version*  
Peer reviewed version

### *Citation for published version (APA):*

Carkaci, D., Højholt, K., Nielsen, X. C., Dargis, R., Rasmussen, S., Skovgaard, O., Fuursted, K., Andersen, P. S., Stegger, M., & Christensen, J. J. (2017). Genomic characterization, phylogenetic analysis, and identification of virulence factors in *Aerococcus sanguinicola* and *Aerococcus urinae* strains isolated from infection episodes. *Microbial Pathogenesis*, 112, 327-340. <https://doi.org/10.1016/j.micpath.2017.09.042>

### **General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

### **Take down policy**

If you believe that this document breaches copyright please contact [rucforsk@ruc.dk](mailto:rucforsk@ruc.dk) providing details, and we will remove access to the work immediately and investigate your claim.

1 **TITLE**

2 Genomic Characterization, Phylogenetic Analysis, and Identification of Virulence Factors in *Aerococcus*  
3 *sanguinicola* and *Aerococcus urinae* Strains Isolated from Infection Episodes

4

5

6 **AUTHOR NAMES**

7 Derya Carkaci<sup>1,2,3,\*</sup>, Katrine Højholt<sup>1,4,\*</sup>, Xiaohui Chen Nielsen<sup>1</sup>, Rimtas Dargis<sup>1</sup>, Simon Rasmussen<sup>4</sup>, Ole Skovgaard<sup>2</sup>,  
8 Kurt Fuursted<sup>3</sup>, Paal Skytt Andersen<sup>3,5</sup>, Marc Stegger<sup>3</sup> & Jens Jørgen Christensen<sup>1,6</sup>.

9

10 \* Shared authorship.

11

12

13 Derya Carkaci (DC)

derya.carkaci@gmail.com

14 Katrine Højholt (KH)

katrine.hojholt@bioinformatics.dtu.dk

15 Xiaohui Chen Nielsen (XCN)

xcn@regionsjaelland.dk

16 Rimtas Dargis (RD)

rida@regionsjaelland.dk

17 Simon Rasmussen (SR)

simon@bioinformatics.dtu.dk

18 Ole Skovgaard (OS)

olesk@ruc.dk

19 Kurt Fuursted (KF)

kfu@ssi.dk

20 Paal Skytt Andersen (PSA)

psa@ssi.dk

21 Marc Stegger (MS)

mtg@ssi.dk

22 Jens Jørgen Christensen (JJC)

jejc@regionsjaelland.dk

23

24

25 **AFFILIATION**

26 <sup>1</sup> Department of Clinical Microbiology, Slagelse Hospital, Slagelse, Denmark.

27 <sup>2</sup> Department of Science and Environment, Roskilde University, Roskilde, Denmark.

28 <sup>3</sup> Department of Microbiology & Infection Control, Statens Serum Institut, Copenhagen, Denmark.

29 <sup>4</sup> Department of Bio and Health Informatics, Technical University of Denmark, Kongens Lyngby, Denmark

30 <sup>5</sup> Department of Veterinary Disease Biology, Faculty of Health and Medical Sciences, University of Copenhagen,  
31 Copenhagen, Denmark.

32 <sup>6</sup> Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark.

33

34

35 **CORRESPONDING AUTHOR INFORMATION**

36 Address correspondence to: Jens Jørgen Christensen, Department of Clinical Microbiology, Slagelse Hospital,  
37 Ingemannsvej 46, DK-4200 Slagelse, Denmark. Email: jejc@regionsjaelland.dk.

38

39 **HIGHLIGHTS**

40 Bacterial adhesion gene homologs were identified in *A. sanguinicola* (*htpB*, *fbpA*, *lmb*, and *ilpA*) and *A. urinae*  
41 (*htpB*, *lap*, *lmb*, *fbp54*, and *ilpA*) genomes.

42

43 Capsular polysaccharide (CPS) gene homologs were identified in *A. sanguinicola* (15 genes) and *A. urinae* (11-  
44 16 genes) strains, giving rise to one and five types of putative CPS loci, respectively.

45

46 Marked differences were observed within *A. urinae* 1984-2004 and 2010-2015 strains in regards to genome  
47 sizes, core-genomes, proteome conservations, and phylogenetic analysis.

48

ACCEPTED MANUSCRIPT

49 **ABSTRACT**

50 *Aerococcus sanguinicola* and *Aerococcus urinae* are emerging pathogens in clinical settings mostly being  
51 causative agents of urinary tract infections (UTIs), urogenic sepsis and more seldomly complicated infective  
52 endocarditis (IE). Limited knowledge exists concerning the pathogenicity of these two species. Eight clinical  
53 *A. sanguinicola* (isolated from 2009-2015) and 40 clinical *A. urinae* (isolated from 1984-2015) strains from  
54 episodes of UTIs, bacteremia, and IE were whole-genome sequenced (WGS) to analyze genomic diversity and  
55 characterization of virulence genes involved in the bacterial pathogenicity.

56 *A. sanguinicola* genome sizes were 2.06-2.12 Mb with a 47.4-47.6 % GC-contents, and 1,783-1,905 genes  
57 were predicted whereof 1,170 were core-genes. In case of *A. urinae* strains, the genome sizes were 1.93-  
58 2.44 Mb with 41.6-42.6 % GC-contents, and 1,708-2,256 genes of which 907 were core-genes.

59 Marked differences were observed within *A. urinae* strains with respect to the average genome sizes,  
60 number and sequence identity of core-genes, proteome conservations, phylogenetic analysis, and putative  
61 capsular polysaccharide (CPS) loci sequences. Strains of *A. sanguinicola* showed high degree of homology.  
62 Phylogenetic analyses showed the 40 *A. urinae* strains formed two clusters according to two time periods:  
63 1984-2004 strains and 2010-2015 strains.

64 Genes that were homologs to virulence genes associated with bacterial adhesion and antiphagocytosis were  
65 identified by aligning *A. sanguinicola* and *A. urinae* pan- and core-genes against Virulence Factors of Bacterial  
66 Pathogens (VFDB). Bacterial adherence associated gene homologs were present in genomes of *A.*  
67 *sanguinicola* (*htpB*, *fbpA*, *lmb*, and *ilpA*) and *A. urinae* (*htpB*, *lap*, *lmb*, *fbp54*, and *ilpA*). Fifteen and 11-16  
68 CPS gene homologs were identified in genomes of *A. sanguinicola* and *A. urinae* strains, respectively. Analysis  
69 of these genes identified one type of putative CPS locus within all *A. sanguinicola* strains. In *A. urinae*  
70 genomes, five different CPS loci types were identified with variations in CPS locus sizes, genetic content, and  
71 structural organization.

72 In conclusion, this is the first study dealing with WGS and comparative genomics of clinical *A. sanguinicola*  
73 and *A. urinae* strains from episodes of UTIs, bacteremia, and IE. Gene homologs associated with  
74 antiphagocytosis and bacterial adherence were identified and genetic variability was observed within *A.*  
75 *urinae* genomes. These findings contributes with important knowledge and basis for future molecular and  
76 experimental pathogenicity study of UTIs, bacteremia, and IE causing *A. sanguinicola* and *A. urinae* strains.

77

78 **KEYWORDS**

79 *Aerococcus sanguinicola*; *Aerococcus urinae*; Infective endocarditis; Urinary tract infections; Capsular  
80 Polysaccharide; Bacterial adherence.

81

## 82 1. INTRODUCTION

83 The genus *Aerococcus* was first described in 1953 and consists nowadays of eight species of  
84 which *Aerococcus viridans* for a long time was the only species within the genus [1,2].

85 *Aerococcus urinae* was isolated in 1984 from a urine sample from a patient with verified urinary tract  
86 infection (UTI). This strain was characterized in 1989 as an *Aerococcus*-like organism and reclassified into its  
87 own species designation in 1992 [3,4]. *Aerococcus sanguinicola* was isolated in 1999 from an infective  
88 endocarditis (IE) suspected patient and in 2001 designated into its own species [5]. Both species are  
89 associated with UTIs worldwide, especially in elderly patients with predisposing conditions [6,7].

90 The prevalence of *A. urinae* in urine samples vary from 0.25 % to 4 % [7,8]. Both species were isolated from  
91 blood of patients suffering from urogenic sepsis, in few cases from patients with complicating IE and  
92 casuistically isolated from other foci [9]. Recognition of both species may be limited by their fastidious  
93 growth, often requiring supplementation with CO<sub>2</sub> for optimal growth [6,10]. Aerococci share colony  
94 morphology with  $\alpha$ -hemolytic streptococci and have a microscopic appearance similar to staphylococci,  
95 which adds to the risk of misinterpretation and misidentification [9]. At present, very limited knowledge  
96 exists regarding the bacterial pathogenicity and virulence mechanisms that lead to and maintain infections.

97 In clinical microbiology laboratories, diagnosing *A. urinae* and *A. sanguinicola* infections have been  
98 challenging [9]. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF  
99 MS), however, identifies both species rapidly and accurately, allowing clinical laboratories to correctly  
100 identify strains with increasing frequency of detection [11,12]. The species identifications can also be  
101 achieved with analysis of the 16S rRNA gene sequence [13] or the 16S-23S rRNA Intergenic Spacer Region  
102 [14].

103 Bacterial adherence and invasion to host tissue and cells increases the bacterial pathogenicity in infectious  
104 diseases as UTI [15] and IE [16]. Several host cell surface molecules are involved in the adhesive process in  
105 other pathogenic species, including fibronectin-binding proteins of *Streptococcus pyogenes* (*fbp54*) [17] and  
106 *Listeria monocytogenes* (*fbpA*) [18], laminin-binding protein of *Streptococcus agalactiae* (*lmb*) [19], and the  
107 *Listeria* adhesion protein (*lap*) [20]. A study from Shannon *et al.* (2010) described for the first time biofilm  
108 formation and stimulated biofilm production of *A. urinae* during exposure to human plasma [21]. The same  
109 study showed activation and aggregation of human platelets by *A. urinae*. Similarly, Senneby *et al.* (2014)  
110 demonstrated biofilm production in *A. sanguinicola* strains [22].

111  
112 Expression of capsular polysaccharide (CPS) facilitates bacterial protection against host immune phagocytosis  
113 [23]. Within genus *Aerococcus*, CPS expression were reported in a variant of *A. viridans*, *A. viridans* var.  
114 *homari*, which is a lobster pathogen causing gaffkemia [24]. The same study group showed upregulated

115 expression of molecular heat shock protein 60 (Hsp60) in virulent *A. viridans* strains compared to an  
116 avirulent strain [25].

117 A study from Christensen *et al.* showed genetic heterogeneity within a group of *A. urinae* strains. Fourteen  
118 Danish strains from 1984 to 1994 constituted a homogeneous group compared to seven heterogeneous  
119 non-Danish strains from 1985 to 1995 using DNA hybridization and phenotypic analysis [26].

120 Application of WGS has drastically expanded the understanding of the microbial world. The availability of  
121 bacterial genome data enables comprehensive bacterial comparisons and provides a better understanding of  
122 genome structures, evolutionary diversity, pathogenicity, and antimicrobial resistance [27]. In order to  
123 obtain further understanding of the genetic context of genes and to have a suitable high quality reference  
124 strain for the comparative genomics, complete and closed genomes of six *Aerococcus* type strains were  
125 recently achieved [28].

126 No whole-genome comparisons and genomic characterizations of *A. urinae* and *A. sanguinicola* have  
127 previously been performed. The aim of this study was to investigate the genomes of 40 *A. urinae* and eight *A.*  
128 *sanguinicola* strains in order to gain insight into their pan- and core-genome content and to identify putative  
129 virulence mechanisms that may be associated with human disease. Moreover, we compared WGS data and  
130 inferred phylogenetic relationships of the 40 clinical *A. urinae* strains from two different time periods of  
131 1984-2004 and 2010-2015, to analyze if the genomic diversity may be specific for the time period of strain  
132 isolations and type of infections.

133

## 134 2. MATERIALS AND METHODS

### 135 2.1. Bacterial strain characteristics, identifications, DNA isolation, genome sequencing, and verification of 136 species identifications

#### 137 2.1.1. Bacterial strains and species level identifications.

138 Eight clinical *A. sanguinicola* strains were collected between 2009 and 2015. Four isolates from two patients  
139 (one urine and one blood isolate for each patient), two isolates from two patients (one urine and one blood  
140 isolate), and two urine isolates from one patient (Supplementary material A).

141 Forty clinical *A. urinae* strains were collected from 32 patients between 1984 and 2015, twenty of these  
142 strains from 1984-2004 and the remaining 20 strains from 2010-2015. Twenty-four strains were isolated  
143 from 24 individual patients: From urine samples of UTI verified patients ( $n = 9$ ), from positive blood cultures  
144 of patients with bacteremia ( $n = 9$ ) and with verified IE ( $n = 6$ ). Fourteen strains were isolated from seven  
145 patients, both from urine ( $n = 7$ ) and blood culture ( $n = 7$ ) of each patient (paired strains). Two strains were

146 isolated as a pair from one patient, one blood isolate and one post mortem heart valve sample  
147 (Supplementary material A).

148 All strains were received from departments of clinical microbiology in Denmark. Identification to the species  
149 level was accomplished using MALDI-TOF MS v4.0.0.1 (5627 reference entries) (Bruker Daltonics, Germany)  
150 with a score above 2.0 at the Department of Clinical Microbiology, Slagelse Hospital, Denmark. Clinical  
151 strains were stored at -80 °C in bovine broth with 10 % glycerol (SSI Diagnostica, Denmark) until use.

152 Type strains of *A. sanguinicola* CCUG 43001<sup>T</sup> and *A. urinae* CCUG 36881<sup>T</sup> were obtained from the Culture  
153 Collection, University of Göteborg (www.ccug.se) and used as reference strains for the comparative genomic  
154 analyses. *A. sanguinicola* CCUG 43001<sup>T</sup> (isolated in 2001) and *A. urinae* CCUG 36881<sup>T</sup> (isolated in 1984,  
155 characterized in 1989, and reclassified in 1992) were isolated from a positive blood culture from a patient  
156 having bacteremia and from urine sample of a patient having UTIs, respectively [28].

157 The bacterial species identification and strain characteristics were denominated in a three-part identifier,  
158 such as "Au-01-U13". The initial two letter refers to the species identification (As for *A. sanguinicola* and Au  
159 for *A. urinae*), followed by a strain specific number. The final three characters describe the source of  
160 isolation (blood (B), urine (U) or heart valve (H)), and the year of strain isolation. "Au-01-U13" is a strain of *A.*  
161 *urinae* from a positive urine sample which was isolated in 2013.

162 Numbering of the paired *A. sanguinicola* strains, pair no. 1) As-24-U13 & As-25-U14, 2) As-41-B14 & As-46-  
163 U14, and 3) As-55-B15 & As-56-U15. Numbering of the paired *A. urinae* strains, pair no. 1) Au-02-B96 & Au-  
164 03-U96, 2) Au-44-B14 & Au-47-U14, 3) Au-49-B14 & Au-50-U14, 4) Au-51-B15 & Au-52-U15, 5) Au-53-B14 &  
165 Au-54-U14, 6) Au-57-B15 & Au-58-U15, 7) Au-59-B15 & Au-60-U15, and 8) Au-18-B93 & Au-19-H93.

166 Genomes of *A. urinae* CCUG 36881<sup>T</sup> (CP014161), *A. urinae* ACS-120-V-Col10a (CP002512), and *A. urinae* AU3  
167 (LUKP00000000.1) strains were obtained from NCBI GenBank for comparative analyses. *A. urinae* CCUG  
168 36881<sup>T</sup> was isolated from a positive human urine of a UTI infected person in 1984. *A. urinae* ACS-120-V-  
169 Col10a was isolated from a human vagina sample in Belgium in 2007. *A. urinae* AU3 was isolated from the  
170 human blood of a patient with bacteremia in Sweden in 2010.

171

#### 172 2.1.2. DNA isolation and extraction.

173 Strains were maintained by no more than three-to-four serial overnight passages at 35-37 °C in ambient air  
174 with 5 % CO<sub>2</sub> enrichment on 5 % blood agar plates (SSI Diagnostica, Denmark). Extraction of genomic DNA  
175 was carried out at Department of Microbiology and Infection Control, Statens Serum Institut, Denmark using  
176 the DNeasy Blood & Tissue kit, as described by the manufacturer (Qiagen, Denmark). Extraction of genomic  
177 DNA and WGS of *A. sanguinicola* CCUG 43001<sup>T</sup> and *A. urinae* CCUG 36881<sup>T</sup> were described in Carkaci *et al.*  
178 [28].

179

180 *2.1.3. Genome sequencing and pre-processing of sequence data.*

181 Fragment libraries were constructed using the Nextera XT DNA Sample Preparation Kit (Illumina, USA)  
182 followed by 251-bp or 150-bp paired-end sequencing on MiSeq or NextSeq sequencers (Illumina, USA),  
183 respectively, according to manufacturer's instructions. The Illumina demultiplexing process removed adapter  
184 sequences.

185 Quality of sequence reads were validated using FastQC v0.11.2 [29] and filtered using PRINSEQ v0.20.4 [30].  
186 High-quality sequence reads were *de novo* assembled using SPAdes v3.6.0 [31] with default *k*-mer settings.  
187 Enabling of the "careful" option minimized errors during genome assembly followed by Quast v3.1 quality  
188 assessment of assemblies [32]. Sequence reads were preprocessed according to the following criteria; 1)  
189 minimum sequence quality Q20, 2) minimum read lengths of 35 bp, and 3) removal of low quality reads from  
190 the 5'-end (20 bp) and 3'-end (5 bp). Minimum scaffold length was set as 200 bp and scaffolds having mean  
191 assembly coverage lower than 5x were discarded. The sequence coverage was set to 50x.

192

193 *2.1.4. Verification of species identifications.*

194 The bacterial identities were post-sequencing verified using the 16S rRNA gene sequence. The 16S rRNA  
195 gene sequences of clinical strains were predicted using SpeciesFinder [33] and used for nucleotide BLAST  
196 [34] against NCBI GenBank. The identifications were evaluated using BLAST percent identities, differences  
197 between maximum score of best and second best taxon matches, and minimum E-values of 0.001.

198

199

## 200 **2.2. Pan- and core-genome characterizations**

201 *2.2.1. Genome annotations and identification of pan- and core-genomes.*

202 Pan- and core-genomes were defined using PAN-genome analysis based on FUNctional PROfiles, PanFunPro  
203 [35]. Genes were predicted and translated into amino acid sequences using Prodigal v2.5 [36]. Each protein  
204 sequence was scanned against three protein databases with InterProScan [37] in the following order; PfamA  
205 [38], TIGRFAM [39], and SUPERFAMILY [40] to identify functional protein domains. Genes translated into  
206 protein sequences with identical functional protein domains were categorized as belonging to the same  
207 protein family. Proteins without identified functional domains were clustered using CD-hit [41] according to  
208 at least 60 % amino acid identities. For each genome, a collection of the annotated genes and the CD-hit  
209 clustered sequences constituted the genome profiles, and the complete collection of genome profiles from  
210 all strains represented the pan-genome.



211 The number of predicted genes for each strain was visualized in a genome plot along with the fraction of  
212 genes with protein domains of annotated function, protein domains with unknown function, and with no  
213 functional protein domains identified.

214 Genes found to be present in all of the analyzed genomes were categorized as belonging to the core-  
215 genome using PanFunPro2apply of PanFunPro [35] and visualized in a genome plot. Each collection of  
216 translated core-gene sequences were clustered using CD-hit [41] to ensure homology according to at least  
217 60 % amino acid identities and 60 % coverage. Core-genes passing the clustering criteria were globally  
218 aligned in MUSCLE v3.8.425 [42] and translated core-genes with less than 30 % conserved amino acid sites  
219 were not taken into considerations as core-genes.

220

#### 221 2.2.2. Pan-genomic proteome comparison.

222 Genomic relationships of strains were analyzed using PanFunPro predicted pan-genes. These genes were  
223 used for construction of a presence-absence matrix of genes within all genomes using  
224 PanGenome2Abundance of PanFunPro [35]. Genomic clustering of strains were statistically analyzed using  
225 Pearson correlation of the matrix. The correlation was illustrated as a heatmap where the correlation  
226 coefficient was color assigned.

227

#### 228 2.2.3. Proteome conservations.

229 The level of proteome conservations within each species were analyzed by pairwise all-against-all  
230 comparisons of protein domain annotations. For each comparison, the absolute number of shared protein  
231 families out of the total number of protein families were shown and converted into percentages. The  
232 genomic relatedness of two proteomes were demonstrated as a color assigned matrix plot, and the darker  
233 coloring, the higher percent identities and the higher degree of proteome conservations.

234

235

### 236 2.3. Phylogenetic relationships

#### 237 2.3.1. Core-gene phylogeny.

238 The phylogenetic relationships of the clinical *A. urinae* strains were analyzed using common core-genes  
239 within all 40 clinical *A. urinae* genomes. The PanFunPro predicted and subsequent homology verified protein  
240 sequences, encoded by the core-genes, were concatenated and multiple sequence aligned using MUSCLE  
241 v.3.8.425 [42]. jModelTest v2.1.10 [43] predicted the *Le & Gascuel* amino acid substitution model as the  
242 best-fit substitution model for the core-tree construction. PhyML v3.1 [44] generated the maximum

243 likelihood phylogenetic tree and the tree robustness was evaluated using 100 bootstrap replicates. The tree  
244 was visualized in CLC bio's Genomics Workbench v9.0 ([www.qiagenbioinformatics.com](http://www.qiagenbioinformatics.com)).

245

#### 246 2.3.2. SNPs phylogeny.

247 The phylogenetic relationships of the 40 *A. urinae* strains were verified using single-nucleotide  
248 polymorphisms (SNPs). SNPs were determined using the CSI Phylogeny  
249 ([www.cge.cbs.dtu.dk/services/CSIPhylogeny](http://www.cge.cbs.dtu.dk/services/CSIPhylogeny)) [45] by mapping of raw sequence reads against a reference  
250 genome. Three phylogenetic trees were generated, either by using the *A. urinae* CCUG 36881<sup>T</sup> type strain  
251 (complete genome), the clinical *A. urinae* ACS-120V-Col10a (complete genome), or the clinical *A. urinae* AU3  
252 (draft genome) as reference genomes. Calling of SNPs and validations were performed according to default  
253 settings of CSI Phylogeny.

254 SNPs passing the quality thresholds were concatenated to SNP sequences. Phylogenetic trees were created  
255 using the jModelTest [43] which predicted *generalized time reversible* nucleotide substitution model, as the  
256 most suitable substitution model for the dataset. The maximum likelihood trees in was generated using  
257 PhyML v3.1 [44]. Robustness of tree topologies were evaluated using bootstrap replicates of 100 and  
258 visualized in CLC bio's Genomics Workbench v9.0.

259

260

#### 261 2.4. Comparison of pan- and core-genes with Virulence Factors of Bacterial Pathogens

262 PanFunPro predicted pan- and core-genes were translated into protein sequence and aligned against the  
263 protein dataset of Virulence Factors of Bacterial Pathogens (VFDB) [46] using BLASTP v2.2.31 [34]. The  
264 protein dataset, only composed of experimentally verified virulence factors, was downloaded May 27<sup>th</sup> 2016.

265 Translated pan- and core-genes with VFDB hit bitscore values higher than 90, E-values lower than 0.001 and  
266 BLASTP amino acid sequence identities higher than 30 % were included in the analysis. Pan-genes with  
267 multiple VFDB hits were manually curated using at least 30 % BLASTP amino acid identities between the  
268 query and subject sequence. The query sequences were the PfamA, TIGRFAM, and SUPERFAMILY annotated  
269 and CD-hit clustered translated genes. Subject sequences were VFDB virulence protein sequences. Only  
270 translated pan-gene homologs with the highest bitscore values against a translated VFDB virulence gene  
271 were taken into account. Core-genes with multiple VFDB hits were sorted using an in-house Perl script, in  
272 which only gene with the highest bitscore values were taken into account.

273 Grouping of *A. sanguinicola* and *A. urinae* putative virulence gene homologs were accomplished according to  
274 VFDB assigned functional keywords for an overall genomic characterization of putative virulence genes.

275

276

## 277 2.5. Bacterial Capsular Polysaccharide

278 2.5.1 Search for CPS gene homologs within genomes of *A. urinae* ACS-120-V-Col10a and *A. urinae* AU3.

279 CPS associated gene homologs were searched within the public available *A. urinae* ACS-120-V-Col10a and *A.*  
280 *urinae* AU3 genomes. These genomes were subjected to BLASTX analysis against CPS associated genes of  
281 VFDB [46]. The BLASTX analysis was performed in CLC bio's Genomics Workbench v9.0 using E-values of  
282 0.001, bitscore values higher than 90, and minimum amino acid sequence identities of 30 %. Genes with  
283 multiple VFDB CPS gene mappings were sorted by only taking the BLAST hit with the highest bitscore value.

284

285 2.5.2. Mapping of CPS gene homologs within assembled genomes for prediction of putative CPS loci.

286 All the identified CPS gene homologs were plotted against the assembled *A. sanguinicola* and *A. urinae*  
287 genomes according to gene positions. Genomic regions with high abundance of CPS associated gene  
288 homologs were extracted and identified as putative CPS loci.

289

290 2.5.3. CPS structural organization analysis.

291 Mapping of gene homologs to the same VFDB CPS gene homologs were color assigned with the same color  
292 and side-by-side visualized in Geneious v9.1.6 [47].

293 Protein sequences of the initial four *A. urinae* gene homologs of *cps4A*, *cap8A*, *cap8B*, and *cap8C*, which  
294 constituted the common CPS loci region were subjected to four global protein sequence alignments to  
295 determine sequence identities using the MUSCLE v.9.1.6 [42].

296 The common CPS regions were followed by regions of variable sizes and genetic contents, hence defined as  
297 the variable CPS region. Genes positioned within the variable CPS loci regions and without VFDB assigned  
298 CPS annotations were subjected to BLASTX analysis for functional characterizations against the non-  
299 redundant protein sequence database of NCBI [34]. Only BLAST hits with E-values lower than 0.001 were  
300 taken into considerations.

301

302

## 303 2.6. Heat shock protein 60

304 The PanFunPro predicted *A. sanguinicola* and *A. urinae* Hsp60 homolog protein sequences (541-542 amino  
305 acids), encoded by the *htpB* gene, were compared against the Hsp60 protein sequence of the virulent *A.*  
306 *viridans* var. *homari* (184 amino acid partial sequence, AAM88526.1) to calculate sequence identities. The  
307 comparisons were made using the protein BLAST implementation in CLC bio's Genomics Workbench v9.0.

308

309

## 310 2.7. Adhesion associated gene homologs and cell wall signaling and anchoring

311 The presence of signal peptides were predicted using SignalP v4.1 ([www.cbs.dtu.dk/services/SignalP/](http://www.cbs.dtu.dk/services/SignalP/)) [48]  
312 and PSORTb v3 ([www.psort.org/](http://www.psort.org/)) [49]. The presence of cell wall anchoring protein domains were predicted  
313 using the TMHMM Server v2.0 ([www.cbs.dtu.dk/services/TMHMM/](http://www.cbs.dtu.dk/services/TMHMM/)) [50].

314

315 This study was approved by the Danish Data Protection Agency (J.nr. 2012-41-0240).

316

## 317 3. RESULTS

### 318 3.1. Species verification by 16S rRNA gene sequence analysis and features of genomic sequence data

#### 319 3.1.1. Confirmation of species identifications.

320 Forty-eight Danish clinical strains of *A. sanguinicola* ( $n = 8$ ) and *A. urinae* ( $n = 40$ ) (Supplementary A) were  
321 subjected to whole-genome analysis and genomic characterizations, including the corresponding type  
322 strains.

323 Identification to the species level using MALDI-TOF MS (score above 2.0) were post-sequencing verified using  
324 BLASTN sequence analysis of the 16S rRNA gene sequence against NCBI GenBank.

325 More than 99 % sequence identities were observed between the clinical *A. sanguinicola* 16S rRNA gene  
326 sequence and the public available type strain *A. sanguinicola* CCUG 43001<sup>T</sup> (BLAST maximum alignment  
327 score 2,835-2,841), and between the clinical *A. urinae* strains and the public available type strain *A. urinae*  
328 CCUG 36881<sup>T</sup> (BLAST maximum alignment score 2,804-2,837). BLAST maximum alignment score value  
329 differences between the best and second best taxon matches were 316-366.

330

#### 331 3.1.2. Features of genomic sequence data.

332 The number of *de novo* assembled scaffolds ranged from 17-44 and 12-58 for the clinical *A. sanguinicola* and  
333 *A. urinae* strains, respectively (Table 1). Genome sizes of *A. sanguinicola* strains were between 2.06 Mb to  
334 2.12 Mb with GC-contents of 47.4-47.6 %. *A. urinae* genome sizes ranged from 1.93 Mb to 2.44 Mb with GC-  
335 contents of 41.6-42.6 %. The 1984-2004 and 2010-2015 strains had average genome sizes of 1,947,525 bp  
336 (range 1.93-2.01 Mb) and 2,032,841 bp (1.93-2.44 Mb), respectively, which corresponded to an average  
337 increase of 86,000 bp genetic material in the 2010-2015 strains.

338 The type strains of *A. sanguinicola* CCUG 43001<sup>T</sup> and *A. urinae* CCUG 36881<sup>T</sup> had genome sizes of 2.03 Mb  
339 (GC-content 47.6 %) and 1.97 Mb (GC-content 42.6 %), respectively (Table 1).

340 Genomes of all *A. sanguinicola* strains and the corresponding type strain consisted of 1,783-1,905 genes and  
341 1,708-2,256 genes were identified within the genomes of *A. urinae*. The genome annotations revealed a high  
342 proportion of genes which encoded proteins with known annotated functional protein domains (78-84 %),  
343 with protein domains of unknown function (7-8 %), and proteins without annotated protein domains (8-14  
344 %).

345

ACCEPTED MANUSCRIPT

346 **Table 1.** Clinical and genomic characteristics of all clinical and type strains belonging to the *A. sanguinicola* and *A.*  
 347 *urinae* species.

Characteristics	<i>A. sanguinicola</i> CCUG 43001 <sup>T</sup>	<i>A. sanguinicola</i> (all strains)	<i>A. urinae</i> CCUG 36881 <sup>T</sup>	<i>A. urinae</i> (all strains)	<i>A. urinae</i> 1984-2004	<i>A. urinae</i> 2010-2015
<b>Clinical feature</b>						
Strain category	Type strain	Clinical strains	Type strain	Clinical strains	Clinical strains	
Country of isolation	Denmark	Denmark	Denmark	Denmark	Denmark	
Year of isolation	1999 <sup>1</sup>	2009 to 2015	1984 <sup>2</sup>	1984 to 2015	1984 to 2004	2010 to 2015
Strains (patients)	1	8 (5)	1	40 (32)	20 (18)	20 (14)
Patient mean age yrs. (range)	-	75 (62-87)	-	73 (10-94)	74.8 (56-85)	70.7 (10-94)
Gender ratio Male:Female:Unknown	-	2:3:0	-	18:8:6	8:4:6	10:4:0
Source of isolation	Blood	Urine and blood	Human urine	Urine, blood and heart valve	Urine, blood and heart valve	Urine and blood
Type of infection	Sepsis	UTI and bacteremia	UTI	UTI, bacteremia, and IE	UTI, bacteremia, and IE	UTI and bacteremia
<b>Genomic feature</b>						
Genome size (Mb)	2.03	2.06-2.12	1.97	1.93-2.44	1.93-2.01	1.93-2.44
Average genome size (bp)	-	-	-	-	1,947,525	2,032,841
Scaffolds	1	17-44	1	12-58	26-40	12-58
GC-content (%)	47.6	47.4-47.6	42.6	41.6-42.6	42.4-42.6	41.6-42.5
Genes	1,783 <sup>3</sup> / 1,838 <sup>4</sup>	1,783-1,905 <sup>3</sup>	1,739 <sup>3</sup> / 1,801 <sup>4</sup>	1,708-2,256 <sup>3</sup>	1,725-1,800 <sup>3</sup>	1,708-2,256 <sup>3</sup>
Core-genes (amino acid percent identity)	-	1,170	-	907	1,191 (99.4-100 %)	1,011 (96.6-100 %)
Unique intra-period core-genes	-	-	-	-	204	24
Common core-genes (amino acid length)	-	-	-	-	987 (312,235 amino acids)	

348 UTI, Urinary tract infection.

349 IE, Infective endocarditis.

350 <sup>1</sup> Isolated in 1999 and characterized in 2001.

351 <sup>2</sup> Isolated in 1984, characterized in 1989, and reclassified in 1992.

352 <sup>3</sup> Number of genes according to genome annotation using the PanFunPro pipeline [35].

353 <sup>4</sup> Number of genes according to genome annotation using the NCBI Prokaryotic Genome Annotation Pipeline [51].

354

355

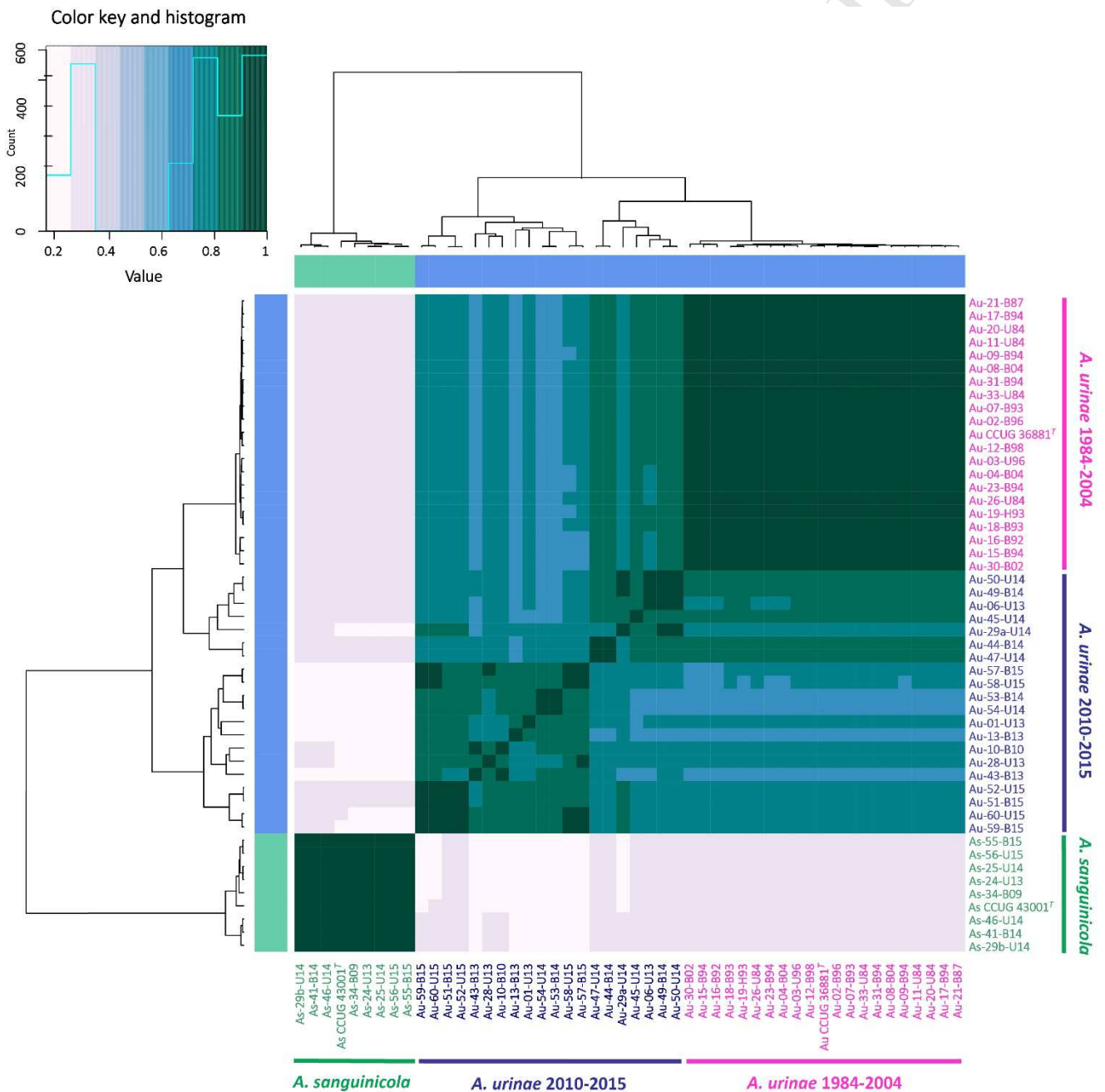
### 356 3.2. Pan- and core-genome characterizations, proteome conservations, and phylogeny

#### 357 3.2.1. Pan-genome analysis.

358 The total number of genes for all strains of *A. sanguinicola* were 16,678 genes and for strains of *A. urinae*  
 359 72,930 genes, including the type strains in both cases. The total number of genes for both species was  
 360 89,608 genes, of which 2,360 unique pan-genes. These genes were used to analyze the genomic relatedness  
 361 of all strains with a presence-absence analysis of the pan-genes across all strains (Figure 1).

362 Overall, high intra-species clustering was observed within both species and low clustering was observed  
 363 between both species (correlation coefficient below 0.4). The intra-species clustering was highest within  
 364 strains of *A. sanguinicola* (green, correlation coefficient 0.9-1) and within 1984-2004 isolated *A. urinae*  
 365 strains (pink, correlation coefficient 0.9-1). The 2010-2015 isolated *A. urinae* strains (blue) showed internal  
 366 heterogeneity (correlation coefficient 0.6-1). All the paired strains showed very high genomic clustering  
 367 (correlation coefficient 0.9-1).

368



369

370 **Figure 1.** Clustering of *A. sanguinicola* and *A. urinae* strains using Pearson correlation of the presence-absence  
371 matrix of the 2,360 unique pan-genes within both species. The highest correlation and genomic clustering was  
372 observed at correlation coefficient 1 (darkest coloring) and lowest at 0 (brightest coloring). Strains of *A.*  
373 *sanguinicola* showed high genomic clustering (green, correlation coefficient 0.9-1) and internal heterogeneity  
374 within *A. urinae* strains (blue and pink, correlation coefficient 0.6-1). The *A. urinae* 1984-2004 showed high  
375 genomic clustering (pink, correlation coefficient 0.9-1) and heterogeneity within the *A. urinae* 2010-2015 strains  
376 (blue, correlation coefficient 0.6-1). Low clustering was observed between the two species (correlation coefficient  
377 below 0.4). All the paired strains showed very high genomic clustering (correlation coefficient 0.9-1).

378

### 379 3.2.2 Core-genome analysis.

380 Highly conserved core-genomes were observed within both species as the number of core-genes decreased  
381 slightly as more genomes were added. The core-genomes reached a plateau stage through both species.

382 The number of PanFunPro predicted core-genes for clinical and the type strain of *A. sanguinicola* started  
383 from 1,359 core-genes and dropped to 1,260 core-genes when genomes of all *A. sanguinicola* strains were  
384 included. Core-gene homology was further verified using 60 % protein sequence identity across 60 %  
385 sequence coverage and more than 30 % sequence identities, which reduced the core-gene number to 1,170  
386 genes for *A. sanguinicola* strains (Table 1). In case of the clinical and the *A. urinae* type strain, the number of  
387 core-genes started from 1,314 genes and dropped to 1,023 genes when genomes of all *A. urinae* strains  
388 were included. Using the same homology verification criteria as in case of *A. sanguinicola* core-genes, the  
389 number was reduced to 907 core-genes (Table 1). Without the *A. urinae* type strain, the remaining 40 clinical  
390 *A. urinae* strains shared 987 core-genes (312,235 amino acids with overall 95.7-100 % amino acid identities).  
391 In case of the 1984-2004 and 2010-2015 *A. urinae* strains, the number of core-genes were determined as  
392 1,191 core-genes (99.4-100 % amino acid identity) and 1,011 core-genes (96.6-100 % amino acid identity),  
393 respectively. A total number of 204 core-genes were unique for only the 1984-2004 strains and 24 core-  
394 genes for the 2010-2015 strains.

395 The number of common core-genes, which fulfilled the homology verification criteria using 60 % sequence  
396 identities, were 81 genes for all *A. sanguinicola* and *A. urinae* strains.

397

### 398 3.2.3. *A. urinae* proteome conservations of 1984-2004 and 2010-2015 *A. urinae* strains.

399 Between 1,725-1,800 and 1,708-2,256 genes were predicted within the 1984-2004 and the 2010-2015  
400 strains, respectively (Table 1). These genes were evaluated and classified into 1,208 and 1,347 protein  
401 families for both species, respectively. Intra-period comparison of protein families showed high degree of  
402 proteome conservations as 96.4 to 99.7 % protein families were shared within the 1984-2004 strains  
403 (Supplementary material B). Higher proteome variations were observed within the 2010-2015 strains as



404 74.3-99.8 % of the protein families were shared. Inter-period comparison of the 1984-2004 and 2010-2015  
405 strains showed 74.7-87.8 % identities of shared protein families. Each of the paired strains exhibited 99.2-  
406 99.8 % identities.

407

408 *3.2.4. A. urinae phylogeny based on common core-genes and SNPs.*

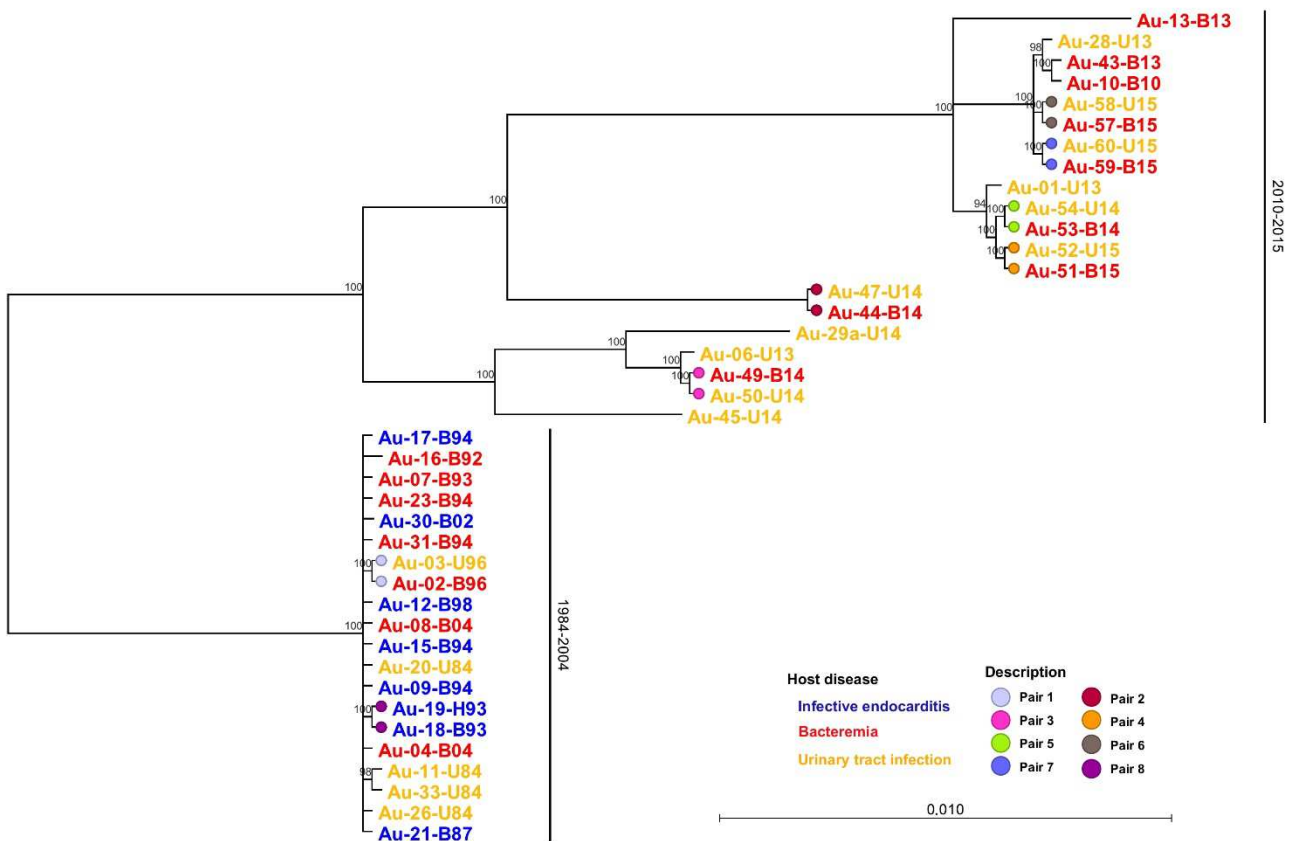
409 The 987 common core-genes within all 40 clinical *A. urinae* strains were used to demonstrate the  
410 phylogenetic relatedness (Figure 2). These 987 core-genes corresponded to 312,234 amino acids and with  
411 95.7-100 % sequence identities. Strains were color assigned according to type of infection: UTIs (yellow),  
412 bacteremia (red), and IE (blue). For the 1984-2004 and 2010-2015 strains, these 987 core-genes showed  
413 99.4-100 % and 96.6-100 % amino acid sequence identities, respectively.

414 The phylogenetic analysis showed no clustering related to the disease entity (UTIs, bacteremia, and IE). Two  
415 major clustering were observed, one consisting of the 1984-2004 strains and the second cluster consisted of  
416 the 2010-2015 strains, of which the main branch separating these two groups of strains was supported by  
417 bootstrap values of 100. Sub-clusterings were shown within the 2010-2015 cluster and also supported by  
418 bootstrap values of 100. Each of the eight paired *A. urinae* strains (marked with colored dots), from blood  
419 and urine samples from seven patients and from one blood and heart valve sample from one patient,  
420 clustered very close to each other and supported by bootstrap values of 100.

421 Identical clustering patterns of the 1984-2004 and 2010-2015 *A. urinae* strains were observed when SNPs  
422 were used to generate the phylogenetic relationships, showing two major clusters (Supplementary material  
423 C, Figure A, B, and C). Each of the paired *A. urinae* isolates were likewise clustered very close to each other.

424 When using the *A. urinae* CCUG 36881<sup>T</sup> genome (isolated in 1984) as a reference genome for SNP callings,  
425 20,694 SNPs were predicted and this reference strain clustered within the 1984-2004 cluster with strains  
426 from the same time period of isolation (Supplementary materials C, Figure A). *A. urinae* ACS-120-V-Col10a  
427 (isolated in 2007) and *A. urinae* AU3 (isolated in 2010) showed 22,608 SNPs and 21,302 SNPs, respectively,  
428 and clustered within the 2010-2015 cluster (Supplementary materials C, Figure B and C).

429



430  
 431 **Figure 2.** Core-genome phylogeny of the 40 clinical *A. urinae* strains based on the 987 translated common core-  
 432 genes (corresponding to 312,235 amino acids). The tree showed two major clustering of strains, one with the  
 433 1984-2004 strains and the other with strains from 2010-2015. Sub-clustering was observed within the 2010-2015  
 434 cluster. Strains were color assigned according to type of infections of UTIs (yellow), bacteremia (red), and IE  
 435 (blue). The last three characters of each strain identifier represented the source of strain isolation, blood (B),  
 436 urine (U) or heart valve (H) followed by the year of strain isolations. Branching of the maximum likelihood tree  
 437 was supported by bootstrap replicates of 100 and only bootstrap values higher than 90 were shown. Branch  
 438 lengths were given as substitutions per site. Clustering of the eight paired strains (marked with colored dots and  
 439 isolated from blood and urine samples of seven patients and blood and heart valve sample of one patient) were  
 440 very close to each other and supported by bootstrap values of 100.

441

442

### 443 3.3. Comparison of pan- and core-genes with Virulence Factors of Bacterial Pathogens

#### 444 3.3.1. Virulence gene homologs from the pan- and core-genomes.

445 The 16,678 pan-genes of *A. sanguinicola* and 72,930 pan-genes of *A. urinae* contained 12 and 20 VFDB  
 446 homolog virulence genes, respectively. Thirty-four out of 1,170 *A. sanguinicola* core-genes were identified as  
 447 VFDB homologs and similarly 24 genes out of 907 *A. urinae* core-genes. Only one common core-gene, which

448 encodes a HtpB protein (around 53-56 % protein sequence identities, Table 2), was predicted as a putative  
449 virulence gene of the 81 common core-genes of *A. urinae* and *A. sanguinicola* using at least 60 % protein  
450 sequence identities.

451 VFDB assigned keywords for functional characterization were used for an overall distribution of *A.*  
452 *sanguinicola* and *A. urinae* specific pan- and core-genes (Supplementary material D). The highest number of  
453 genes within one category was observed for genes associated with antiphagocytosis (15 genes in *A.*  
454 *sanguinicola* and between 11-16 genes in *A. urinae* strains). This was followed by genes associated with  
455 adherence (four genes in *A. sanguinicola* and five genes in *A. urinae*) and endotoxins (six genes in *A.*  
456 *sanguinicola* and five genes in *A. urinae*). Genes were also associated with intracellular growth/survival  
457 (three genes in *A. sanguinicola* and two genes in *A. urinae*) and stress proteins (four genes in *A. sanguinicola*  
458 and three genes in *A. urinae*). According to VFDB keywords, only strains of *A. sanguinicola* encoded gene  
459 homologs associated with biofilm formation (one gene) and beta-hemolysin/cytolysin (three genes). The  
460 miscellaneous group included genes related to iron and magnesium uptake/acquisition, surface protein  
461 anchoring, secretion system, regulation, and genes with uncharacterized function according to VFDB  
462 keyword designations (10 genes in *A. sanguinicola* and eight genes in *A. urinae*).

463 Antiphagocytosis, adherence, and biofilm formation associated proteins are known important virulence  
464 factors during bacterial infections. Translated pan- and core-gene homologs associated with these three  
465 virulence properties were selected for further characterizations. Each VFDB homolog pan- and core-gene is  
466 represented with protein sequence identities against the respective VFDB hit along with VFDB annotations  
467 and keyword designations (Table 2).

468

469 **Table 2.** *A. sanguinicola* and *A. urinae* virulence gene homologs of pan- and core-genes (protein level), involved in  
 470 antiphagocytosis, adherence, and biofilm formation.

Reference strain	VFDB annotation	VFDB gene	<i>A. sanguinicola</i> <sup>1</sup>	<i>A. urinae</i> <sup>2</sup>
			Sequence identity in % (n)	Sequence identity in % (n)
<b>VFDB category: Antiphagocytosis</b>				
<i>S. aureus</i> ssp. <i>aureus</i> MW2	CPS protein Cap8A	<i>cap8A</i>	34.3 (9)	30.4-32.0 (41)
	CPS protein Cap8B	<i>cap8B</i>	36.0-36.2 (9)	37.9-39.2 (41)
	CPS protein Cap8C	<i>cap8C</i>	-	43.6-45.6 (41) <sup>3a</sup>
	CPS protein Cap8D	<i>cap8D</i>	48.0-48.3 (9)	46.9-47.4 (24) & 63.7 (3) <sup>4a</sup>
	CPS protein Cap8F	<i>cap8F</i>	54.7 (9)	53.7-53.9 (22)
	CPS protein Cap8G	<i>cap8G</i>	50.8 (9)	50.8-51.9 (22)
	CPS protein Cap8N	<i>cap8N</i>	38.4 (9)	38.9-40.7 (27)
<i>S. pneumoniae</i> TIGR4	CPS protein Cps4A	<i>cps4A</i>	-	35.3-36.1 (40) & 33.3-42.9 (1) <sup>4b</sup>
	CPS protein Cps4E	<i>cps4E</i>	60.4 (9)	57.8-59.4 (23) & 57.3 (4) <sup>3b</sup>
	CPS protein Cps4F	<i>cps4F</i>	33.9-34.2 (9)	33.2-33.4 (22)
	CPS protein Cps4H	<i>cps4H</i>	-	30.6-31.4 (5)
	CPS protein Cps4I	<i>cps4I</i>	-	63.0 (2)
	CPS protein Cps4J	<i>cps4J</i>	70.6-70.9 (9)	70.6 (21) & 74.4 (1) <sup>4c</sup>
<i>E. faecalis</i> V583	Undecaprenyl diphosphate synthase	<i>cpsA</i>	49.8 (9)	51.4 (41)
	Phosphatidate cytidyltransferase	<i>cpsB</i>	41.7 (9)	42.2-42.9 (41)
	UDP-galactopyranose mutase	<i>cpsI</i>	-	60.5 (14)
<i>S. agalactiae</i> 2603V/R	Glycosyl transferase CpsE	<i>cpsE</i>	-	33.9 (12) & 58.5-71.4 (2) <sup>4d</sup>
	Glycosyl transferase CpsJ	<i>cpsJ</i>	34.9-35.3 (9)	-
	CPS protein CpsL	<i>cpsL</i>	-	32.7 (14)
	Glycosyl transferase CpsO	<i>cpsO</i>	45.7 (9)	-
	N-acetyl neuramic acid synthetase NeuB	<i>neuB</i>	-	39.8-40.4 (41)
<i>S. pyogenes</i> M1	UDP-glucose 6-dehydrogenase HasB	<i>hasB</i>	-	52.3 (24)
	UDP-glucose pyrophosphorylase HasC	<i>hasC</i>	66.2-66.6 (9)	50.7-51.9 (41)
<i>C. jejuni</i> ssp. <i>jejuni</i> NCTC 11168	UDP-glucose 6-dehydrogenase KfiD	<i>kfiD</i>	49.6-49.8 (9)	-
<b>VFDB category: Adherence</b>				
<i>L. pneumophila</i> ssp. <i>pneumophila</i> str. <i>Philadelphia 1</i>	Hsp60, 60K heat shock protein HtpB	<i>htpB</i>	56.1-56.3 (9)	53.6-54.0 (41)
<i>L. monocytogenes</i> EGD-e	Fibronectin-binding protein FbpA	<i>fbpA</i>	41.5-41.8 (9)	-
	<i>Listeria</i> adhesion protein LAP	<i>lap</i>	-	54.4-54.7 (41)
<i>S. agalactiae</i> 2603V/R	Laminin-binding surface protein Lmb	<i>lmb</i>	32.4 (9)	56.2-56.9 (41)
<i>S. pyogenes</i> M1	Fibronectin-binding protein Fbp54	<i>fbp54</i>	-	42.2-43.1 (41)
<i>V. vulnificus</i> YJ016	Immunogenic lipoprotein A IlpA	<i>ilpA</i>	38.0 (9)	38.1-39.2 (41)
<b>VFDB category: Biofilm formation</b>				
<i>E. faecalis</i> V583	Sugar-binding transcriptional regulator	<i>bopD</i>	31.8-32.2(9)	-

471 CPS, Capsular polysaccharide.

472 <sup>1</sup> *A. sanguinicola* strains: Eight clinical and one type strain.

473 <sup>2</sup> *A. urinae* strains: Forty clinical and one type strain.

474 <sup>3a/b</sup> Gene homologs of a) *cap8C* (Au-18-B93 and Au-19-H93) and b) *cps4E* (Au-02-B96, Au-03-U96, Au-12-B98, and  
 475 Au-15-B94) were predicted as shorter genes compared to the remaining *cap8C* and *cps4E* homolog genes of *A.*  
 476 *urinae* strains, respectively.

477 <sup>4a/b/c/d</sup> Gene homologs of a) *cap8D* (Au-06-U13, Au-49-B14, and Au-50-U14), b) *cps4A* (Au-06-U13), c) *cps4J* (Au-  
478 45-U14), and d) *cpsE* (Au-43-B13 and Au-10-B10) were predicted as two partial and shorter genes instead of one  
479 full length gene compared to the remaining *A. urinae* genes of the particular gene homolog.  
480

481 3.3.2. *Bacterial capsular polysaccharide gene homologs involved in evasion of immune phagocytosis.*

482 The CPS gene homologs as identified in *A. sanguinicola* and *A. urinae* strains were described in six bacterial  
483 species; *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Enterococcus faecalis*, *S. agalactiae*, *S. pyogenes*,  
484 and *Campylobacter jejuni* (Table 2). *A. sanguinicola* strains consisted of 15 CPS gene homologs and between  
485 11-16 CPS gene homologs were identified in *A. urinae* strains. The public available *A. urinae* ACS-120-V-  
486 Col10a and *A. urinae* AU3 consisted of 13 and 16 CPS gene homologs, respectively. The majority of the CPS  
487 gene homologs were described in *S. aureus* ssp. *aureus* MW2 (*cap8* genes) and *S. pneumoniae* TIGR4 (*cps4*  
488 genes). The highest percent identity was observed for the *S. pneumoniae* TIGR4 *cps4J* gene homolog with  
489 70.6-70.9 % for *A. sanguinicola* and 70.6-74.4 % for *A. urinae* strains.

490 Mapping of CPS gene homologs within the assembled genomes demonstrated regions with high abundance  
491 of CPS gene homologs in all the strains, whereof identified as putative CPS loci (Figure 3). These genes were  
492 positioned in the same orientation of translation and ordered behind each other with short distances to  
493 neighboring genes. Four CPS gene homologs of *A. sanguinicola* strains (*cpsA*, *cpsB*, *hasC*, and *kfiD*) and four  
494 of *A. urinae* strains (*cpsA*, *cpsB*, *neuB*, and *hasC*) were located outside of the putatively predicted CPS loci  
495 regions and presumable not involving in CPS.  
496

497 The CPS loci sizes were estimated between 12,800 to 19,500 bp, from positioning of CPS gene homologs  
498 until flanking by non-CPS associated genes. The number of genes within the CPS loci varied from 13 to 19  
499 genes, of which 7-12 genes were identified as CPS gene homologs. The genetic CPS loci arrangements  
500 showed one type of CPS loci for *A. sanguinicola* and five different types for *A. urinae* strains, the latter  
501 allocated into two major and three minor groups (Figure 3). Major group I was composed of all *A. urinae*  
502 strains from 1984-2004 and the *A. urinae* CCUG 36881<sup>T</sup> and major group II of 14 of the 20 strains from 2010-  
503 2015. The three minor groups were composed of one 2014 isolate (minor group I), two 2014 isolates (minor  
504 group II), and one 2013 and two 2014 isolates (minor group III). The *A. urinae* ACS-120-V-Col10a constituted  
505 a different CPS locus type and due to contig truncation the CPS locus of *A. urinae* AU3 was only partially  
506 identified.

507 Analysis of the CPS loci throughout all *A. sanguinicola* strains showed the initial two CPS gene homologs,  
508 *cap8A* (100 % protein sequence identity) and *cap8B* (99.9-100 %) to hold annotation of transcriptionally  
509 regulatory function. The remaining CPS gene homologs within the putative CPS loci showed higher than 97.9  
510 % protein sequence identities within all *A. sanguinicola* strains. In case of *A. urinae* strains, the initial four CPS

511 gene homolog were identified as transcriptionally regulator proteins in all strains and identified as the  
 512 common CPS region, *cps4A* (88.8-100 % protein sequence identity), *cap8A* (92.9-100 %), *cap8B* (94.9-100 %),  
 513 and *cap8C* (86.3-100 %). Higher protein identities were observed when the four common region CPS gene  
 514 were compared within strains of major group I and within major group II (Table 3).

515

516 **Table 3.** Sequence identities of the four translated CPS gene homologs constituting the common CPS region of all  
 517 *A. urinae* strains.

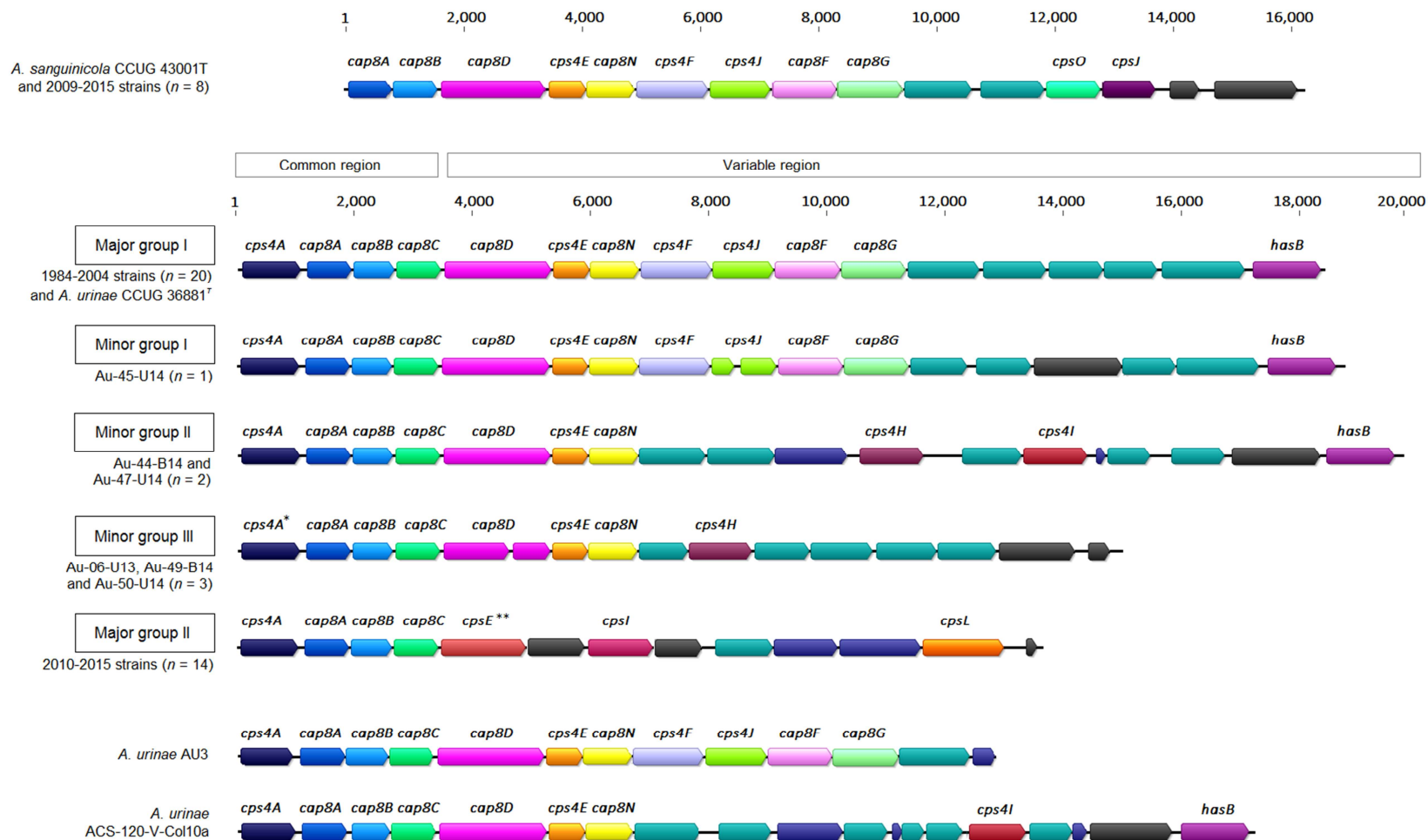
CPS loci	CPS loci common region			
	<i>cps4A</i>	<i>cap8A</i>	<i>cap8B</i>	<i>cap8C</i>
All <i>A. urinae</i> strains	88.8-100 %	92.9-100 %	94.9-100 %	86.3-100%
Major group I - <i>A. urinae</i> strains from 1984-2004 ( <i>n</i> = 20)	99.7-100 %	100 %	99.1-100 %	100 %
Major group II - <i>A. urinae</i> strains from 2010-2015 ( <i>n</i> = 14)	100 %	100 %	99.6-100 %	99.6-100 %




518

519 The common CPS loci region of *A. urinae* strains were followed by a variable region with variations in size,  
 520 number of genes and genetic arrangements. This region was consisting of CPS gene homologs and genes not  
 521 matching any of the CPS genes of the VFDB database. The latter genes were classified into three categories  
 522 by evaluation of the genome annotations and further characterizations using BLASTX against the NCBI  
 523 protein database. The three categories were consisting of I) CPS associated glycosyl transferases and  
 524 hypothetical glycosyl transferases; II) cell surface polysaccharide biosynthesis and CPS synthesis related  
 525 proteins; and III) hypothetical proteins and proteins with unknown function. The cell surface polysaccharide  
 526 biosynthesis and CPS synthesis related proteins were among others epimerases and dehydrogenases.  
 527 Similarly, the *A. sanguinicola* CPS loci gene homologs were annotated as cell surface polysaccharide  
 528 biosynthesis and CPS synthesis related proteins, glycosyl transferases, epimerases, and dehydrogenases.

529 The *hasB* gene homolog (UDP-glucose dehydrogenase) was positioned as the terminal CPS locus gene for all  
 530 1984-2004 strains (major group I), three 2014 strains (minor group I-II), the *A. urinae* CCUG 36881<sup>T</sup>, and the  
 531 *A. urinae* ACS-120-V-Col10a strains. Search for the *hasB* gene homolog within genomes of major group II and  
 532 minor group III strains showed no *hasB* gene homologs. A *hasB* gene homolog was also identified in the *A.*  
 533 *urinae* AU3 genome, although not positioned within the same CPS locus encoding contig.

534



-  Category I CPS associated glycosyl transferase and hypothetical glycosyl transferase
-  Category II Cell surface polysaccharide biosynthesis and CPS synthesis related proteins
-  Category III Hypothetical proteins and proteins with unknown function

	Total no. of CPS gene homologs in genomes	CPS locus size (bp)	No. of genes within CPS locus	No. of VFDB CPS gene homologs within CPS locus
<i>A. sanguinicola</i>	15	16,000	15	11
<i>A. urinae</i>	Major group I	16	18,300	12
	Major group II	11	13,500	7
	Minor group I	16	18,500	12
	Minor group II	14	19,500	10
	Minor group III	12	14,700	8
	<i>A. urinae</i> CCUG 36881 <sup>T</sup>	16	18,300	17
<i>A. urinae</i> AU3	16	12,800	13	11
<i>A. urinae</i> ACS-120-V-Col10a	13	17,100	19	9

536 **Figure 3.** Genomic organization of CPS loci of clinical and type strains of *A. sanguinicola* and *A. urinae*, including  
537 the public available *A. urinae* CCUG 36881<sup>T</sup>, *A. urinae* ACS-120-V-Col10a, and *A. urinae* AU3 strains. All *A.*  
538 *sanguinicola* strains were constituted of the same genomic organization of the putative predicted CPS loci. The 40  
539 *A. urinae* strains and *A. urinae* CCUG 36881<sup>T</sup> constituted five different CPS loci, grouped into two major and three  
540 minor groups. \* The Au-06-U13 *cps4A* gene homolog was predicted as two partial and shorter genes compared to  
541 the remaining *cps4A* gene homolog. \*\* The Au-10-B10 and Au-43-B13 *cpsE* gene homologs were predicted as two  
542 partial and shorter genes compared to the remaining *cpsE* gene homologs.

543

544 3.3.3. Bacterial gene homologs involved in adhesion to host cells and biofilm formation.

545 Six gene homologs related to bacterial adherence were identified in *A. sanguinicola* and *A. urinae* genomes  
546 (Table 2). Among these, four gene homologs were present in *A. sanguinicola* genomes and encoded the  
547 immunogenic lipoprotein A (IlpA), laminin-binding surface protein (Lmb), fibronectin-binding protein (FbpA),  
548 and the 60K heat shock protein (HtpB). The *A. urinae* strains were containing five gene homologs which  
549 encoded the fibronectin-binding protein (Fbp54), *Listeria* adhesion protein (LAP), and IlpA, Lmb, and HtpB as  
550 with *A. sanguinicola* strains. VFDB categorized *htpB* of *Legionella pneumophila* as a bacterial adhesion  
551 protein.

552 A signal peptide was only identified in IlpA and Lmb proteins of *A. sanguinicola* and *A. urinae* strains, and no  
553 LPXTG motif containing anchoring domains were predicted in any of the identified adhesion protein  
554 homologs.

555 Comparison of Hsp60 from the virulent *A. viridans* var. *homari* strain and the HtpB protein of *A. sanguinicola*  
556 and *A. urinae* strains showed between 79.4-82.0 % protein sequence identities.

557 According to VFDB, only *A. sanguinicola* strains contained a biofilm-associated transcriptional regulator *bopD*  
558 gene homolog.

559



560 **4. DISCUSSION**

561 In the present study, WGS of eight *A. sanguinicola* and 40 *A. urinae* strains were analyzed to characterize  
562 these genomes and to identify the potential virulence genes that cause bacterial pathogenicity.

563

564 *4.1. Genomic analysis.*

565 The varying number of pan- and core-genes are highly affected by the number of strains included, the  
566 degree of bacterial heterogeneity and the predefined cut-off thresholds for defining core-genes [52] as also  
567 illustrated for the strains from the two *Aerococcus* species examined in this study. The genetic pool of genes  
568 were lower for *A. sanguinicola* strains (16,678 genes) than for the *A. urinae* strains (72,930 genes), whereas  
569 the number of core-genes were higher for the *A. sanguinicola* strains (1,170 core-genes) than for strains of  
570 *A. urinae* strains (907 core-genes). All *A. sanguinicola* strains showed very close relationships taken into  
571 account of only being represented by one type strain and eight clinical strains from five patients. Marked  
572 differences were observed within all *A. urinae* strains, with respect to the average genome sizes, genomic  
573 clustering, number and sequence identity of core-genes, proteome conservations, phylogenetic analysis, and  
574 CPS loci sequences. The 20 *A. urinae* 1984-2004 strains, from 18 patients, were highly homogeneous  
575 compared to the 20 *A. urinae* 2010-2015 strains from 14 patients.

576 Evolution of bacteria is highly affected through genetic alternations during evolutionary processes which  
577 shapes the bacterial genomes. Homologous recombination, lateral gene transfer, as well as indel and SNP  
578 mutations are genetic events responsible for genomic diversity and shaping of bacterial populations [53,54].  
579 These events can give rise to selective advantages in a bacterial species such as increased bacterial  
580 pathogenicity and adaptation for a host environment under selection pressure. In our study, analysis of  
581 unique core-genes and the subsequent core-genome phylogeny showed high genomic conservations within  
582 the 1984-2004 *A. urinae* strains compared to 2010-2015 strains with internal diversity. These findings were  
583 interesting in the way that these strains were belonging to the same bacterial species and only being  
584 separated by a period of six years in the strain collections. In *A. urinae*, a selective pressure, that might have  
585 taken place after 2004, could potentially explain the presence of multiple sub-clusters within the short-time  
586 span isolated 2010-2015 strains (5 years) compared to the 1984-2004 strains (20 years). Both the host-  
587 pathogen interaction, selective pressure through the use of antibiotics, and competition between microbial  
588 pathogens are factors that adds to the selectivity of beneficial genetic variations within a population [55].  
589 Acquisition of genetic material could support an average gain of 86,000 bp in genomes of the 2010-2015  
590 strains compared to the 1984-2004 strains, potentially increasing the genetic and proteomic variation as  
591 shown in the study.

592 In comparison, high level of recombination and positive selection was observed within streptococcal core-  
593 genomes. Low degree of recombination was observed in *S. agalactiae* core-genomes compared to *S.*

594 *pyogenes* with high degree of core-genome recombination [56]. In *S. aureus*, low level of recombination was  
595 observed in the core-genomes even though being a highly pathogenic species [57]. Variations within the  
596 genomes could be dispersed across the entire genome or concentrated within specific core-genes with a  
597 selective advantages. In case of *S. aureus* genomes, recombination was often taking part in genes related to  
598 bacterial pathogenicity [57]. This kind of findings could suggest a bacterial fitness for survival and host  
599 adaptation, as suggested for *Clostridium perfringens* strains in an evolutionary lineage study [58].

600 Another aspect was if the genetic variability only were seen in Danish *A. urinae* isolates (local environmental  
601 pressure) of which we performed the SNPs based phylogenetic analysis. These showed the two foreign *A.*  
602 *urinae* isolates, one from Belgium in 2007 and one from Sweden in 2010, clustering with the Danish 2010-  
603 2015 isolated *A. urinae* strains. These findings may suggest that the genetic changes observed, within the  
604 recently isolated Danish *A. urinae* genomes, might be a result of a general evolutionary event. Similarly, a  
605 study from de Been *et al.* showed phylogenetic clustering of modern *Enterococcus faecium* with modern  
606 clinical isolates, by analyzing adaptive recombination events in terms of SNPs within core-genomes [59].  
607 Marvig *et al.* demonstrated within-host bacterial adaptation to changing host environments and  
608 accumulation of SNPs in favor for bacterial survival and fitness of *Pseudomonas aeruginosa* in patients with  
609 cystic fibrosis [60]. In the latter study, SNPs were localized within the regulatory part of the bacterial  
610 genomes and in pathoadaptive genes among others CPS genes, demonstrating how positive selection for  
611 mutations might have aimed in bacterial adaptation to its host [60].

612 A large number of UTI causing bacteria is often associated with urosepsis, in which the pathogenic strains  
613 gets access into the bloodstream. A mortality rate of 33 % was observed in hospitalized patients with cases  
614 of uncomplicated UTIs causing pathogenic *Escherichia coli*, leading to bacteremia [61]. The transition of a  
615 superficial site of infection to a deep site of infection is important in regards to which bacterial virulence  
616 mechanisms the UTI pathogens are taking advantages of. McNally *et al.* analyzed the genomic diversity of  
617 blood and urine isolates of *E. coli* from five patients with urosepsis, like we did in the current study with the  
618 eight paired *A. urinae* isolates. In four of the paired set of *E. coli* strains, the urine and blood isolates had the  
619 same sequence type, no variations were observed between each set of isolates, and only a minimal set of  
620 virulence genes were needed to establish bacteremia [62]. In the fifth *E. coli* urosepsis patient, two different  
621 *E. coli* sequence types were identified in the same urine sample and a third serotype was causing  
622 bacteremia. Based on results from McNally *et al.*, we were not expecting to observe genomic differences  
623 within each set of the paired *A. urinae* strains and results from the current study showed highly similar set of  
624 *A. urinae* isolates. This indicates that superficial site of infection causing *A. urinae* isolates (from urine) were  
625 the same isolate causing a deep site of infection within the bloodstream.

626

627 4.2. VFDB predicted putative virulence genes.

628 The current study attempted to characterize the clinical strains for the presence of virulence associated  
629 genes by comparison against a database collection of virulence factors, VFDB [46]. In this way, we only  
630 expected to identify already known virulence genes and factors as the VFDB database was consisting of. Until  
631 now, no UTI or IE associated virulence genes were characterized within genomes of *A. sanguinicola* and *A.*  
632 *urinae* strains.

633

634 4.2.1. Bacterial capsular polysaccharide genes.

635 Within genus *Aerococcus* expression of CPS has only been described in *A. viridans* var. *homari*, the causative  
636 agent of the lobster disease gaffkemia. The study were studying the relationship between bacterial virulence  
637 and CPS thickness in a virulent and avirulent *A. viridans* var. *homari* strain [24]. In our study, the majority of  
638 *A. sanguinicola* and *A. urinae* CPS gene homologs were described in genomes of *S. aureus* ssp. *aureus* MW2  
639 (*cap8* genes) and *S. pneumoniae* TIGR4 (*cps4* genes), which are two well-known CPS expressing bacterial  
640 species [63–65].

641 Skov Sørensen *et al.* investigated expression of CPS of *S. pneumoniae* and mitis group streptococci [66].  
642 Previously, it was assumed that CPS expression does not take place in commensal organisms as mitis group  
643 streptococci. Surprisingly, in a high number of the commensal mitis group streptococci, both the presence of  
644 CPS loci and subsequent CPS expression were observed [66]. Based on these results and identification of  
645 VFDB gene homologs associated with CPS, we were analyzing how these genes were dispersed within each  
646 of the *A. urinae* and *A. sanguinicola* genomes. Very surprisingly, we were identifying putative CPS loci in all  
647 the WGS genomes with high certainties of being a real CPS loci due to a number of findings. First, all *A.*  
648 *sanguinicola* and *A. urinae* CPS loci were divided into a highly common (regulatory part) and variable region  
649 (CPS biosynthesis) [67,68], as seen with CPS loci of *S. agalactiae* [69] and *S. pneumoniae* strains [66]. In *S.*  
650 *agalactiae* strains, the regulatory function of the common region was, among others, demonstrated with a  
651 functional knock-out mutation analysis in which the common region regulated CPS expression and its fine-  
652 tuning [70].

653 Secondly, CPS gene homologs of the variable region of *A. sanguinicola* and *A. urinae* CPS loci were encoding  
654 cell surface polysaccharide biosynthesis proteins as glycosyl transferases, epimerases, and dehydrogenases,  
655 which was in line with CPS genes of the variable region of streptococcal and staphylococcal CPS loci. Skov  
656 Sørensen *et al.* [66] and O’Riordan & Lee [71] described the structural organization of streptococcal and *S.*  
657 *aureus* CPS locus organization, which consisted of polymerases, epimerases, flippases, dehydrogenases, and  
658 sugar transferases such as glycosyl transferase.

659 Thirdly, *A. urinae* CPS loci showed structural variations with different CPS locus sizes, genetic content, and  
660 organization genetic. The observed genetic CPS loci diversity as five different CPS loci types, mainly  
661 separated the 1984-2004 *A. urinae* CPS loci from the highly diverse 2010-2015 *A. urinae* CPS loci. This type of  
662 structural complexity and organization of CPS genes were also shown within *S. pneumoniae* [68], *S. aureus*  
663 [71], and *Klebsiella* ssp. [72] CPS loci.

664

#### 665 4.2.2. Bacterial adherence.

666 In this study, the presence of core-genes that were homologs to genes linked to bacterial adherence of *A.*  
667 *sanguinicola* (*htpB*, *fbpA*, *lmb*, and *ilpA*) and *A. urinae* (*htpB*, *lap*, *lmb*, *fbp54*, and *ilpA*) indicates adhesion as  
668 an important virulence factor within strains causing UTIs, bacteremia, and IE.

669 These genes were homologs to FbpA of *L. monocytogenes* [73] and Fbp54 of *S. pyogenes* [17], Lmb of *S.*  
670 *agalactiae* [19], and IlpA of *Vibrio vulnificus* [74]. The importance of these genes have been demonstrated  
671 with reduced adhesion using mutants due to no expression of fibronectin-binding proteins (*L.*  
672 *monocytogenes* FbpA [73] and *S. pyogenes* Fbp54 [17]), poor adhesion to immobilized placental laminin and  
673 subsequent reduced invasiveness (*S. agalactiae* Lmb) [19,75], and decreased adhesion to intestinal cells and  
674 reduced mortality in mice models (*V. vulnificus* IlpA) [74,76].

675 The *Listeria* adhesion protein LAP is an essential adhesion factor [20,77], which has been demonstrated as a  
676 cell surface protein [78,79], and binds Hsp60 [80]. A *lap*-deficient *L. monocytogenes* showed reduced  
677 adherence and unable to translocate into intestinal cells [77,80]. Hsp60 associated cell adherence was also  
678 described for *Clostridium difficile* [81]. In genus *Aerococcus*, upregulated Hsp60 expression was previously  
679 described in *A. viridans* var. *homari* [25]. In the current study, both *Aerococcus* species were having a Hsp60  
680 encoding *htpB* gene homolog, whereas only a *lap* gene homolog in *A. urinae* strains. The presence of *lap*  
681 gene and *htpB* gene homologs within *A. urinae* genomes enhances the need for further enlightening of a  
682 putative bacterial adherence interaction between these two gene products.

683 In Gram-positive bacteria, a cell surface exposure of bacterial adhesion proteins can be achieved through a  
684 signal peptide sequence and a LPXTG containing cell wall anchoring protein domain [82]. A new class of  
685 anchorless and surface exposed Gram-positive proteins lacks the signal peptide and/or the LPXTG motif [82].  
686 In the current study, no adhesion associated gene homologs contained a LPXTG anchoring motif and only *A.*  
687 *sanguinicola* and *A. urinae* Lmb and IlpA homolog protein coding genes consisted of a signal peptide  
688 sequence, which was in line with the laminin-binding protein Lmb of *S. agalactiae* [19] and Lbp of *S.*  
689 *pyogenes* [83], and with the IlpA protein of *V. vulnificus* [74,76].

690 Neither the *A. sanguinicola* nor *A. urinae* gene homologs of fibronectin-binding proteins, the LAP protein, or  
691 the Hsp60 (HtpB) proteins contained a signal sequence nor the LPXTG motif. This was indeed in line with

692 other atypical and surface exposed adhesion proteins that binds fibronectin (FbpA of *L. monocytogenes* [73],  
693 FbpA of *Streptococcus gordonii* [84], and PavA of *S. pneumoniae* [85]), the *Listeria* adhesion protein LAP of *L.*  
694 *monocytogenes* [79], and heat shock proteins (Hsp60 of *Legionella pneumophila* [86] and *C. difficile* [81]).

#### 695 4.2.3. Biofilm formation.

696 Only *A. sanguinicola* strains contained a biofilm associated transcriptional regulator gene homolog (*bopD*)  
697 with low sequence identities. The *bopD* gene of *E. faecalis* is one out of four *bopABCD* genes associated with  
698 biofilm formation [87,88]. We find it questionable whether the *A. sanguinicola bopD* gene homolog is a  
699 biofilm associated gene or simply a transcriptional regulator gene, since the *bopABCD* locus also contains  
700 three other genes. As *in vitro* biofilm production previously was observed in *A. sanguinicola* [22] and *A.*  
701 *urinae* strains [21], the search for gene homologs associated with biofilm production may be a key step to  
702 increase the bacterial pathogenicity understanding.

703

704

#### 705 5. Future perspectives.

706 With the development of sequencing technologies and the presence of genomes from pathogenic bacteria, a  
707 broad range of analyses for a better understanding of bacterial pathogenicity are facilitated. More attention  
708 can be subjected to *A. sanguinicola* and *A. urinae* pathogenicity in order to further step into how these  
709 clinical strains may cause infections as UTIs, bacteremia, and IE.

710 Experimental animal models could be one way to analyze the current pathogenic status of recent 2010-2015  
711 *A. urinae* strains compared to 1984-2004 strains and how the bacterial pathogenicity and host adaptation  
712 may have evolved after the first time period of strain collections. Inclusion of more clinical strains, from even  
713 broader time periods, and from geographical different locations are needed to extend these analysis. This  
714 also in regards to demonstrate if CPS expression takes place, even though both species only were considered  
715 as low pathogenic. The functional meaning of gene homologs which were associated with bacterial adhesion  
716 needs to be verified and to reveal if the expressed gene products were bacterial cell surface exposed to  
717 maintain the adherence function.

718 Introduction of WGS in clinical laboratories will illuminate the fully genomic repertoire of these strains and  
719 enhance the clinical importance of these strains, including identification of the natural habitat of these  
720 bacterial species.

721

722 **6. CONCLUSIONS**

723 This is the first study dealing with comparative WGS analysis of clinical and type strain genomes of *A.*  
724 *sanguinicola* and *A. urinae*. High degree of genomic clustering was observed for strains of *A. sanguinicola*  
725 and marked differences within genomes of *A. urinae* strains with regards to the average genome sizes,  
726 number and sequence identity of core-genes, proteome conservations, genomic clustering, and phylogenetic  
727 analysis.

728 Gene homologs associated with antiphagocytosis and bacterial adherence were identified and putative CPS  
729 loci were identified within both species.

730 These findings contributes with novel genetic information of *A. sanguinicola* and *A. urinae* strains which  
731 provides an important basis for future understanding of UTIs, bacteremia, and IE pathogenicity caused by  
732 these two *Aerococcus* species.

733

**734 COMPETING INTERESTS**

735 The authors declare no competing interest.

736

**737 FUNDING**

738 The work was supported by the Region Zealand Research Unit (13-000835); The Foundation of Director  
739 Jacob Madsen and Spouse Olga Madsen (May 2014); The Foundation of Kurt Bønnelycke and Spouse Mrs  
740 Grethe Bønnelycke (10053030); The Council of Research in Region Zealand Næstved/Slagelse/Ringsted  
741 Hospitals (December 2013 and October 2014); The Foundation of Helge Peetz and Verner Peetz & Spouse  
742 Vilma Peetz (December 2013); The Region Zealand Foundation for Health Research (12-000095); The  
743 Common Research Foundation of Region Zealand and Roskilde University (November 2013); and The Danish  
744 Heart Foundation (15-R99-A6040-22951).

745

**746 ACKNOWLEDGMENTS**

747 Thanks to medical laboratory technicians at the Department of Clinical Microbiology, Slagelse Hospital,  
748 Denmark for collaboration during collection and identification of clinical strains. Thanks to Ulrik Stenz  
749 Justesen (Department of Clinical Microbiology, Odense University Hospital, Odense, Denmark) for providing  
750 clinical strains; and thanks to Elvira Chapka (Department of Microbiology & Infection Control, Statens Serum  
751 Institut, Denmark) for technical assistance in the WGS process.

752

**753 AUTHORS' CONTRIBUTIONS**

754 DC, XCN, and JJC designed the overall study. RD, PSA, MS, and Elvira Chapka contributed to the laboratory  
755 work and WGS process. XCN, JJC and SR guided the bioinformatic analysis and DC and KH performed the  
756 bioinformatic data analysis. DC wrote the manuscript and all authors contributed to the critical reading.

757

## 758 REFERENCES

- 759 [1] R.E.O. Williams, A. Hirsch, S.T. Cowan, *Aerococcus*, a New Bacterial Genus, *J. Gen. Microbiol.* 8 (1953) 475–480.  
760 doi:10.1099/00221287-8-3-475.
- 761 [2] M. Tohno, M. Kitahara, S. Matsuyama, K. Kimura, M. Ohkuma, K. Tajima, *Aerococcus vaginalis* sp. nov., isolated  
762 from the vaginal mucosa of a beef cow, and emended descriptions of *Aerococcus suis*, *Aerococcus viridans*,  
763 *Aerococcus urinae*, *Aerococcus urinaehominis*, *Aerococcus urinae*, *Aerococcus christensenii* and  
764 *Aerococcus sa*, *Int. J. Syst. Evol. Microbiol.* 64 (2014) 1229–1236. doi:10.1099/ijvs.0.058081-0.
- 765 [3] J.J. Christensen, B. Korner, H. Kjaergaard, *Aerococcus*-like organism - an unnoticed urinary tract pathogen,  
766 *APMIS.* 97 (1989) 539–546. doi:10.1111/j.1699-0463.1989.tb00828.x.
- 767 [4] M. Aguirre, M.D. Collins, Phylogenetic analysis of some *Aerococcus*-like organisms from urinary tract infections:  
768 description of *Aerococcus urinae* sp. nov., *J. Gen. Microbiol.* 138 (1992) 401–405. doi:10.1099/00221287-138-2-  
769 401.
- 770 [5] P.A. Lawson, E. Falsen, K. Truberg-Jensen, M.D. Collins, *Aerococcus sanguicola* sp. nov., isolated from a human  
771 clinical source, *Int. J. Syst. Evol. Microbiol.* 51 (2001) 475–479. doi:10.1099/00207713-51-2-475.
- 772 [6] J.J. Christensen, I.P. Jensen, J. Faerk, B. Kristensen, R. Skov, B. Korner, the D.A.S. Group, Bacteremia/Septicemia  
773 Due to *Aerococcus*-Like Organisms: Report of Seventeen Cases, *Clin. Infect. Dis.* 21 (1995) 943–947.  
774 doi:10.1093/clinids/21.4.943.
- 775 [7] M. Rasmussen, *Aerococci* and aerococcal infections, *J. Infect.* 66 (2013) 467–474.  
776 doi:10.1016/j.jinf.2012.12.006.
- 777 [8] Y.N. Guilarte, R. Tinguely, A. Lupo, A. Endimiani, Prevalence and characteristics of fluoroquinolone-resistant  
778 *Aerococcus urinae* isolates detected in Switzerland, *Int. J. Antimicrob. Agents.* 43 (2014) 474–483.  
779 doi:10.1016/j.ijantimicag.2014.01.022.
- 780 [9] M. Rasmussen, *Aerococcus*: an increasingly acknowledged human pathogen, *Clin. Microbiol. Infect.* 22 (2016)  
781 22–27. doi:10.1016/j.cmi.2015.09.026.
- 782 [10] J.J. Christensen, M. Kilian, V. Füssing, K. Andresen, J. Blom, B. Korner, A.G. Steigerwald, *Aerococcus urinae*:  
783 Polyphasic characterization of the species, *APMIS.* 113 (2005) 517–525. doi:10.1111/j.1600-  
784 0463.2005.apm\_183.x.
- 785 [11] E. Senneby, L. Göransson, S. Weiber, M. Rasmussen, A population-based study of aerococcal bacteraemia in the  
786 MALDI-TOF MS-era, *Eur. J. Clin. Microbiol. Infect. Dis.* 35 (2016) 755–762. doi:10.1007/s10096-016-2594-z.
- 787 [12] J.J. Christensen, R. Dargis, M. Hammer, U.S. Justesen, X.C. Nielsen, M. Kemp, T.D.M.-T.M.S. Group, Matrix-  
788 assisted laser desorption ionization-time of flight mass spectrometry analysis of gram-positive, catalase-  
789 negative cocci not belonging to the *Streptococcus* or *Enterococcus* genus and benefits of database extension, *J.*  
790 *Clin. Microbiol.* 50 (2012) 1787–1791. doi:10.1128/JCM.06339-11.
- 791 [13] O. Opota, G. Prod'homme, C. Andreutti-Zaugg, M. Dessauges, L. Merz, G. Greub, J.P. Chave, K. Jaton, Diagnosis of  
792 *Aerococcus urinae* infections: Importance of matrix-assisted laser desorption ionization time-of-flight mass  
793 spectrometry and broad-range 16S rDNA PCR, *Clin. Microbiol. Infect.* 22 (2016) e1–e2.  
794 doi:10.1016/j.cmi.2015.08.026.
- 795 [14] X.C. Nielsen, D. Carkaci, R. Dargis, L. Hannecke, J.U. Stenz, M. Kemp, M. Hammer, J.J. Christensen, 16S-23S  
796 Intergenic Spacer (ITS) Region Sequence Analysis: Applicability and Usefulness in Identifying Genera and Species  
797 Resembling Non-Hemolytic *Streptococci*, *Clin. Microbiol. Open Access.* 2 (2013) 1–8. doi:10.4172/2327-  
798 5073.1000130.
- 799 [15] A.L. Flores-Mireles, J.N. Walker, M. Caparon, S.J. Hultgren, Urinary tract infections: epidemiology, mechanisms  
800 of infection and treatment options, *Nat. Rev. Microbiol.* 13 (2015) 269–284. doi:10.1038/nrmicro3432.
- 801 [16] M.A. Kielhofner, R.J. Hamill, Role of Adherence in Infective Endocarditis, *Texas Hear. Inst. J.* 16 (1989) 239–249.



- 802 [17] H.S. Courtney, Y. Li, J.B. Dale, D.L. Hasty, Cloning, sequencing, and expression of a fibronectin/fibrinogen-  
803 binding protein from group A streptococci, *Infect. Immun.* 62 (1994) 3937–3946.
- 804 [18] A. Osanai, L. Sheng-Jun, K. Asano, H. Sashinami, D.L. Hu, A. Nakane, Fibronectin-binding protein, FbpA, is the  
805 adhesin responsible for pathogenesis of *Listeria monocytogenes* infection, *Microbiol. Immunol.* 57 (2013) 253–  
806 262. doi:10.1111/1348-0421.12030.
- 807 [19] B. Spellerberg, E. Rozdzinski, S. Martin, J. Weber-Heynemann, N. Schnitzler, R. Lütticken, A. Podbielski, Lmb, a  
808 protein with similarities to the Lral adhesin family, mediates attachment of *Streptococcus agalactiae* to human  
809 laminin, *Infect. Immun.* 67 (1999) 871–878.
- 810 [20] K.M. Burkholder, A.K. Bhunia, *Listeria monocytogenes* uses *Listeria* adhesion protein (LAP) to promote bacterial  
811 transepithelial translocation and induces expression of LAP receptor Hsp60, *Infect. Immun.* 78 (2010) 5062–  
812 5073. doi:10.1128/IAI.00516-10.
- 813 [21] O. Shannon, M. Mörgelin, M. Rasmussen, Platelet activation and biofilm formation by *Aerococcus urinae*, an  
814 endocarditis-causing pathogen, *Infect. Immun.* 78 (2010) 4268–4275. doi:10.1128/IAI.00469-10.
- 815 [22] E. Senneby, B. Eriksson, E. Fagerholm, M. Rasmussen, Bacteremia with *Aerococcus sanguinicola*: Case Series  
816 with Characterization of Virulence Properties, *Open Forum Infect. Dis.* 1 (2014) 1–6. doi:10.1093/ofid/ofu025.
- 817 [23] I.S. Roberts, The biochemistry and genetics of capsular polysaccharide production in bacteria, *Annu. Rev.*  
818 *Microbiol.* 50 (1996) 285–315. doi:10.1146/annurev.micro.50.1.285.
- 819 [24] K.F. Clark, D. Wadowska, S.J. Greenwood, *Aerococcus viridans* var. *homari*: The presence of capsule and the  
820 relationship to virulence in American lobster (*Homarus americanus*), *J. Invertebr. Pathol.* 133 (2016) 20–26.  
821 doi:10.1016/j.jip.2015.11.007.
- 822 [25] K.F. Clark, S.J. Greenwood, *Aerococcus viridans* expression of Cpn60 is associated with virulence during  
823 infection of the American lobster, *Homarus americanus* Milne Edwards, *J. Fish Dis.* 34 (2011) 831–843.  
824 doi:10.1111/j.1365-2761.2011.01300.x.
- 825 [26] J.J. Christensen, A.M. Whitney, L.M. Teixeira, A.G. Steigerwalt, R.R. Facklam, B. Korner, D.J. Brenner, *Aerococcus*  
826 *urinae*: Intraspecies Genetic and Phenotypic Relatedness, *Int. J. Syst. Bacteriol.* 47 (1997) 28–32.  
827 doi:10.1099/00207713-47-1-28.
- 828 [27] R.J. Olsen, S.W. Long, J.M. Musser, Bacterial Genomics in Infectious Disease and the Clinical Pathology  
829 Laboratory, *Arch. Pathol. Lab. Med.* 136 (2012) 1414–1422. doi:10.5858/arpa.2012-0025-RA.
- 830 [28] D. Carkaci, R. Dargis, X.C. Nielsen, O. Skovgaard, K. Fursted, J.J. Christensen, Complete Genome Sequences of  
831 *Aerococcus christensenii* CCUG 28831T, *Aerococcus sanguinicola* CCUG 43001T, *Aerococcus urinae* CCUG  
832 36881T, *Aerococcus urinaequi* CCUG 28094T, *Aerococcus urinaehominis* CCUG 42038 BT, and *Aerococcus*  
833 *viridans* CCUG 4311T, *Genome Announc.* 4 (2016) 1–2. doi:10.1128/genomeA.00302-16.
- 834 [29] S. Andrews, FastQC: a quality control tool for high throughput sequence data,  
835 www.bioinformatics.babraham.ac.uk/projects/fastqc, (2010).
- 836 [30] R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets, *Bioinformatics.* 27  
837 (2011) 863–864. doi:10.1093/bioinformatics/btr026.
- 838 [31] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham,  
839 A.D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev, P.A. Pevzner, SPAdes: A New  
840 Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing, *J. Comput. Biol.* 19 (2012) 455–477.  
841 doi:10.1089/cmb.2012.0021.
- 842 [32] A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies,  
843 *Bioinformatics.* 29 (2013) 1072–1075. doi:10.1093/bioinformatics/btt086.
- 844 [33] M. V Larsen, S. Cosentino, O. Lukjancenko, D. Saputra, S. Rasmussen, H. Hasman, T. Sicheritz-Pontén, F.M.  
845 Aarestrup, D.W. Ussery, O. Lund, Benchmarking of methods for genomic taxonomy, *J. Clin. Microbiol.* 52 (2014)

- 846 1529–1539. doi:10.1128/JCM.02981-13.
- 847 [34] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, *J. Mol. Biol.* 215  
848 (1990) 403–410. doi:10.1016/S0022-2836(05)80360-2.
- 849 [35] O. Lukjancenko, M.C. Thomsen, M.V. Larsen, D.W. Ussery, PanFunPro: PAN-genome analysis based on  
850 FUNctional PROfiles, *F1000Research*. 2 (2013) 1–13. doi:10.12688/f1000research.2-265.v1.
- 851 [36] D. Hyatt, G.-L. Chen, P.F. LoCascio, M.L. Land, F.W. Larimer, L.J. Hauser, Prodigal: prokaryotic gene recognition  
852 and translation initiation site identification, *BMC Bioinformatics*. 11 (2010) 1–11. doi:10.1186/1471-2105-11-  
853 119.
- 854 [37] E.M. Zdobnov, R. Apweiler, InterProScan - an integration platform for the signature-recognition methods in  
855 InterPro, *Bioinformatics*. 17 (2001) 847–848. doi:10.1093/bioinformatics/17.9.847.
- 856 [38] M. Punta, P.C. Coghill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements,  
857 A. Heger, L. Holm, E.L.L. Sonnhammer, S.R. Eddy, A. Bateman, R.D. Finn, The Pfam protein families databases,  
858 *Nucleic Acids Res.* 30 (2012) D290–D301. doi:10.1093/nar/gkr1065.
- 859 [39] D.H. Haft, J.D. Selengut, O. White, The TIGRFAMs database of protein families, *Nucleic Acids Res.* 31 (2003)  
860 371–373. doi:10.1093/nar/gkg128.
- 861 [40] D. Wilson, R. Pethica, Y. Zhou, C. Talbot, C. Vogel, M. Madera, C. Chothia, J. Gough, SUPERFAMILY -  
862 Sophisticated comparative genomics, data mining, visualization and phylogeny, *Nucleic Acids Res.* 37 (2009)  
863 D380–D386. doi:10.1093/nar/gkn762.
- 864 [41] W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide  
865 sequences, *Bioinformatics*. 22 (2006) 1658–1659. doi:10.1093/bioinformatics/btl158.
- 866 [42] R.C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.*  
867 32 (2004) 1792–1797. doi:10.1093/nar/gkh340.
- 868 [43] D. Darriba, G.L. Taboada, R. Doallo, D. Posada, jModelTest 2: more models, new heuristics and parallel  
869 computing, *Nat. Methods*. 9 (2012) 772. doi:10.1038/nmeth.2109.
- 870 [44] S. Guindon, O. Gascuel, A Simple, Fast, and Accurate Method to Estimate Large Phylogenies by Maximum  
871 Likelihood, *Syst. Biol.* 52 (2003) 696–704. doi:10.1080/10635150390235520.
- 872 [45] R.S. Kaas, P. Leekitcharoenphon, F.M. Aarestrup, O. Lund, Solving the problem of comparing whole bacterial  
873 genomes across different sequencing platforms, *PLoS One*. 9 (2014) 1–8. doi:10.1371/journal.pone.0104984.
- 874 [46] L. Chen, D. Zheng, B. Liu, J. Yang, Q. Jin, VFDB 2016: hierarchical and refined dataset for big data analysis-10  
875 years on., *Nucleic Acids Res.* 44 (2016) D694–D697. doi:10.1093/nar/gkv1239.
- 876 [47] M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C.  
877 Duran, T. Thierer, B. Ashton, P. Meintjes, A. Drummond, Geneious Basic: An integrated and extendable desktop  
878 software platform for the organization and analysis of sequence data, *Bioinformatics*. 28 (2012) 1647–1649.  
879 doi:10.1093/bioinformatics/bts199.
- 880 [48] T.N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: discriminating signal peptides from  
881 transmembrane regions, *Nat. Methods*. 8 (2011) 785–786. doi:10.1038/nmeth.1701.
- 882 [49] N.Y. Yu, J.R. Wagner, M.R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S.C. Sahinalp, M. Ester, L.J. Foster, F.S.L.  
883 Brinkman, PSORTb 3.0: Improved protein subcellular localization prediction with refined localization  
884 subcategories and predictive capabilities for all prokaryotes, *Bioinformatics*. 26 (2010) 1608–1615.  
885 doi:10.1093/bioinformatics/btq249.
- 886 [50] A. Krogh, B. Larsson, G. von Heijne, E.L.L. Sonnhammer, Predicting transmembrane protein topology with a  
887 hidden Markov model: application to complete genomes, *J. Mol. Biol.* 305 (2001) 567–580.  
888 doi:10.1006/jmbi.2000.4315.

- 889 [51] T. Tatusova, M. Dicuccio, A. Badretdin, V. Chetvernin, E.P. Nawrocki, L. Zaslavsky, A. Lomsadze, K.D. Pruitt, M.  
890 Borodovsky, J. Ostell, NCBI prokaryotic genome annotation pipeline, *Nucleic Acids Res.* 44 (2016) 6614–6624.  
891 doi:10.1093/nar/gkw569.
- 892 [52] D.W. Ussery, T.M. Wassenaar, S. Borini, *Computing for Comparative Microbial Genomics - Bioinformatics for*  
893 *Microbiologists - Chapter 12*, Springer-Verlag London, UK, 2009. doi:10.1007/978-1-84800-255-5\_2.
- 894 [53] E.J. Feil, Small change: keeping pace with microevolution, *Nat. Rev. Microbiol.* 2 (2004) 483–495.  
895 doi:10.1038/nrmicro904.
- 896 [54] B. Segerman, The genetic integrity of bacterial species: the core genome and the accessory genome, two  
897 different stories, *Front. Cell. Infect. Microbiol.* 2 (2012) 1–8. doi:10.3389/fcimb.2012.00116.
- 898 [55] K.A. Bliven, A.T. Maurelli, Evolution of Bacterial Pathogens within the Human Host, *Microbiol. Spectr.* 4 (2016)  
899 1–13. doi:10.1128/microbiolspec.
- 900 [56] T. Lefébure, M.J. Stanhope, Evolution of the core and pan-genome of *Streptococcus*: positive selection,  
901 recombination, and genome composition, *Genome Biol.* 8 (2007) R71.1-R71.17. doi:10.1186/gb-2007-8-5-r71.
- 902 [57] Y. Feng, C.J. Chen, L.H. Su, S. Hu, J. Yu, C.H. Chiu, Evolution and pathogenesis of *Staphylococcus aureus*: Lessons  
903 learned from genotyping and comparative genomics, *FEMS Microbiol. Rev.* 32 (2008) 23–37.  
904 doi:10.1111/j.1574-6976.2007.00086.x.
- 905 [58] A.P. Rooney, J.L. Swezey, R. Friedman, D.W. Hecht, C.W. Maddox, Analysis of core housekeeping and virulence  
906 genes reveals cryptic lineages of *Clostridium perfringens* that are associated with distinct disease presentations,  
907 *Genetics.* 172 (2006) 2081–2092. doi:10.1534/genetics.105.054601.
- 908 [59] M. de Been, W. Van Schaik, L. Cheng, J. Corander, R.J. Willems, Recent recombination events in the core  
909 genome are associated with adaptive evolution in *Enterococcus faecium*, *Genome Biol. Evol.* 5 (2013) 1524–  
910 1535. doi:10.1093/gbe/evt111.
- 911 [60] R.L. Marvig, L.M. Sommer, S. Molin, H.K. Johansen, Convergent evolution and adaptation of *Pseudomonas*  
912 *aeruginosa* within patients with cystic fibrosis, *Nat. Genet.* 47 (2015) 57–65. doi:10.1038/ng.3148.
- 913 [61] S. Tal, V. Guller, S. Levi, R. Bardenstein, D. Berger, I. Gurevich, A. Gurevich, Profile and prognosis of febrile  
914 elderly patients with bacteremic urinary tract infection, *J. Infect.* 50 (2005) 296–305.  
915 doi:10.1016/j.jinf.2004.04.004.
- 916 [62] A. McNally, F. Alhashash, M. Collins, A. Alqasim, K. Paszckiewicz, V. Weston, M. Diggle, Genomic analysis of  
917 extra-intestinal pathogenic *Escherichia coli* urosepsis, *Clin. Microbiol. Infect.* 19 (2013) E328–E334.  
918 doi:10.1111/1469-0691.12202.
- 919 [63] T.T. Luong, C.Y. Lee, Overproduction of Type 8 Capsular Polysaccharide Augments *Staphylococcus aureus*  
920 Virulence Overproduction of Type 8 Capsular Polysaccharide Augments *Staphylococcus aureus* Virulence, *Infect.*  
921 *Immun.* 70 (2002) 3389–3395. doi:10.1128/IAI.70.7.3389–3395.2002.
- 922 [64] S. Sau, J. Sun, C.Y. Lee, Molecular characterization and transcriptional analysis of type 8 capsule genes in  
923 *Staphylococcus aureus*, *J. Bacteriol.* 179 (1997) 1614–1621. doi:10.1128/jb.179.5.1614-1621.1997.
- 924 [65] R. Jothi, S. Parthasarathy, K. Ganesan, Comparison of the Virulence Factors and Analysis of Hypothetical  
925 Sequences of the Strains TIGR4, D39, G54 and R6 of *Streptococcus Pneumoniae*, *J. Comput. Sci. Syst. Biol.* 1  
926 (2008) 103–118. doi:10.4172/jcsb.1000010.
- 927 [66] U.B. Skov Sørensen, K. Yao, Y. Yang, H. Tettelin, M. Kilian, Capsular Polysaccharide Expression in Commensal  
928 *Streptococcus* Species: Genetic and Antigenic Similarities to *Streptococcus pneumoniae*, *MBio.* 7 (2016) e01844-  
929 16. doi:10.1128/mBio.01844-16.
- 930 [67] W. Zhongsong, J.-R. Zhang, Bacterial Capsules - Chapter 3, in: *Mol. Med. Microbiol.*, 2nd ed., Elsevier Ltd, 2015:  
931 pp. 33–53. doi:http://dx.doi.org/10.1016/B978-0-12-397169-2.00003-2.

- 932 [68] S.D. Bentley, D.M. Aanensen, A. Mavroidi, D. Saunders, E. Rabinowitsch, M. Collins, K. Donohoe, D. Harris, L.  
 933 Murphy, M.A. Quail, G. Samuel, I.C. Skovsted, M.S. Kalltoft, B. Barrell, P.R. Reeves, J. Parkhill, B.G. Spratt, Genetic  
 934 analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes, *PLoS Genet.* 2 (2006) e31.  
 935 doi:10.1371/journal.pgen.0020031.
- 936 [69] D.O. Chaffin, S.B. Beres, H.H. Yim, C.E. Rubens, The Serotype of Type Ia and III Group B Streptococci Is  
 937 Determined by the Polymerase Gene within the Polycistronic Capsule Operon, *J. Bacteriol.* 182 (2000) 4466–  
 938 4477. doi:10.1128/JB.182.16.4466-4477.2000.
- 939 [70] C. Toniolo, E. Balducci, M.R. Romano, D. Proietti, I. Ferlenghi, G. Grandi, F. Berti, I.M. Ros, R. Janulczyk,  
 940 *Streptococcus agalactiae* capsule polymer length and attachment is determined by the proteins CpsABCD, *J.*  
 941 *Biol. Chem.* 290 (2015) 9521–9532. doi:10.1074/jbc.M114.631499.
- 942 [71] K. O’Riordan, J.C. Lee, *Staphylococcus aureus* Capsular Polysaccharides, *Clin. Microbiol. Rev.* 17 (2004) 218–234.  
 943 doi:10.1128/CMR.17.1.218–234.2004.
- 944 [72] Y.-J. Pan, T.-L. Lin, C.-T. Chen, Y.-Y. Chen, P.-F. Hsieh, C.-R. Hsu, M.-C. Wu, J.-T. Wang, Genetic analysis of  
 945 capsular polysaccharide synthesis gene clusters in 79 capsular types of *Klebsiella* spp., *Sci. Rep.* 5 (2015) 15573.  
 946 doi:10.1038/srep15573.
- 947 [73] S. Dramsi, F. Bourdichon, D. Cabanes, M. Lecuit, H. Fsihi, P. Cossart, FbpA, a novel multifunctional *Listeria*  
 948 *monocytogenes* virulence factor, *Mol. Microbiol.* 53 (2004) 639–649. doi:10.1111/j.1365-2958.2004.04138.x.
- 949 [74] K.J. Lee, N.Y. Lee, Y.S. Han, J. Kim, K.H. Lee, S.J. Park, Functional characterization of the IIPa protein of *Vibrio*  
 950 *vulnificus* as an adhesin and its role in bacterial pathogenesis, *Infect. Immun.* 78 (2010) 2408–2417.  
 951 doi:10.1128/IAI.01194-09.
- 952 [75] T. Tenenbaum, B. Spellerberg, R. Adam, M. Vogel, K.S. Kim, H. Schroten, *Streptococcus agalactiae* invasion of  
 953 human brain microvascular endothelial cells is promoted by the laminin-binding protein Lmb, *Microbes Infect.* 9  
 954 (2007) 714–720. doi:10.1016/j.micinf.2007.02.015.
- 955 [76] S.Y. Goo, Y.S. Han, W.H. Kim, K.-H. Lee, S.-J. Park, *Vibrio vulnificus* IIPa-induced cytokine production is mediated  
 956 by toll-like receptor 2, *J. Biol. Chem.* 282 (2007) 27647–27658. doi:10.1074/jbc.M701876200.
- 957 [77] Z.W. Jaradat, J.L. Wampler, A.K. Bhunia, A *Listeria* adhesion protein-deficient *Listeria monocytogenes* strain  
 958 shows reduced adhesion primarily to intestinal cell lines, *Med. Microbiol. Immunol.* 192 (2003) 85–91.  
 959 doi:10.1007/s00430-002-0150-1.
- 960 [78] B. Jagadeesan, O.K. Koo, K.P. Kim, K.M. Burkholder, K.K. Mishra, A. Aroonnu, A.K. Bhunia, LAP, an alcohol  
 961 acetaldehyde dehydrogenase enzyme in *Listeria*, promotes bacterial adhesion to enterocyte-like Caco-2 cells  
 962 only in pathogenic species, *Microbiology.* 156 (2010) 2782–2795. doi:10.1099/mic.0.036509-0.
- 963 [79] K.M. Burkholder, K.P. Kim, K.K. Mishra, S. Medina, B.K. Hahm, H. Kim, A.K. Bhunia, Expression of LAP, a SecA2-  
 964 dependent secretory protein, is induced under anaerobic environment, *Microbes Infect.* 11 (2009) 859–867.  
 965 doi:10.1016/j.micinf.2009.05.006.
- 966 [80] J.L. Wampler, K. Kim, Z. Jaradat, A.K. Bhunia, Heat Shock Protein 60 Acts as a Receptor for the *Listeria* Adhesion  
 967 Protein in Caco-2 Cells, *Infect. Immun.* 72 (2004) 931–936. doi:10.1128/IAI.72.2.931–936.2004.
- 968 [81] C. Hennequin, F. Porcheray, A.-J. Waligora-Dupriet, A. Collignon, M. Barc, P. Bourlioux, T. Karjalainen, GroEL  
 969 (Hsp60) of *Clostridium difficile* is involved in cell adherence, *Microbiology.* 147 (2001) 87–96.  
 970 doi:10.1099/00221287-147-1-87.
- 971 [82] G.S. Chhatwal, Anchorless Adhesins and Invasins of Gram-Positive Bacteria: a New Class of Virulence Factors,  
 972 *Trends Microbiol.* 10 (2002) 205–208. doi:10.1016/S0966-842X(02)02351-X.
- 973 [83] Y. Terao, S. Kawabata, E. Kunitomo, I. Nakagawa, S. Hamada, Novel laminin-binding protein of *Streptococcus*  
 974 *pyogenes*, Lbp, is involved in adhesion to epithelial cells, *Infect. Immun.* 70 (2002) 993–997.  
 975 doi:10.1128/IAI.70.2.993–997.2002.

- 976 [84] J. Christie, R. McNab, H.F. Jenkinson, Expression of fibronectin-binding protein FbpA modulates adhesion in  
977 *Streptococcus gordonii*, *Microbiology*. 148 (2002) 1615–1625. doi:10.1099/00221287-148-6-1615.
- 978 [85] A.R. Holmes, R. McNab, K.W. Millsap, M. Rohde, S. Hammerschmidt, J.L. Mawdsley, H.F. Jenkinson, The *pavA*  
979 gene of *Streptococcus pneumoniae* encodes a fibronectin-binding protein that is essential for virulence, *Mol.*  
980 *Microbiol.* 41 (2001) 1395–1408. doi:10.1046/j.1365-2958.2001.02610.x.
- 981 [86] R.A. Garduño, E. Garduño, P.S. Hoffman, Surface-Associated Hsp60 Chaperonin of *Legionella pneumophila*  
982 Mediates Invasion in a HeLa Cell Model, *Infect. Immun.* 66 (1998) 4602–4610.
- 983 [87] M. Hufnagel, S. Koch, R. Creti, L. Baldassarri, J. Huebner, A Putative Sugar-Binding Transcriptional Regulator in a  
984 Novel Gene Locus in *Enterococcus faecalis* Contributes to Production of Biofilm and Prolonged Bacteremia in  
985 Mice, *J. Infect. Dis.* 189 (2004) 420–430. doi:10.1086/381150.
- 986 [88] R. Creti, S. Koch, F. Fabretti, L. Baldassarri, J. Huebner, Enterococcal Colonization of the Gastro-Intestinal Tract:  
987 Role of Biofilm and Environmental Oligosaccharides, *BMC Microbiol.* 6 (2006) 1–8. doi:10.1186/1471-2180-6-  
988 60.
- 989