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

Link to article, DOI:
10.1016/j.micpath.2017.09.042

Publication date:
2017

Document Version
Peer reviewed version

TITLE

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

AUTHOR NAMES

Derya Carkaci[1,2,3,*], Katrine Højholt[1,4,*], Xiaohui Chen Nielsen[1], Rimtas Dargis[1], Simon Rasmussen[4], Ole Skovgaard[2], Kurt Fuursted[3], Paal Skytt Andersen[3,5], Marc Stegger[3] & Jens Jørgen Christensen[1,6].

* Shared authorship.

Derya Carkaci (DC)	derya.carkaci@gmail.com
Katrine Højholt (KH)	katrine.hojholt@bioinformatics.dtu.dk
Xiaohui Chen Nielsen (XCN)	xcn@regionsjaelland.dk
Rimtas Dargis (RD)	rida@regionsjaelland.dk
Simon Rasmussen (SR)	simon@bioinformatics.dtu.dk
Ole Skovgaard (OS)	olesk@ruc.dk
Kurt Fuursted (KF)	kfu@ssi.dk
Paal Skytt Andersen (PSA)	psa@ssi.dk
Marc Stegger (MS)	mtg@ssi.dk
Jens Jørgen Christensen (JJC)	jejc@regionsjaelland.dk

AFFILIATION

[1] Department of Clinical Microbiology, Slagelse Hospital, Slagelse, Denmark.
[2] Department of Science and Environment, Roskilde University, Roskilde, Denmark.
[3] Department of Microbiology & Infection Control, Statens Serum Institut, Copenhagen, Denmark.
[4] Department of Bio and Health Informatics, Technical University of Denmark, Kongens Lyngby, Denmark
[5] Department of Veterinary Disease Biology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
[6] Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark.

CORRESPONDING AUTHOR INFORMATION

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

HIGHLIGHTS

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

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

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

ABSTRACT

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

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

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

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

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

KEYWORDS

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

1. INTRODUCTION

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

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

The prevalence of *A. urinae* in urine samples vary from 0.25 % to 4 % [7,8]. Both species were isolated from blood of patients suffering from urogenic sepsis, in few cases from patients with complicating IE and casuistically isolated from other foci [9]. Recognition of both species may be limited by their fastidious growth, often requiring supplementation with CO_2 for optimal growth [6,10]. Aerococci share colony morphology with α-hemolytic streptococci and have a microscopic appearance similar to staphylococci, which adds to the risk of misinterpretation and misidentification [9]. At present, very limited knowledge exists regarding the bacterial pathogenicity and virulence mechanisms that lead to and maintain infections.

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

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

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

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

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

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

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

2. MATERIALS AND METHODS

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

2.1.1. Bacterial strains and species level identifications.

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

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

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

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

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

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

Numbering of the paired *A. sanguinicola* strains, pair no. 1) As-24-U13 & As-25-U14, 2) As-41-B14 & As-46-U14, and 3) As-55-B15 & As-56-U15. Numbering of the paired *A. urinae* strains, pair no. 1) Au-02-B96 & Au-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 & 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.

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

2.1.2. DNA isolation and extraction.

Strains were maintained by no more than three-to-four serial overnight passages at 35-37 °C in ambient air with 5 % CO_2 enrichment on 5 % blood agar plates (SSI Diagnostica, Denmark). Extraction of genomic DNA was carried out at Department of Microbiology and Infection Control, Statens Serum Institut, Denmark using the DNeasy Blood & Tissue kit, as described by the manufacturer (Qiagen, Denmark). Extraction of genomic DNA and WGS of *A. sanguinicola* CCUG 43001^T and *A. urinae* CCUG 36881^T were described in Carkaci *et al.* [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.

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

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

2.2.2. Pan-genomic proteome comparison.

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

2.2.3. Proteome conservations.

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

2.3. Phylogenetic relationships

2.3.1. Core-gene phylogeny.

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

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

2.3.2. SNPs phylogeny.

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

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

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

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

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

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

2.5. Bacterial Capsular Polysaccharide

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

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

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

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

2.5.3. CPS structural organization analysis.

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

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

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

2.6. Heat shock protein 60

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

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

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

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

3. RESULTS

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

3.1.1. Confirmation of species identifications.

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

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

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

3.1.2. Features of genomic sequence data.

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

The type strains of *A. sanguinicola* CCUG 43001^T and *A. urinae* CCUG 36881^T had genome sizes of 2.03 Mb (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

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

Characteristics	*A. sanguinicola* CCUG 43001^T	*A. sanguinicola* (all strains)	*A. urinae* CCUG 36881^T	*A. urinae* (all strains)	*A. urinae* 1984-2004	*A. urinae* 2010-2015
Clinical feature						
Strain category	Type strain	Clinical strains	Type strain	Clinical strains	Clinical strains	
Country of isolation	Denmark	Denmark	Denmark	Denmark	Denmark	
Year of isolation	1999 [1]	2009 to 2015	1984 [2]	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
Genomic feature						
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 [3] / 1,838 [4]	1,783-1,905 [3]	1,739 [3] / 1,801 [4]	1,708-2,256 [3]	1,725-1,800 [3]	1,708-2,256 [3]
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 [1] Isolated in 1999 and characterized in 2001.
351 [2] Isolated in 1984, characterized in 1989, and reclassified in 1992.
352 [3] Number of genes according to genome annotation using the PanFunPro pipeline [35].
353 [4] 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).

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

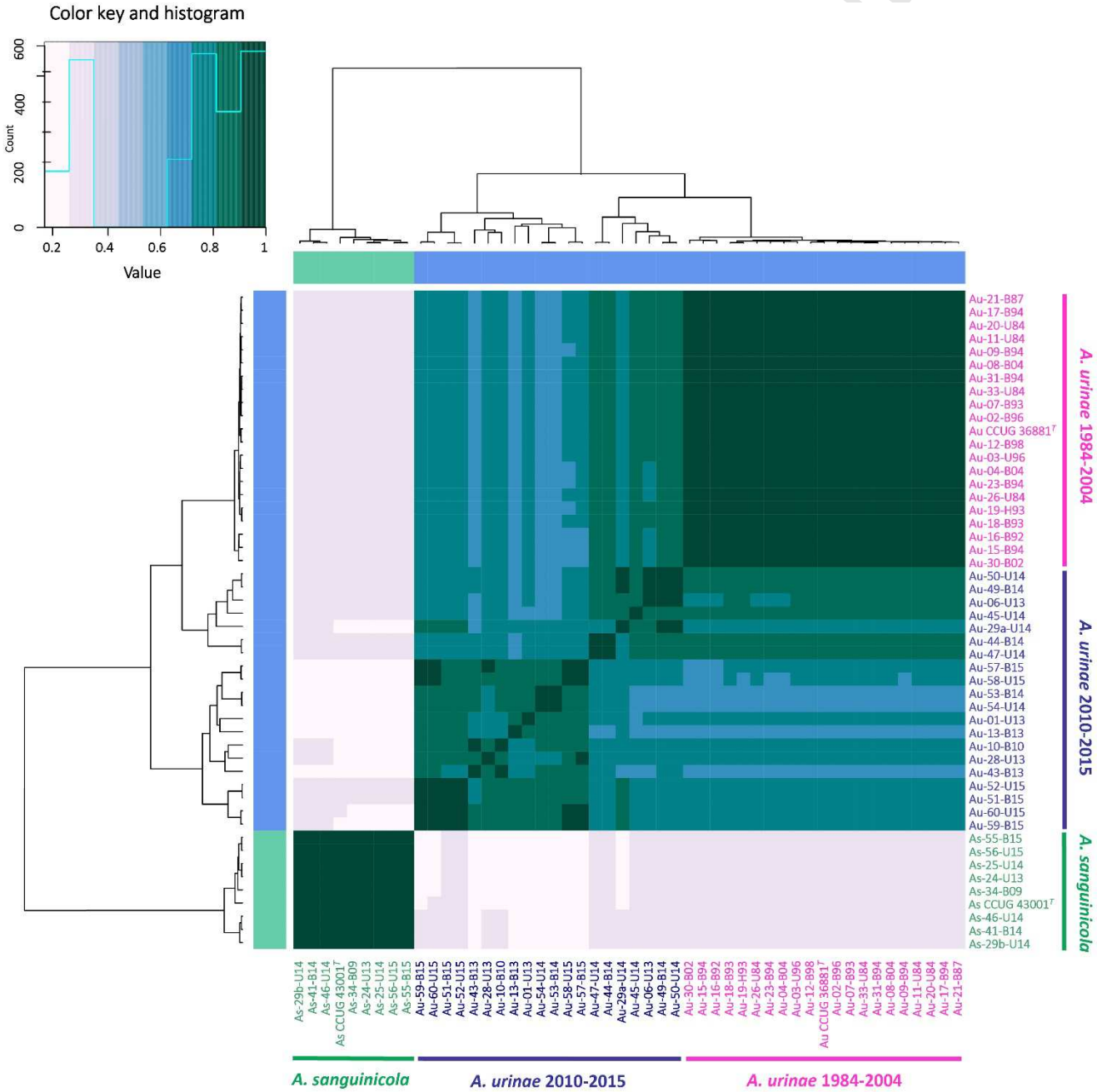


Figure 1. Clustering of *A. sanguinicola* and *A. urinae* strains using Pearson correlation of the presence-absence matrix of the 2,360 unique pan-genes within both species. The highest correlation and genomic clustering was observed at correlation coefficient 1 (darkest coloring) and lowest at 0 (brightest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

3.2.2 Core-genome analysis.

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

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

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

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

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

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

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

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

The phylogenetic analysis showed no clustering related to the disease entity (UTIs, bacteremia, and IE). Two major clustering were observed, one consisting of the 1984-2004 strains and the second cluster consisted of the 2010-2015 strains, of which the main branch separating these two groups of strains was supported by bootstrap values of 100. Sub-clusterings were shown within the 2010-2015 cluster and also supported by bootstrap values of 100. Each of the eight paired *A. urinae* strains (marked with colored dots), from blood and urine samples from seven patients and from one blood and heart valve sample from one patient, clustered very close to each other and supported by bootstrap values of 100.
Identical clustering patterns of the 1984-2004 and 2010-2015 *A. urinae* strains were observed when SNPs were used to generate the phylogenetic relationships, showing two major clusters (Supplementary material C, Figure A, B, and C). Each of the paired *A. urinae* isolates were likewise clustered very close to each other.

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

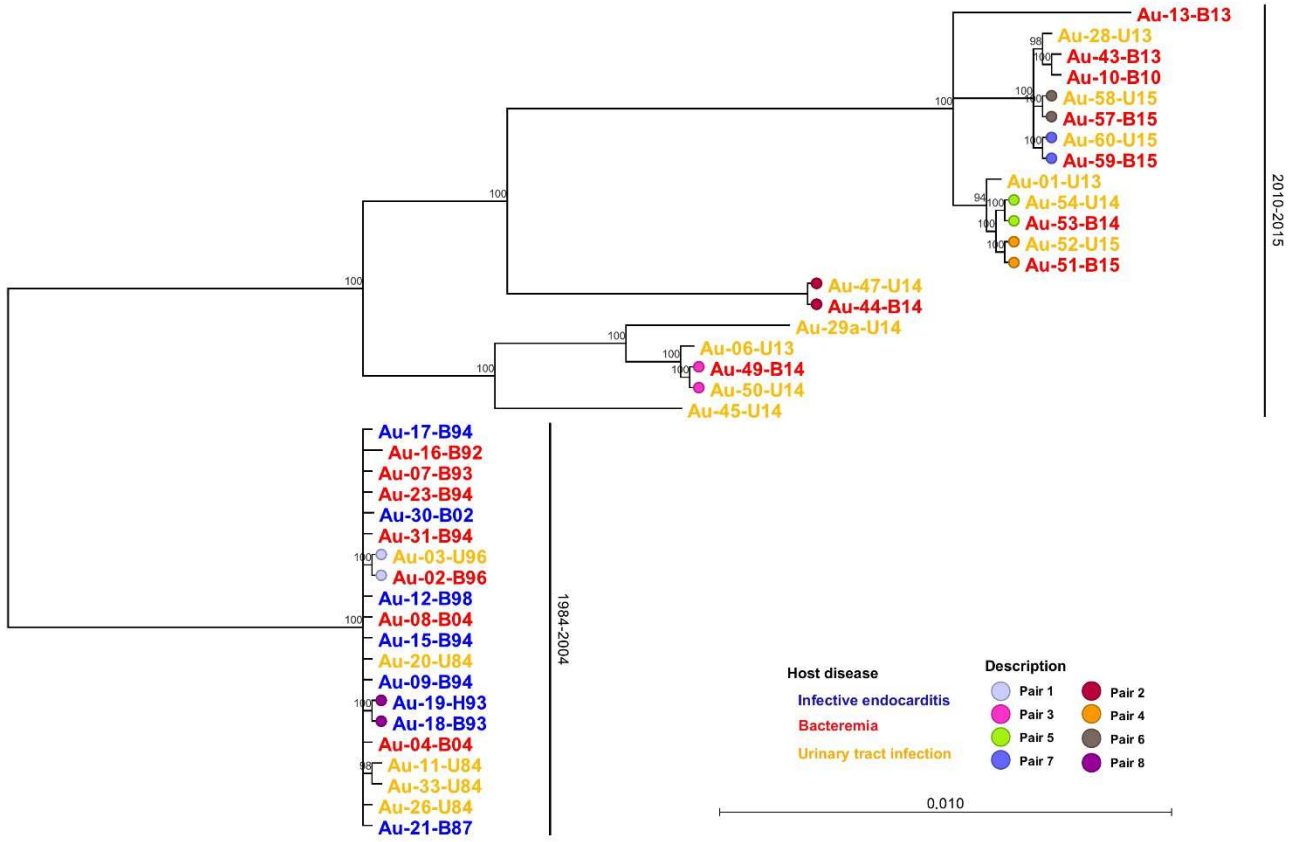


Figure 2. Core-genome phylogeny of the 40 clinical *A. urinae* strains based on the 987 translated common core-genes (corresponding to 312,235 amino acids). The tree showed two major clustering of strains, one with the 1984-2004 strains and the other with strains from 2010-2015. Sub-clustering was observed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

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

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

The 16,678 pan-genes of *A. sanguinicola* and 72,930 pan-genes of *A. urinae* contained 12 and 20 VFDB homolog virulence genes, respectively. Thirty-four out of 1,170 *A. sanguinicola* core-genes were identified as 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	*A. sanguinicola* [1]	*A. urinae* [2]
			Sequence identity in % (*n*)	Sequence identity in % (*n*)
VFDB category: Antiphagocytosis				
S. aureus ssp. *aureus* MW2	CPS protein Cap8A	*cap8A*	34.3 (9)	30.4-32.0 (41)
	CPS protein Cap8B	*cap8B*	36.0-36.2 (9)	37.9-39.2 (41)
	CPS protein Cap8C	*cap8C*	-	43.6-45.6 (41) [3a]
	CPS protein Cap8D	*cap8D*	48.0-48.3 (9)	46.9-47.4 (24) & 63.7 (3) [4a]
	CPS protein Cap8F	*cap8F*	54.7 (9)	53.7-53.9 (22)
	CPS protein Cap8G	*cap8G*	50.8 (9)	50.8-51.9 (22)
	CPS protein Cap8N	*cap8N*	38.4 (9)	38.9-40.7 (27)
S. pneumoniae TIGR4	CPS protein Cps4A	*cps4A*	-	35.3-36.1 (40) & 33.3-42.9 (1) [4b]
	CPS protein Cps4E	*cps4E*	60.4 (9)	57.8-59.4 (23) & 57.3 (4) [3b]
	CPS protein Cps4F	*cps4F*	33.9-34.2 (9)	33.2-33.4 (22)
	CPS protein Cps4H	*cps4H*	-	30.6-31.4 (5)
	CPS protein Cps4I	*cps4I*	-	63.0 (2)
	CPS protein Cps4J	*cps4J*	70.6-70.9 (9)	70.6 (21) & 74.4 (1) [4c]
E. faecalis V583	Undecaprenyl diphosphate synthase	*cpsA*	49.8 (9)	51.4 (41)
	Phosphatidate cytidylyltransferase	*cpsB*	41.7 (9)	42.2-42.9 (41)
	UDP-galactopyranose mutase	*cpsI*	-	60.5 (14)
S. agalactiae 2603V/R	Glycosyl transferase CpsE	*cpsE*	-	33.9 (12) & 58.5-71.4 (2) [4d]
	Glycosyl transferase CpsJ	*cpsJ*	34.9-35.3 (9)	-
	CPS protein CpsL	*cpsL*	-	32.7 (14)
	Glycosyl transferase CpsO	*cpsO*	45.7 (9)	-
	N-acetyl neuramic acid synthetase NeuB	*neuB*	-	39.8-40.4 (41)
S. pyogenes M1	UDP-glucose 6-dehydrogenase HasB	*hasB*	-	52.3 (24)
	UDP-glucose pyrophosphorylase HasC	*hasC*	66.2-66.6 (9)	50.7-51.9 (41)
C. jejuni ssp. *jejuni* NCTC 11168	UDP-glucose 6-dehydrogenase KfiD	*kfiD*	49.6-49.8 (9)	-
VFDB category: Adherence				
L. pneumophila ssp. *pneumophila* str. *Philadelphia 1*	Hsp60, 60K heat shock protein HtpB	*htpB*	56.1-56.3 (9)	53.6-54.0 (41)
L. monocytogenes EGD-e	Fibronectin-binding protein FbpA	*fbpA*	41.5-41.8 (9)	-
	Listeria adhesion protein LAP	*lap*	-	54.4-54.7 (41)
S. agalactiae 2603V/R	Laminin-binding surface protein Lmb	*lmb*	32.4 (9)	56.2-56.9 (41)
S. pyogenes M1	Fibronectin-binding protein Fbp54	*fbp54*	-	42.2-43.1 (41)
V. vulnificus YJ016	Immunogenic lipoprotein A IlpA	*ilpA*	38.0 (9)	38.1-39.2 (41)
VFDB category: Biofilm formation				
E. faecalis V583	Sugar-binding transcriptional regulator	*bopD*	31.8-32.2(9)	-

471 CPS, Capsular polysaccharide.
472 [1] *A. sanguinicola* strains: Eight clinical and one type strain.
473 [2] *A. urinae* strains: Forty clinical and one type strain.
474 [3a/b] 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.

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

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

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

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

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

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

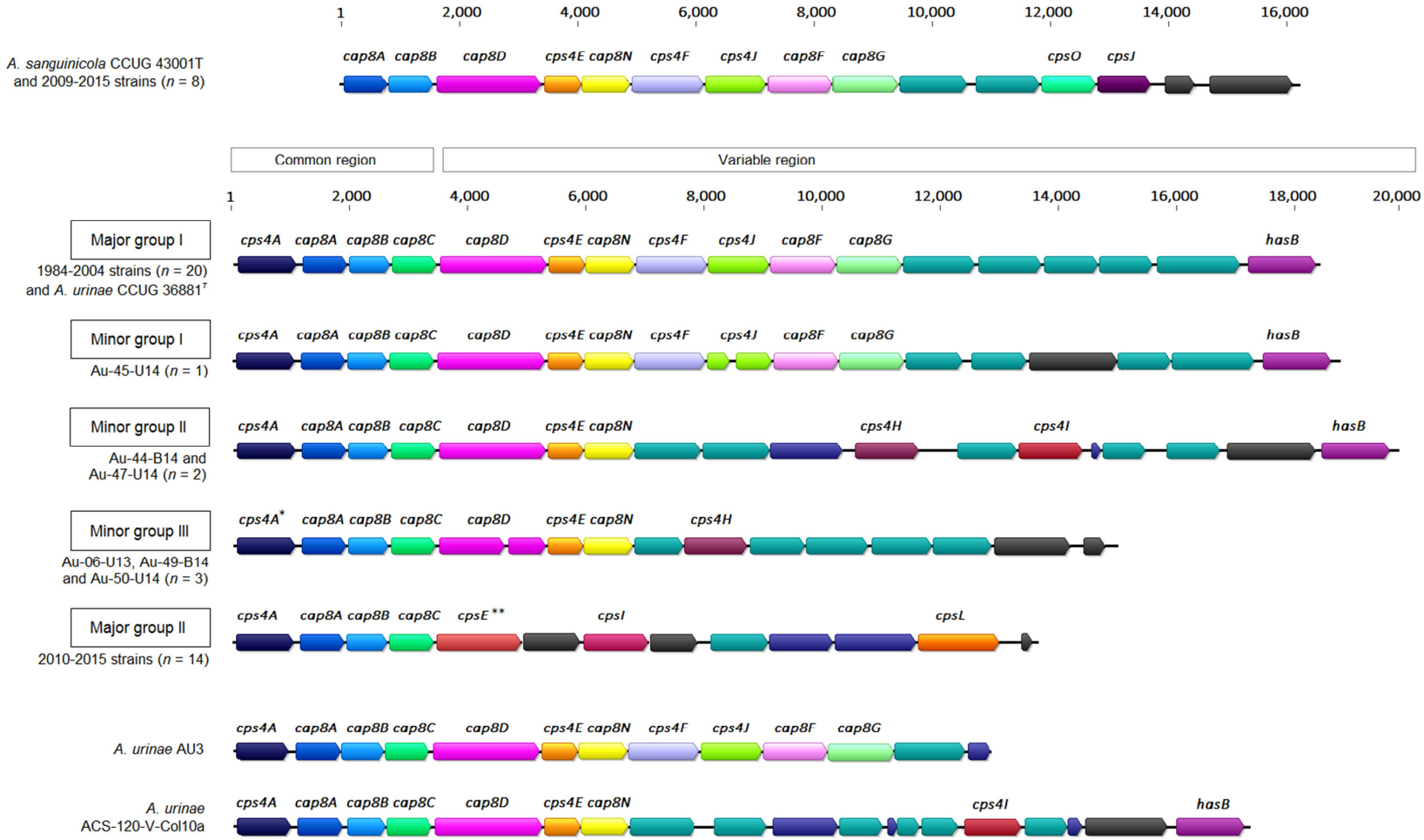
511 gene homolog were identified as transcriptionally regulator proteins in all strains and identified as the
512 common CPS region, *cps4A* (88.8-100 % protein sequence identity), *cap8A* (92.9-100 %), *cap8B* (94.9-100 %),
513 and *cap8C* (86.3-100 %). Higher protein identities were observed when the 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			
	cps4A	*cap8A*	*cap8B*	*cap8C*
All *A. urinae* strains	88.8-100 %	92.9-100 %	94.9-100 %	86.3-100%
Major group I - *A. urinae* strains from 1984-2004 (*n* = 20)	99.7-100 %	100 %	99.1-100 %	100 %
Major group II - *A. urinae* strains from 2010-2015 (*n* = 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^{T}, 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



		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
	A. sanguinicola	15	16,000	15	11
A. urinae	Major group I	16	18,300	17	12
	Major group II	11	13,500	13	7
	Minor group I	16	18,500	17	12
	Minor group II	14	19,500	18	10
	Minor group III	12	14,700	15	8
	A. urinae CCUG 36881^T	16	18,300	17	12
	A. urinae AU3	16	12,800	13	11
	A. urinae ACS-120-V-Col10a	13	17,100	19	9

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

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

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

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

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

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

4. DISCUSSION

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

4.1. Genomic analysis.

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

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

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

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

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

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

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.

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

4.2.2. Bacterial adherence.

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

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

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

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

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

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

4.2.3. Biofilm formation.

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

5. Future perspectives.

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

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

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

6. CONCLUSIONS

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

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

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

COMPETING INTERESTS

The authors declare no competing interest.

FUNDING

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

ACKNOWLEDGMENTS

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

AUTHORS' CONTRIBUTIONS

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

REFERENCES

[1] R.E.O. Williams, A. Hirch, S.T. Cowan, Aerococcus, a New Bacterial Genus, J. Gen. Microbiol. 8 (1953) 475–480. doi:10.1099/00221287-8-3-475.

[2] M. Tohno, M. Kitahara, S. Matsuyama, K. Kimura, M. Ohkuma, K. Tajima, Aerococcus vaginalis sp. nov., isolated from the vaginal mucosa of a beef cow, and emended descriptions of Aerococcus suis, Aerococcus viridans, Aerococcus urinaeequi, Aerococcus urinaehominis, Aerococcus urinae, Aerococcus christensenii and Aerococcus sa, Int. J. Syst. Evol. Microbiol. 64 (2014) 1229–1236. doi:10.1099/ijs.0.058081-0.

[3] J.J. Christensen, B. Korner, H. Kjaergaard, Aerococcus -like organism - an unnoticed urinary tract pathogen, APMIS. 97 (1989) 539–546. doi:10.1111/j.1699-0463.1989.tb00828.x.

[4] M. Aguirre, M.D. Collins, Phylogenetic analysis of some Aerococcus-like organisms from urinary tract infections: description of Aerococcus urinae sp. nov., J. Gen. Microbiol. 138 (1992) 401–405. doi:10.1099/00221287-138-2-401.

[5] P.A. Lawson, E. Falsen, K. Truberg-Jensen, M.D. Collins, Aerococcus sanguicola sp. nov., isolated from a human clinical source, Int. J. Syst. Evol. Microbiol. 51 (2001) 475–479. doi:10.1099/00207713-51-2-475.

[6] J.J. Christensen, I.P. Jensen, J. Faerk, B. Kristensen, R. Skov, B. Korner, the D.A.S. Group, Bacteremia/Septicemia Due to Aerococcus-Like Organisms: Report of Seventeen Cases, Clin. Infect. Dis. 21 (1995) 943–947. doi:10.1093/clinids/21.4.943.

[7] M. Rasmussen, Aerococci and aerococcal infections, J. Infect. 66 (2013) 467–474. doi:10.1016/j.jinf.2012.12.006.

[8] Y.N. Guilarte, R. Tinguely, A. Lupo, A. Endimiani, Prevalence and characteristics of fluoroquinolone-resistant Aerococcus urinae isolates detected in Switzerland, Int. J. Antimicrob. Agents. 43 (2014) 474–483. doi:10.1016/j.ijantimicag.2014.01.022.

[9] M. Rasmussen, Aerococcus: an increasingly acknowledged human pathogen, Clin. Microbiol. Infect. 22 (2016) 22–27. doi:10.1016/j.cmi.2015.09.026.

[10] J.J. Christensen, M. Kilian, V. Fussing, K. Andresen, J. Blom, B. Korner, A.G. Steigerwalt, Aerococcus urinae: Polyphasic characterization of the species, APMIS. 113 (2005) 517–525. doi:10.1111/j.1600-0463.2005.apm_183.x.

[11] E. Senneby, L. Göransson, S. Weiber, M. Rasmussen, A population-based study of aerococcal bacteraemia in the MALDI-TOF MS-era, Eur. J. Clin. Microbiol. Infect. Dis. 35 (2016) 755–762. doi:10.1007/s10096-016-2594-z.

[12] J.J. Christensen, R. Dargis, M. Hammer, U.S. Justesen, X.C. Nielsen, M. Kemp, T.D.M.-T.M.S. Group, Matrix-assisted laser desorption ionization-time of flight mass spectrometry analysis of gram-positive, catalase-negative cocci not belonging to the Streptococcus or Enterococcus genus and benefits of database extension, J. Clin. Microbiol. 50 (2012) 1787–1791. doi:10.1128/JCM.06339-11.

[13] O. Opota, G. Prod'hom, C. Andreutti-Zaugg, M. Dessauges, L. Merz, G. Greub, J.P. Chave, K. Jaton, Diagnosis of Aerococcus urinae infections: Importance of matrix-assisted laser desorption ionization time-of-flight mass spectrometry and broad-range 16S rDNA PCR, Clin. Microbiol. Infect. 22 (2016) e1–e2. doi:10.1016/j.cmi.2015.08.026.

[14] X.C. Nielsen, D. Carkaci, R. Dargis, L. Hannecke, J.U. Stenz, M. Kemp, M. Hammer, J.J. Christensen, 16S-23S Intergenic Spacer (ITS) Region Sequence Analysis: Applicability and Usefulness in Identifying Genera and Species Resembling Non-Hemolytic Streptococci, Clin. Microbiol. Open Access. 2 (2013) 1–8. doi:10.4172/2327-5073.1000130.

[15] A.L. Flores-Mireles, J.N. Walker, M. Caparon, S.J. Hultgren, Urinary tract infections: epidemiology, mechanisms of infection and treatment options, Nat. Rev. Microbiol. 13 (2015) 269–284. doi:10.1038/nrmicro3432.

[16] M.A. Kielhofner, R.J. Hamill, Role of Adherence in Infective Endocarditis, Texas Hear. Inst. J. 16 (1989) 239–249.

[17] H.S. Courtney, Y. Li, J.B. Dale, D.L. Hasty, Cloning, sequencing, and expression of a fibronectin/fibrinogen-binding protein from group A streptococci, Infect. Immun. 62 (1994) 3937–3946.

[18] A. Osanai, L. Sheng-Jun, K. Asano, H. Sashinami, D.L. Hu, A. Nakane, Fibronectin-binding protein, FbpA, is the adhesin responsible for pathogenesis of Listeria monocytogenes infection, Microbiol. Immunol. 57 (2013) 253–262. doi:10.1111/1348-0421.12030.

[19] B. Spellerberg, E. Rozdzinski, S. Martin, J. Weber-Heynemann, N. Schnitzler, R. Lütticken, A. Podbielski, Lmb, a protein with similarities to the LraI adhesin family, mediates attachment of Streptococcus agalactiae to human laminin, Infect. Immun. 67 (1999) 871–878.

[20] K.M. Burkholder, A.K. Bhunia, Listeria monocytogenes uses Listeria adhesion protein (LAP) to promote bacterial transepithelial translocation and induces expression of LAP receptor Hsp60, Infect. Immun. 78 (2010) 5062–5073. doi:10.1128/IAI.00516-10.

[21] O. Shannon, M. Mörgelin, M. Rasmussen, Platelet activation and biofilm formation by Aerococcus urinae, an endocarditis-causing pathogen, Infect. Immun. 78 (2010) 4268–4275. doi:10.1128/IAI.00469-10.

[22] E. Senneby, B. Eriksson, E. Fagerholm, M. Rasmussen, Bacteremia with Aerococcus sanguinicola: Case Series with Characterization of Virulence Properties, Open Forum Infect. Dis. 1 (2014) 1–6. doi:10.1093/ofid/ofu025.

[23] I.S. Roberts, The biochemistry and genetics of capsular polysaccharide production in bacteria, Annu. Rev. Microbiol. 50 (1996) 285–315. doi:10.1146/annurev.micro.50.1.285.

[24] K.F. Clark, D. Wadowska, S.J. Greenwood, Aerococcus viridans var. homari: The presence of capsule and the relationship to virulence in American lobster (Homarus americanus), J. Invertebr. Pathol. 133 (2016) 20–26. doi:10.1016/j.jip.2015.11.007.

[25] K.F. Clark, S.J. Greenwood, Aerococcus viridans expression of Cpn60 is associated with virulence during infection of the American lobster, Homarus americanus Milne Edwards, J. Fish Dis. 34 (2011) 831–843. doi:10.1111/j.1365-2761.2011.01300.x.

[26] J.J. Christensen, A.M. Whitney, L.M. Teixeira, A.G. Steigerwalt, R.R. Facklam, B. Korner, D.J. Brenner, Aerococcus urinae: Intraspecies Genetic and Phenotypic Relatedness, Int. J. Syst. Bacteriol. 47 (1997) 28–32. doi:10.1099/00207713-47-1-28.

[27] R.J. Olsen, S.W. Long, J.M. Musser, Bacterial Genomics in Infectious Disease and the Clinical Pathology Laboratory, Arch. Pathol. Lab. Med. 136 (2012) 1414–1422. doi:10.5858/arpa.2012-0025-RA.

[28] D. Carkaci, R. Dargis, X.C. Nielsen, O. Skovgaard, K. Fuursted, J.J. Christensen, Complete Genome Sequences of Aerococcus christensenii CCUG 28831T, Aerococcus sanguinicola CCUG 43001T, Aerococcus urinae CCUG 36881T, Aerococcus urinaeequi CCUG 28094T, Aerococcus urinaehominis CCUG 42038 BT, and Aerococcus viridans CCUG 4311T, Genome Announc. 4 (2016) 1–2. doi:10.1128/genomeA.00302-16.

[29] S. Andrews, FastQC: a quality control tool for high throughput sequence data, www.bioinformatics.babraham.ac.uk/projects/fastqc, (2010).

[30] R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets, Bioinformatics. 27 (2011) 863–864. doi:10.1093/bioinformatics/btr026.

[31] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev, P.A. Pevzner, SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing, J. Comput. Biol. 19 (2012) 455–477. doi:10.1089/cmb.2012.0021.

[32] A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics. 29 (2013) 1072–1075. doi:10.1093/bioinformatics/btt086.

[33] M. V Larsen, S. Cosentino, O. Lukjancenko, D. Saputra, S. Rasmussen, H. Hasman, T. Sicheritz-Pontén, F.M. Aarestrup, D.W. Ussery, O. Lund, Benchmarking of methods for genomic taxonomy, J. Clin. Microbiol. 52 (2014)

1529–1539. doi:10.1128/JCM.02981-13.

[34] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410. doi:10.1016/S0022-2836(05)80360-2.

[35] O. Lukjancenko, M.C. Thomsen, M.V. Larsen, D.W. Ussery, PanFunPro: PAN-genome analysis based on FUNctional PROfiles, F1000Research. 2 (2013) 1–13. doi:10.12688/f1000research.2-265.v1.

[36] D. Hyatt, G.-L. Chen, P.F. LoCascio, M.L. Land, F.W. Larimer, L.J. Hauser, Prodigal: prokaryotic gene recognition and translation initiation site identification, BMC Bioinformatics. 11 (2010) 1–11. doi:10.1186/1471-2105-11-119.

[37] E.M. Zdobnov, R. Apweiler, InterProScan - an integration platform for the signature-recognition methods in InterPro, Bioinformatics. 17 (2001) 847–848. doi:10.1093/bioinformatics/17.9.847.

[38] M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E.L.L. Sonnhammer, S.R. Eddy, A. Bateman, R.D. Finn, The Pfam protein families databases, Nucleic Acids Res. 30 (2012) D290–D301. doi:10.1093/nar/gkr1065.

[39] D.H. Haft, J.D. Selengut, O. White, The TIGRFAMs database of protein families, Nucleic Acids Res. 31 (2003) 371–373. doi:10.1093/nar/gkg128.

[40] D. Wilson, R. Pethica, Y. Zhou, C. Talbot, C. Vogel, M. Madera, C. Chothia, J. Gough, SUPERFAMILY - Sophisticated comparative genomics, data mining, visualization and phylogeny, Nucleic Acids Res. 37 (2009) D380–D386. doi:10.1093/nar/gkn762.

[41] W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics. 22 (2006) 1658–1659. doi:10.1093/bioinformatics/btl158.

[42] R.C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797. doi:10.1093/nar/gkh340.

[43] D. Darriba, G.L. Taboada, R. Doallo, D. Posada, jModelTest 2: more models, new heuristics and parallel computing, Nat. Methods. 9 (2012) 772. doi:10.1038/nmeth.2109.

[44] S. Guindon, O. Gascuel, A Simple, Fast, and Accurate Method to Estimate Large Phylogenies by Maximum Likelihood, Syst. Biol. 52 (2003) 696–704. doi:10.1080/10635150390235520.

[45] R.S. Kaas, P. Leekitcharoenphon, F.M. Aarestrup, O. Lund, Solving the problem of comparing whole bacterial genomes across different sequencing platforms, PLoS One. 9 (2014) 1–8. doi:10.1371/journal.pone.0104984.

[46] L. Chen, D. Zheng, B. Liu, J. Yang, Q. Jin, VFDB 2016: hierarchical and refined dataset for big data analysis-10 years on., Nucleic Acids Res. 44 (2016) D694–D697. doi:10.1093/nar/gkv1239.

[47] M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, A. Drummond, Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data, Bioinformatics. 28 (2012) 1647–1649. doi:10.1093/bioinformatics/bts199.

[48] T.N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: discriminating signal peptides from transmembrane regions, Nat. Methods. 8 (2011) 785–786. doi:10.1038/nmeth.1701.

[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. Brinkman, PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics. 26 (2010) 1608–1615. doi:10.1093/bioinformatics/btq249.

[50] A. Krogh, B. Larsson, G. von Heijne, E.L.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J. Mol. Biol. 305 (2001) 567–580. doi:10.1006/jmbi.2000.4315.

[51] T. Tatusova, M. Dicuccio, A. Badretdin, V. Chetvernin, E.P. Nawrocki, L. Zaslavsky, A. Lomsadze, K.D. Pruitt, M. Borodovsky, J. Ostell, NCBI prokaryotic genome annotation pipeline, Nucleic Acids Res. 44 (2016) 6614–6624. doi:10.1093/nar/gkw569.

[52] D.W. Ussery, T.M. Wassenaar, S. Borini, Computing for Comparative Microbial Genomics - Bioinformatics for Microbiologists - Chapter 12, Springer-Verlag London, UK, 2009. doi:10.1007/978-1-84800-255-5_2.

[53] E.J. Feil, Small change: keeping pace with microevolution, Nat. Rev. Microbiol. 2 (2004) 483–495. doi:10.1038/nrmicro904.

[54] B. Segerman, The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories, Front. Cell. Infect. Microbiol. 2 (2012) 1–8. doi:10.3389/fcimb.2012.00116.

[55] K.A. Bliven, A.T. Maurelli, Evolution of Bacterial Pathogens within the Human Host, Microbiol. Spectr. 4 (2016) 1–13. doi:10.1128/microbiolspec.

[56] T. Lefébure, M.J. Stanhope, Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition, Genome Biol. 8 (2007) R71.1-R71.17. doi:10.1186/gb-2007-8-5-r71.

[57] Y. Feng, C.J. Chen, L.H. Su, S. Hu, J. Yu, C.H. Chiu, Evolution and pathogenesis of Staphylococcus aureus: Lessons learned from genotyping and comparative genomics, FEMS Microbiol. Rev. 32 (2008) 23–37. doi:10.1111/j.1574-6976.2007.00086.x.

[58] A.P. Rooney, J.L. Swezey, R. Friedman, D.W. Hecht, C.W. Maddox, Analysis of core housekeeping and virulence genes reveals cryptic lineages of Clostridium perfringens that are associated with distinct disease presentations, Genetics. 172 (2006) 2081–2092. doi:10.1534/genetics.105.054601.

[59] M. de Been, W. Van Schaik, L. Cheng, J. Corander, R.J. Willems, Recent recombination events in the core genome are associated with adaptive evolution in Enterococcus faecium, Genome Biol. Evol. 5 (2013) 1524–1535. doi:10.1093/gbe/evt111.

[60] R.L. Marvig, L.M. Sommer, S. Molin, H.K. Johansen, Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis, Nat. Genet. 47 (2015) 57–65. doi:10.1038/ng.3148.

[61] S. Tal, V. Guller, S. Levi, R. Bardenstein, D. Berger, I. Gurevich, A. Gurevich, Profile and prognosis of febrile elderly patients with bacteremic urinary tract infection, J. Infect. 50 (2005) 296–305. doi:10.1016/j.jinf.2004.04.004.

[62] A. McNally, F. Alhashash, M. Collins, A. Alqasim, K. Paszckiewicz, V. Weston, M. Diggle, Genomic analysis of extra-intestinal pathogenic Escherichia coli urosepsis, Clin. Microbiol. Infect. 19 (2013) E328–E334. doi:10.1111/1469-0691.12202.

[63] T.T. Luong, C.Y. Lee, Overproduction of Type 8 Capsular Polysaccharide Augments Staphylococcus aureus Virulence Overproduction of Type 8 Capsular Polysaccharide Augments Staphylococcus aureus Virulence, Infect. Immun. 70 (2002) 3389–3395. doi:10.1128/IAI.70.7.3389–3395.2002.

[64] S. Sau, J. Sun, C.Y. Lee, Molecular characterization and transcriptional analysis of type 8 capsule genes in Staphylococcus aureus, J. Bacteriol. 179 (1997) 1614–1621. doi:10.1128/jb.179.5.1614-1621.1997.

[65] R. Jothi, S. Parthasarathy, K. Ganesan, Comparison of the Virulence Factors and Analysis of Hypothetical Sequences of the Strains TIGR4, D39, G54 and R6 of Streptococcus Pneumoniae, J. Comput. Sci. Syst. Biol. 1 (2008) 103–118. doi:10.4172/jcsb.1000010.

[66] U.B. Skov Sørensen, K. Yao, Y. Yang, H. Tettelin, M. Kilian, Capsular Polysaccharide Expression in Commensal *Streptococcus* Species: Genetic and Antigenic Similarities to *Streptococcus pneumoniae*, MBio. 7 (2016) e01844-16. doi:10.1128/mBio.01844-16.

[67] W. Zhengsong, J.-R. Zhang, Bacterial Capsules - Chapter 3, in: Mol. Med. Microbiol., 2nd ed., Elsevier Ltd, 2015: pp. 33–53. doi:http://dx.doi.org/10.1016/B978-0-12-397169-2.00003-2.

[68] S.D. Bentley, D.M. Aanensen, A. Mavroidi, D. Saunders, E. Rabbinowitsch, M. Collins, K. Donohoe, D. Harris, L. Murphy, M.A. Quail, G. Samuel, I.C. Skovsted, M.S. Kaltoft, B. Barrell, P.R. Reeves, J. Parkhill, B.G. Spratt, Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes, PLoS Genet. 2 (2006) e31. doi:10.1371/journal.pgen.0020031.

[69] D.O. Chaffin, S.B. Beres, H.H. Yim, C.E. Rubens, The Serotype of Type Ia and III Group B Streptococci Is Determined by the Polymerase Gene within the Polycistronic Capsule Operon, J. Bacteriol. 182 (2000) 4466–4477. doi:10.1128/JB.182.16.4466-4477.2000.

[70] C. Toniolo, E. Balducci, M.R. Romano, D. Proietti, I. Ferlenghi, G. Grandi, F. Berti, I.M. Ros, R. Janulczyk, Streptococcus agalactiae capsule polymer length and attachment is determined by the proteins CpsABCD, J. Biol. Chem. 290 (2015) 9521–9532. doi:10.1074/jbc.M114.631499.

[71] K. O´Riordan, J.C. Lee, Staphylococcus aureus Capsular Polysaccharides, Clin. Microbiol. Rev. 17 (2004) 218–234. doi:10.1128/CMR.17.1.218–234.2004.

[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 capsular polysaccharide synthesis gene clusters in 79 capsular types of Klebsiella spp., Sci. Rep. 5 (2015) 15573. doi:10.1038/srep15573.

[73] S. Dramsi, F. Bourdichon, D. Cabanes, M. Lecuit, H. Fsihi, P. Cossart, FbpA, a novel multifunctional Listeria monocytogenes virulence factor, Mol. Microbiol. 53 (2004) 639–649. doi:10.1111/j.1365-2958.2004.04138.x.

[74] K.J. Lee, N.Y. Lee, Y.S. Han, J. Kim, K.H. Lee, S.J. Park, Functional characterization of the IlpA protein of Vibrio vulnificus as an adhesin and its role in bacterial pathogenesis, Infect. Immun. 78 (2010) 2408–2417. doi:10.1128/IAI.01194-09.

[75] T. Tenenbaum, B. Spellerberg, R. Adam, M. Vogel, K.S. Kim, H. Schroten, Streptococcus agalactiae invasion of human brain microvascular endothelial cells is promoted by the laminin-binding protein Lmb, Microbes Infect. 9 (2007) 714–720. doi:10.1016/j.micinf.2007.02.015.

[76] S.Y. Goo, Y.S. Han, W.H. Kim, K.-H. Lee, S.-J. Park, Vibrio vulnificus IlpA-induced cytokine production is mediated by toll-like receptor 2, J. Biol. Chem. 282 (2007) 27647–27658. doi:10.1074/jbc.M701876200.

[77] Z.W. Jaradat, J.L. Wampler, A.K. Bhunia, A Listeria adhesion protein-deficient Listeria monocytogenes strain shows reduced adhesion primarily to intestinal cell lines, Med. Microbiol. Immunol. 192 (2003) 85–91. doi:10.1007/s00430-002-0150-1.

[78] B. Jagadeesan, O.K. Koo, K.P. Kim, K.M. Burkholder, K.K. Mishra, A. Aroonnual, A.K. Bhunia, LAP, an alcohol acetaldehyde dehydrogenase enzyme in Listeria, promotes bacterial adhesion to enterocyte-like Caco-2 cells only in pathogenic species, Microbiology. 156 (2010) 2782–2795. doi:10.1099/mic.0.036509-0.

[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-dependent secretory protein, is induced under anaerobic environment, Microbes Infect. 11 (2009) 859–867. doi:10.1016/j.micinf.2009.05.006.

[80] J.L. Wampler, K. Kim, Z. Jaradat, A.K. Bhunia, Heat Shock Protein 60 Acts as a Receptor for the Listeria Adhesion Protein in Caco-2 Cells, Infect. Immun. 72 (2004) 931–936. doi:10.1128/IAI.72.2.931–936.2004.

[81] C. Hennequin, F. Porcheray, A.-J. Waligora-Dupriet, A. Collignon, M. Barc, P. Bourlioux, T. Karjalainen, GroEL (Hsp60) of Clostridium difficile is involved in cell adherence, Microbiology. 147 (2001) 87–96. doi:10.1099/00221287-147-1-87.

[82] G.S. Chhatwal, Anchorless Adhesins and Invasins of Gram-Positive Bacteria: a New Class of Virulence Factors, Trends Microbiol. 10 (2002) 205–208. doi:10.1016/S0966-842X(02)02351-X.

[83] Y. Terao, S. Kawabata, E. Kunitomo, I. Nakagawa, S. Hamada, Novel laminin-binding protein of Streptococcus pyogenes, Lbp, is involved in adhesion to epithelial cells, Infect. Immun. 70 (2002) 993–997. 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