Identification of two abundant Aerococcus urinae cell wall-anchored proteins

Senneby, Erik; Sunnerhagen, Torgny; Hallström, Björn; Lood, Rolf; Malmström, Johan; Karlsson, Christofer; Rasmussen, Magnus

Published in:
International Journal of Medical Microbiology

DOI:
10.1016/j.ijmm.2019.06.005

2019

Document Version:
Peer reviewed version (aka post-print)

Identification of two abundant *Aerococcus urinae* cell wall-anchored proteins.

Erik Senneby[a*], Torgny Sunnerhagen[a], Björn Hallström[b], Rolf Lood[a], Johan Malmström[a], Christofer Karlsson[a] and Magnus Rasmussen[a].

[a]Division of Infection Medicine, Department of Clinical Sciences, BMC B14, 221 85, Lund University, Lund, Sweden.

[b]Centre for Translational Genomics, Division of Clinical Genetics, BMC B10, 221 85, Lund University, Lund, Sweden.

*Corresponding author: erik.senneby@med.lu.se

E-mail addresses: torgny.sunnerhagen@med.lu.se, bjorn.hallstrom@skane.se, rolf.lood@med.lu.se, johan.malmstrom@med.lu.se, christofer.karlsson@med.lu.se, magnus.rasmussen@med.lu.se

Running title: The surface proteome of *A. urinae*

Declaration of interest: none.

Abstract

Aerococcus urinae is an emerging pathogen that causes urinary tract infections, bacteremia and infective endocarditis. The mechanisms through which *A. urinae* cause infection are largely unknown. The aims of this study were to describe the surface proteome of *A. urinae* and to analyse *A. urinae* genomes in search for genes encoding surface proteins. Two proteins, denoted Aerococcal surface protein (Asp) 1 and 2, were through the use of mass spectrometry based proteomics found to quantitatively dominate the aerococcal surface. The presence of these proteins on the surface was also shown using ELISA with serum from rabbits immunized with the recombinant Asp. These proteins had a signal sequence in the amino-terminal end and a cell wall-sorting region in the carboxy-terminal end, which contained an LPATG-motif, a hydrophobic domain and a positively charged tail. Twenty-three additional *A. urinae* genomes were sequenced using Illumina HiSeq technology. Six different variants of *asp* genes were found (denoted *asp*1-6). All isolates had either one or two of these *asp*-genes located in a conserved locus, designated Locus encoding Aerococcal Surface Proteins (LASP). The 25 genomes had in median 13 genes encoding LPXTG-proteins (range 6-24). For other Gram-positive bacteria, cell wall-anchored surface proteins with an LPXTG-motif play a key role for virulence. Thus, it will be of great interest to explore the function of the Asp proteins of *A. urinae* to establish a better understanding of the molecular mechanisms by which *A. urinae* cause disease.

Key Words: *Aerococcus urinae*, surface proteome, LPXTG-motif, genes encoding surface proteins.

Introduction

Aerococcus urinae is a Gram-positive coccus that belongs to the phylum of Firmicutes. *A. urinae* is known to cause urinary tract infection (Christensen et al., 1989; Schuur et al., 1997; Senneby et al., 2015), bacteremia (Christensen et al., 1995; Senneby et al., 2016) and infective endocarditis (de Jong et al., 2010; Ebnother et al., 2002; Kristensen and Nielsen, 1995; Sunnerhagen et al., 2016), mostly in the elderly male population. Since the introduction of MALDI-TOF MS as a diagnostic tool in clinical microbiology laboratories, *A. urinae* is more frequently identified in clinical samples, especially in urine cultures, and has attracted attention as an important emerging pathogen (Rasmussen, 2016). A few studies have reported that *A. urinae* is part of a urinary tract microbiome (Hilt et al., 2014; Pearce et al., 2014). The mechanisms through which *A. urinae* establishes colonization and causes infection are largely unknown. Only one study has previously targeted the question of *A. urinae* virulence mechanisms (Shannon et al., 2010), demonstrating that it produces biofilm and is able to induce human platelet activation and aggregation. These are properties of potential importance in the process of establishing infections in the human host (Donlan and Costerton, 2002).

At the molecular level, little is known regarding aerococcal virulence factors (Carkaci et al., 2017). For other pathogenic Firmicutes, such as *Streptococcus pyogenes* and *Staphylococcus aureus*, cell wall-attached proteins with an LPXTG-motif play key roles for virulence. For instance, the M protein of *S. pyogenes* is an abundant surface protein, which possesses a broad spectrum of functions, such as the inhibition of phagocytosis and binding of several plasma proteins (Fischetti, 1989; Smeesters et al., 2010). The amino (NH_2)-terminal part of the protein displays hypervariability, resulting in antigenic variation (Cunningham, 2000). The LPXTG-proteins share common features, such as a signal sequence in the NH_2-terminal end and three characteristics in the carboxy (COOH)-terminal end; a cell wall-sorting region containing the LPXTG-motif, a hydrophobic membrane-spanning domain and a positively charged tail positioned in the cytoplasm. The LPXTG-motif is recognized by the membrane bound enzyme sortase and is, after cleavage between the threonine and glycine residues, covalently attached to the cell wall (Schneewind and Missiakas,

2014). It is at present unclear if *A. urinae* express proteins attached to the cell wall through the LPXTG-motif.

Mass spectrometry (MS) based proteomics has previously been shown to be a useful method to detect bacterial surface-associated proteins with protein copies per cell accuracy (Malmstrom et al., 2009) and to determine the surface protein composition of gram-positive bacteria (Kilsgard et al., 2016; Rodriguez-Ortega et al., 2006; Severin et al., 2007). However, a presumption to utilize this technique is the genomic sequence for the analysed bacterial species or strain. Hence, the combination of next generation sequencing and MS-based proteomics constitutes a powerful strategy to search for potential novel virulence factors.

The aims with this study were to describe the surface proteome of *A. urinae* and to describe the genes encoding surface proteins.

Material and Methods

Bacteria and culturing conditions

The strain ACS-120-V-Col10a (Col10a) was retrieved from the Culture collection of Gothenburg. The AU3 strain was collected at the Clinical Microbiology laboratory, Lund and originated from a blood culture as described previously (Senneby et al., 2012). Forty-six *A. urinae* isolates from blood cultures had been described previously (Senneby et al., 2016) and were designated as AuB followed by a number. *A. urinae* isolates were cultivated in Tryptic Soy Broth with 0.25% glucose (TSBG) for approximately 24 hours at 35° C in 5 % CO_2.

Bacterial surface digestion and MS sample preparation

Surface proteins were released from AU3 and Col10 stationary phase cells in triplicates with a modified protocol as previously described (Rodriguez-Ortega et al., 2006; Severin et al., 2007). The cells (~8 x10^8 colony forming units (CFU)) were washed with 20 mM Tris-HCl, 150 mM NaCl, pH 7.6 (TBS) and resuspended in 1 M d-arabinose, 10 mM $CaCl_2$ in TBS and 5 µg sequencing grade trypsin (Promega) and incubated at 37 °C with 500 rpm shaking for 20 min. The digested mixture was centrifuged with swing-out rotor at 2000 x g for 15 min at 4 °C. ProteaseMAX

(Promega) was added to the supernatants (surface fraction) to a final concentration of 0.01 % following by heating to 80 °C for denaturation. Cysteine residues were reduced with 25 mM tris(2-carboxyethyl)phosphine (Sigma-Aldrich, TCEP) and alkylated with 25 mM 2-Iodoacetamide (Sigma-Aldrich, IAA). Surface fraction samples were fully digested with 2 µg sequencing grade trypsin (Promega) for 10 h at 37 °C and the sample acidified with 0.5 % Trifluoroacetic acid. The digested cells (cellular fraction) were washed with TBS and resuspended in water and homogenized using a Fastprep-96 beadbeater (MPBio) with Lysing Matrix B tubes (MPBio). The cell lysates were denatured with 8 M Urea in 100 mM ammonium bicarbonate (ABC) and then reduced with 25 mM TCEP for 1 h at 37°C, and alkylated with 25 mM iodoacetamide for 45 min before diluting the sample with 100 mM ABC to a final urea concentration below 1.5 M. Proteins were digested by incubation with trypsin (trypsin:protein ratio of 1:100 (w:w)) for 10 h at 37°C. The peptides from both fractions were desalted and cleaned-up with reversed-phase spin columns (Vydac UltraMicroSpin Silica C18 300Å Columns, Harvard Apparatus) according to the manufacturer's instructions.

MS data acquisition

Peptide analyses were performed on a Q Exactive Plus mass spectrometer (Thermo Scientific) connected to an EASY-nLC 1000 ultra-high-performance liquid chromatography system (Thermo Scientific). Peptides were separated on an EASY-Spray ES802 columns (Thermo Scientific) using a linear gradient from 3 to 35% acetonitrile in aqueous 0.1% formic acid during 2 h. Data-dependent acquisition mode (DDA) and Data-independent acquisition mode (DIA) instrument settings were identical to as described in (Malmstrom et al., 2016). From the DDA data, spectral libraries were built using the TPP Fraggle workflow (Teleman et al., 2017) using NCBI fasta files GCA_000193205.1_ASM19320 and GCA_001649715.1_ASM164971 respectively concatenated with iRT peptides, contaminants and decoys. The generated spectral libraries were used to extract the DIA data with DIANA v2.0.0 (Teleman et al., 2015) using a 1% peptide false discovery rate.

MS data analysis

Data was processed with custom R-scripts using the tidyverse package collection together with broom package for statistical functions. For the quantification of proteins, all integrated peptide ion intensities extracted from the MS2 spectra was summed up by protein and then divided by protein length. Peptides matching more than one protein were not included in the analysis. Protein quantification data was normalized based on the sample total intensity. Proteins from Col10a and AU3 were classified into orthologous pairs using ProteinOrtho.pl v5.16b (Lechner et al., 2011) and protein domains predicted using standalone InterProScan v5.11-50 (Jones et al., 2014).

Genome sequencing

Bacteria were pelleted through centrifugation and stored at -70° C until shipment on dry ice to GATC Biotech (Konstanz, Germany). Library preparation was performed using an optimized protocol and standard Illumina adapter sequences were used. Sequencing was performed with Illumina HiSeq 2500 (Illumina, Inc., San Diego, US) with paired-end reads, 2 x 125 bp. The genomes of AU3 (ASM164971v1, GenBank assembly accession GCA_001649715.1) and ACS-120-V-Col10a (ASM19320v1, GenBank assembly accession GCA_00193205.1) had been previously published. For assembly of the genomes, SPAdes 3.9 with the careful-mode on, was used through services on the Center for Genomic Epidemiology's website (Nurk et al., 2013). The computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). Assembly quality was evaluated using QUAST (Gurevich et al., 2013). Sample contamination was evaluated with Kraken (cross-species) (Wood and Salzberg, 2014) and by mapping the raw sequence data to a reference genome using BWA MEM (Li and Durbin, 2009) and evaluating the presence of minority bases with samtools mpileup (within-species) (Li et al., 2009). In cases with within-species contamination the contaminant gene were eliminated by coverage analysis, i.e. genes with low coverage were excluded. The BLAST software (Zhang et al., 2000) was used for searches in the genomes. Sequence alignment was performed with Clustal Omega. The maximum likelihood method was used for building phylogenetic tress through the use of the MEGA software (version 7.0.26).

Annotation of the genomes was performed with RASTtk on the PATRIC platform (Wattam et al., 2017). SignalP 4.1 was used to locate signal peptide cleavage sites (Petersen et al., 2011). Hydrophobic plots were performed using services on the Expasy website (Kyte and Doolittle, 1982). The ClustalW multiple sequence alignment was used to produce identity scores (%) for the amino acid sequences with the following parameters: similarity matrix: gonnet, open gap penalty 10, extend gap penalty 0.1-0.2, gap distance 4, delay divergent 30%. The signal sequences and the N-terminal end (starting with the LPXTG-motive) were deleted prior to analysis. Searches were also performed in 26 *A. urinae* genomes belonging to Bioproject PRJNA315093, that were accessible on the NCBI webpage. The bacterial isolates originated from the female urinary tract.

Expression cloning and protein production

The MacVector software (v. 14.5.3) was used to analyse the bacterial genome for suitable primer sites for amplification of *asp* genes. Primers were ordered from Eurofins and can be found in Table 1. Chromosomal DNA from AU3 was extracted using the innuPRerp Bacterial DNA System (Analytik Jena AG). PCR products were inserted into the pGEX-6P-1 expression vector. One Shot TOP10 *E. coli* (Thermo Fisher Scientific) were transformed and used for plasmid amplification. Plasmids were extracted using Qiagen Plasmid Plus kits. The plasmids were sent to GATC GmbH for sequencing to ensure that the sequence was correct. BL21 cells with pLysS were then transformed according to the manufacturer's protocol, with selection for pGEX-6-p-1 (100 μg/ ml ampicillin) and for pLysS (32 μg/ml chloramphenicol).

For protein expression, an overnight *E. coli* culture carrying the construct was diluted 1:25 in fresh TSBG supplemented with antibiotics and grown for 3 h. IPTG (1 mM) was then added and incubation continued for 3 h at 37° C. The bacteria were pelleted and the supernatant discarded. Pellet was dissolved in BugBuster (Merck Millipore) suspended in 20 mM TRIS (pH 7.4) supplemented with 1:100 (v/v) benzonase. After 10 minutes of incubation at room temperature bacterial cell debris was collected by centrifugation and the supernatant was loaded onto an equilibrated GST-column, washed, and finally eluted (elution buffer). The sequence of the affinity-purified protein was confirmed by mass-spectrometry at the SciBlu core facility at Lund University.

Antiserum generation and ELISA

To generate antiserum against the affinity-purified recombinant proteins, the proteins were sent to Biogenes GmbH for immunization. To assess binding of antibodies to the surface proteins an ELISA was performed. Nunc MaxiSorp (ThermoFisher Scientific) 96-well plates were coated with either a suspension of bacteria (AU3, Col10a or *A. viridans*), or a solution of Asp-proteins in coating buffer (1.69 g Na_2CO_3 and 2.94 g $NaHCO_3$ in 1 L H_2O). As a control, bacteria with reduced surface protein expression were generated by incubating a bacterial solution with trypsin (0.0004% w/v) at 37° C for 20 minutes after which the reaction was stopped by adding trypsin inhibitor at a molar ratio of 8:1. The trypsinated bacteria and control bacteria (where trypsin inhibitor was added before the trypsin) were then pelleted by centrifugation, washed (PBS), and finally resuspended in coating buffer. The bacteria were then heat-killed at 85° C for 30 minutes before adding the samples to the Nunc plate for coating over night at 4° C. The plates were washed (PBST), rabbit serum in different dilutions added, followed by further washes and addition of protein G coupled HRP (1:3000). As an HRP substrate, ABTS dissolved 1:20 in a substrate solution (21 g citric acid monohydrate and 17.8 g Na_2HPO_4 2 H_2O in 1 L H_2O, pH 4.5) was used in combination with H_2O_2; incubated in room temperature for 15 minutes. The plates were then analyzed with the iMark Microplate reader (Bio-Rad) at 415 nm.

Sequence similarity network (SNN) of LPXTG-containing ORFs.

In total, 106 285 ORFs from the 25 *A. urinae* genomes were analyzed with standalone InterProScan v5.11-50 (Jones et al., 2014) and included the following analyses: Pfam, PANTHER, Gene3D, CDD, TIGRFAM, ProSitePatterns, SUPERFAMILY, Hamap, ProSiteProfiles, Coils, SMART, MobiDBLite, PRINTS, PIRSF, SignalP, SFLD and ProDom. LPXTG-motifs were predicted with InterPro signature accessions PS50847 and/or TIGR01167 and/or PF00746 and sequences identified through this procedure were analyzed by all-by-all BLAST comparisons using blastp (v2.7.1+ with default settings). The blast result was filtered with the following thresholds: *E*-value < 1×10^{-10}, >50 % sequence identity and >75 alignment length.

The SNN was visualized using Cytoscape (v3.5.1) (Shannon et al., 2003), where each node represents a protein/ORF and an edge or line between the nodes denotes a

similarity relationship between the proteins. The “organic” layout was used whereby nodes are clustered more tightly if they are more highly interconnected and all edges were bundled.

Results

Quantitative profiling of *A. urinae* surface proteomes.

To search for putative novel surface-associated proteins, we released surface exposed proteins from hypotonically swelled *A. urinae* cells of the strains Col10a and AU3 with a short trypsin digestion step. This produced a surface protein fraction and a cell protein fraction. This method has previously been used for releasing surface exposed proteins from cells of other gram-positive bacteria (Rodriguez-Ortega et al., 2006; Solis et al., 2010). The trypsinated cells (the cell fraction) were homogenized and used as control samples hypothetically containing mostly intracellular and membrane proteins. Both surface and cell fractions were then prepared for MS analysis. All derived peptide samples (n = 12) were analysed using data-independent-acquisition (DIA) followed by targeted data extraction (SWATH-MS) (Fig. 1A). In this analysis we quantified a total of 1118 and 1168 proteins of which 765 and 682 proteins were detected in the surface fractions from strains Col10a and AU3 respectively (Fig. 1B). Essentially all surface fraction proteins (99.8 %) were also detected in the cellular fractions (Fig. 1C-D), but with large differences of abundances between fractions. The majority of proteins in the surface fraction were much less abundant in the cellular fraction (Fig. 2A). We defined surface associated proteins based on two criteria; 10-fold enriched in the surface associated fractions and a Hochberg adjusted p-value < 0.01 compared to the cellular fraction (Col10a, n= 24; AU3, n = 30) (Fig. 2A and B). The stringent cut-off values for the surface associated protein groups were selected to distinctly separate proteins identified in the two fractions. The defined surface associated proteins included the most abundant proteins in the surface fraction (Fig. 2C) and 70% of these proteins had a predicted signal peptide domain as determined by InterProScan analysis (Fig. 2D). Of the proteins identified in the surface fraction but not defined as surface associated, most were ribosomal proteins,

metabolic enzymes and transcriptional regulators and only ~9 % had a predicted signal peptide (Fig. 2D).

Next we compared the surface associated protein orthology between the two strains as defined by the bidirectional best-hit (Lechner et al., 2011). Of the total 54 surface associated proteins from both strains, 17 pairs were determined to be orthologous by the Proteinortho software (Lechner et al., 2011) (Fig. 2E). Based on this analysis we identified two orthologous protein pairs with predicted LPXTG-sequences and signal peptide protein domains that were abundantly produced on the surface (Fig. 2E, red nodes with dotted edges). The proteins were denoted Aerococcal surface protein (Asp) 1 and 2 and the corresponding genes were accordingly named (*asp1* and *asp2).* The Asp1 was the most abundant protein on the surface, with a relative abundance of approximately 73 % and 26 % of the total proteins on the surface (for Col10a and AU3 respectively). No proteins similar to the Asp proteins were found in the InterPro database that includes high-level structure-based classifications and sub-family classifications. Furthermore, InterPro analysis of motifs and/or domains only identified LPXTG and signal peptide sequences. Additionally, four other surface associated proteins were classified as LPXTG-anchored. However, these proteins lacked an orthologous counterpart or the orthologous counterpart did not harbour a predicated LPXTG sequence (see Fig. 2E). The total number of predicted LPXTG proteins in the genomes was 8 for AU3 and 10 for Col10a (not shown). In summary, the quantitative profiling of *A. urinae* surface proteomes revealed the that less than half all theoretical LPXTG proteins are expressed during the experimental conditions used and that both surfaces are dominated by an orthologous pair of proteins (Asp1).

Genomic and protein sequence analysis of Aerococcal surface protein 1 and 2.

Based on the outcome from the MS-analysis, we continued with acquiring 23 additional novel whole genome sequences of *A. urinae* isolates to investigate the presence of genes encoding Asp proteins and also other LPXTG-containing proteins. The genomes of the AU3 and the Col10a isolates were included in the analysis. Thus, 25 *A. urinae* genomes were studied. Variants of the *asp*-genes were found in all isolates sequenced. The genes were denoted *asp1-6,* based on a phylogenetic analysis (Figure 3), and the corresponding gene products were denoted Asp1-6. The *asp*-genes were located adjacent to each other in a chromosomal locus of the aerococcal

genome, in this study designated as Locus encoding Aerococcal Surface Proteins (LASP). In all isolates it was constituted by, in a 5' – 3' order, a pyruvate carboxylase gene, a gene designated as a "hypothetical cytosolic protein", the asp-gene(s) and a 16S rRNA methyltransferase gene. The LASP is schematically depicted in Figure 4. The primary structure of Asp1 contained 284-298 amino acids and Asp2 contained 307-382 amino acids. The Asp3 contained 376-407 amino acids, all Asp4 had 425 amino acids and Asp5 and Asp6 had 483 amino acids. The comparison of the amino acid sequences for Asp1-6 is presented in Table 2 with identity scores (%). The amino acid sequence of the LPXTG-motif was LPATG in all isolates. The aerococcal isolates had combinations of several different *asp*-genes in their LASP. Four variants were found, designated $LASP_{1-4}$. In $LASP_1$ only the *asp1* was present, $LASP_2$ had both *asp1* and *asp2*, $LASP_3$ had *asp3* and *asp6* whereas $LASP_4$ contained *asp4* and *asp5*.

The 26 *A. urinae* genomes in the bio project PRJNA315093 also included *asp* genes and these isolates had either $LASP_1$, $LASP_2$ or $LASP_4$.

Expression cloning and immunological assays

The predicted mature forms of Asp1 and Asp2 were recombinantly expressed in *E. coli* and were used for production of two polyclonal antisera in rabbits. ELISA demonstrated that the post-immune antisera from both rabbits reacted with Asp1 and Asp2, whereas there was no reaction between the pre-immune sera (data not shown). To confirm that Asp1 and Asp2 were present on the surface of *A. urinae*, ELISA using anti-Asp1 and anti-Asp2 antisera against whole bacteria immobilized in the wells was performed. This showed that both the antisera reacted with *A. urinae* AU3 and Col10a but not with the control *A. viridans* that lacked *asp* homologs in the genome (Figure 5A and data not shown). The signal was reduced when the bacteria were treated with trypsin, further indicating that the proteins are located on the surface (Figure 5B).

Genome-wide associations of LPXTG-containing proteins in *A. urinae*

To expand the information on other potential surface anchored proteins in addition to the Asp-proteins we analysed all ORFs (n=106 285) from the 25 genomes for LPXTG motifs. From this analysis we identified 289 ORFs that contained LPXTG-motifs. In order to cluster these ORFs into groups we compared the translated sequences of the

289 ORFs using all-by-all BLAST comparisons, which in turn yielded 83 521 comparisons. After applying a BLAST filter as previously described (Mashiyama et al., 2014), 9 647 sequence comparisons remained and these were selected for sequence similarity network (SNN) visualization (Figure 6A). Eleven apparent clusters were identified and these clusters were numbered and colored (Figure 6B). The functional domains that were predicted in the analysis are presented in Figure 6C. Our analysis of the 25 genomes revealed that these isolates had in median 13 genes encoding LPXTG-proteins (range 6-24). Predicted domains of these proteins included the G5 domain, which is widely found in Gram-positive bacteria such as streptococcal species (Lin et al., 2012) and has been proposed to be involved in biofilm formation of *S. aureus* (Bateman et al., 2005). Also, collagen binding domains and domains with collagen triple helix repeats were predicted. Furthermore, Mucin-binding protein domains, which play a role in adhesion of Lactobacillus species to mucin (Chatterjee et al., 2018) and domains with Rib/alpha-like repeats were predicted. Proteins with Rib and alpha repeats can be found in surface proteins of group B *Streptococcus* (Wastfelt et al., 1996).

Discussion

In this study, we aimed to describe the surface proteome of *A. urinae* and also to describe aerococcal genes encoding surface proteins. We selected the two strains Col10a and AU3 for the proteomic surface profiling using previously published protocols (Rodriguez-Ortega et al., 2006; Severin et al., 2007). A major issue with studying bacterial surface exposed proteins is contamination of intracellular proteins (Solis et al., 2010). Here we performed a subtraction of the contaminating proteins by comparing the protein abundances between the surface fraction and cellular fraction resulting in 54 surface associated proteins from both strains of which 17 were orthologues. A total of 18 genes encoding potential proteins with an LPXTG-motif were identified in the Col10a and AU3 genomes. However, only two orthologous pairs of them were detected in the surface fractions. These two proteins, denoted Asp1 and Asp2 in this study, were among the most abundant proteins on the aerococcal surface. The Asp1 and Asp2 share features with LPXTG-proteins from other bacterial

401 species (for instance streptococcal and staphylococcal species), such as a signal
402 sequence in the NH_2-terminal end, a hydrophobic membrane-spanning domain and a
403 positively charged tail in the COOH-terminal end.
404 In a recent publication by Carkaci *et al*. (Carkaci et al., 2017), the authors described
405 and compared 40 *A. urinae* and eight *A. sanguinicola* genomes of strains isolated
406 from blood, urine and heart valves. In that study, virulence genes encoding LPXTG-
407 containing proteins were not reported. The virulence-associated genes were however
408 identified using a virulence factor database that only included known virulence
409 factors. Thus, aerococcal genes with low similarity to genes in this collection, such as
410 the *asp* genes, were prone to be missed. Prior to our investigation, it was therefore
411 unclear whether all *A. urinae* isolates had genes encoding LPXTG-proteins.
412 Our results show that *A. urinae* genomes of isolates originating from both blood and
413 urine cultures, contain genes encoding LPXTG-proteins and that some of the
414 predicted functional domains of these proteins are of potential importance for
415 virulence. Our results also show that *A. urinae* has a conserved locus, LASP, which
416 comprise one or two of the *asp* genes. The organization of the LASP has a high
417 degree of similarity to the organization of the so-called *mga* regulon of *S. pyogenes*.
418 The genes in the *mga* regulon encode important streptococcal virulence factors (M
419 and M-like proteins and the C5a peptidase (Fischetti, 1989; O'Connor and Cleary,
420 1987)), which are regulated by the transcription factor Mga (Berge et al., 1998;
421 Caparon and Scott, 1987; McIver et al., 1995). Both the aerococcal and the
422 streptococcal loci comprise LPXTG-containing genes, arranged in a conserved
423 fashion, but with a variation between bacterial isolates in the number of genes present
424 and in sequence similarities. The aerococcal protein HCP, with the corresponding
425 gene positioned in the LASP upstream from the *asp* gene(s), was detected in the
426 intracellular fraction in the MS analysis. It is possible that the HCP has a similar
427 regulatory role for the LASP as Mga has for the *mga* regulon of *S. pyogenes*.
428 The function of the aerococcal Asp proteins is unknown. However, LPXTG-proteins
429 in other species are involved in processes such as bacterial adhesion, immune evasion,
430 internalization, iron acquisition and biofilm formation. Since most *A. urinae* isolates
431 are found in urine cultures, it is possible that the Asp proteins act as adhesion
432 molecules to epithelial cells in the human urinary tract or are involved in the process
433 of biofilm formation, which, mentioned previously, has been described as a feature of
434 *A. urinae* (Shannon et al., 2010). A few isolates in this study came from patients

diagnosed with infective endocarditis but no common LASP variant was observed for these isolates (data not shown). Conclusion regarding correlation between LASP variants and clinical features could not be made due to the small sample size.

Conclusion

In this study, we show for the first time that *A. urinae* has genes encoding surface proteins with an LPXTG-motif and that two of these proteins quantitatively dominate the surface. To establish a better understanding of how *A. urinae* colonize humans and cause disease, it is of great interest to explore the function of the Asp proteins of *A. urinae* further.

Acknowledgements

We acknowledge Ann-Cathrine Petersson and Gisela Hovold for important help.

Funding:

This study was supported by the Foundations of Knut and Alice Wallenberg (2016.0023), the Swedish Research Council (project 2015-02481), the Wallenberg Academy Fellow program KAW (2012.0178 and 2017.0271), the Swedish Government Funds for Clinical Research, the Royal Physiographic Society in Lund, and the foundations of Marianne and Marcus Wallenberg, Crafoord, Österlund, Lundgren and Skåne University hospital.

References

Bateman, A., Holden, M.T., Yeats, C., 2005. The G5 domain: a potential N-acetylglucosamine recognition domain involved in biofilm formation. Bioinformatics 21, 1301-1303.

Berge, A., Rasmussen, M., Björck, L., 1998. Identification of an insertion sequence located in a region encoding virulence factors of Streptococcus pyogenes. Infect Immun 66, 3449-3453.

Caparon, M.G., Scott, J.R., 1987. Identification of a gene that regulates expression of M protein, the major virulence determinant of group A streptococci. Proc Natl Acad Sci U S A 84, 8677-8681.

Carkaci, D., Hojholt, 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. Microb Pathog.

Chatterjee, M., Pushkaran, A.C., Vasudevan, A.K., Menon, K.K.N., Biswas, R., Mohan, C.G., 2018. Understanding the adhesion mechanism of a mucin binding domain from Lactobacillus fermentum and its role in enteropathogen exclusion. Int J Biol Macromol 110, 598-607.

Christensen, J.J., Jensen, I.P., Faerk, J., Kristensen, B., Skov, R., Korner, B., 1995. Bacteremia/septicemia due to Aerococcus-like organisms: report of seventeen cases. Danish ALO Study Group. Clin Infect Dis 21, 943-947.

Christensen, J.J., Korner, B., Kjaergaard, H., 1989. Aerococcus-like organism--an unnoticed urinary tract pathogen. APMIS 97, 539-546.

Cunningham, M.W., 2000. Pathogenesis of group A streptococcal infections. Clin Microbiol Rev 13, 470-511.

de Jong, M.F., Soetekouw, R., ten Kate, R.W., Veenendaal, D., 2010. Aerococcus urinae: severe and fatal bloodstream infections and endocarditis. J Clin Microbiol 48, 3445-3447.

Donlan, R.M., Costerton, J.W., 2002. Biofilms: survival mechanisms of clinically relevant microorganisms. Clin Microbiol Rev 15, 167-193.

501 Ebnother, C., Altwegg, M., Gottschalk, J., Seebach, J.D., Kronenberg, A., 2002.
502 Aerococcus urinae endocarditis: case report and review of the literature.
503 Infection 30, 310-313.
504 Fischetti, V.A., 1989. Streptococcal M protein: molecular design and biological
505 behavior. Clin Microbiol Rev 2, 285-314.
506 Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: quality assessment
507 tool for genome assemblies. Bioinformatics 29, 1072-1075.
508 Hilt, E.E., McKinley, K., Pearce, M.M., Rosenfeld, A.B., Zilliox, M.J., Mueller, E.R.,
509 Brubaker, L., Gai, X., Wolfe, A.J., Schreckenberger, P.C., 2014. Urine is not sterile:
510 use of enhanced urine culture techniques to detect resident bacterial flora in the
511 adult female bladder. J Clin Microbiol 52, 871-876.
512 Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H.,
513 Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A.,
514 Scheremetjew, M., Yong, S.Y., Lopez, R., Hunter, S., 2014. InterProScan 5: genome-
515 scale protein function classification. Bioinformatics 30, 1236-1240.
516 Kilsgard, O., Karlsson, C., Malmstrom, E., Malmstrom, J., 2016. Differential
517 compartmentalization of Streptococcus pyogenes virulence factors and host
518 protein binding properties as a mechanism for host adaptation. Int J Med
519 Microbiol 306, 504-516.
520 Kristensen, B., Nielsen, G., 1995. Endocarditis caused by Aerococcus urinae, a
521 newly recognized pathogen. Eur J Clin Microbiol Infect Dis 14, 49-51.
522 Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic
523 character of a protein. J Mol Biol 157, 105-132.
524 Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P.F., Prohaska, S.J., 2011.
525 Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC
526 Bioinformatics 12, 124.
527 Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-
528 Wheeler transform. Bioinformatics 25, 1754-1760.
529 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,
530 Abecasis, G., Durbin, R., Genome Project Data Processing, S., 2009. The Sequence
531 Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.
532 Lin, I.H., Hsu, M.T., Chang, C.H., 2012. Protein domain repetition is enriched in
533 Streptococcal cell-surface proteins. Genomics 100, 370-379.

534 Malmstrom, E., Kilsgard, O., Hauri, S., Smeds, E., Herwald, H., Malmstrom, L.,
535 Malmstrom, J., 2016. Large-scale inference of protein tissue origin in gram-
536 positive sepsis plasma using quantitative targeted proteomics. Nat Commun 7,
537 10261.
538 Malmstrom, J., Beck, M., Schmidt, A., Lange, V., Deutsch, E.W., Aebersold, R., 2009.
539 Proteome-wide cellular protein concentrations of the human pathogen
540 Leptospira interrogans. Nature 460, 762-765.
541 Mashiyama, S.T., Malabanan, M.M., Akiva, E., Bhosle, R., Branch, M.C., Hillerich, B.,
542 Jagessar, K., Kim, J., Patskovsky, Y., Seidel, R.D., Stead, M., Toro, R., Vetting, M.W.,
543 Almo, S.C., Armstrong, R.N., Babbitt, P.C., 2014. Large-scale determination of
544 sequence, structure, and function relationships in cytosolic glutathione
545 transferases across the biosphere. PLoS Biol 12, e1001843.
546 McIver, K.S., Heath, A.S., Green, B.D., Scott, J.R., 1995. Specific binding of the
547 activator Mga to promoter sequences of the emm and scpA genes in the group A
548 streptococcus. J Bacteriol 177, 6619-6624.
549 Nurk, S., Bankevich, A., Antipov, D., Gurevich, A.A., Korobeynikov, A., Lapidus, A.,
550 Prjibelski, A.D., Pyshkin, A., Sirotkin, A., Sirotkin, Y., Stepanauskas, R., Clingenpeel,
551 S.R., Woyke, T., McLean, J.S., Lasken, R., Tesler, G., Alekseyev, M.A., Pevzner, P.A.,
552 2013. Assembling single-cell genomes and mini-metagenomes from chimeric
553 MDA products. J Comput Biol 20, 714-737.
554 O'Connor, S.P., Cleary, P.P., 1987. In vivo Streptococcus pyogenes C5a peptidase
555 activity: analysis using transposon- and nitrosoguanidine-induced mutants. J
556 Infect Dis 156, 495-504.
557 Pearce, M.M., Hilt, E.E., Rosenfeld, A.B., Zilliox, M.J., Thomas-White, K., Fok, C.,
558 Kliethermes, S., Schreckenberger, P.C., Brubaker, L., Gai, X., Wolfe, A.J., 2014. The
559 female urinary microbiome: a comparison of women with and without urgency
560 urinary incontinence. MBio 5, e01283-01214.
561 Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0:
562 discriminating signal peptides from transmembrane regions. Nat Methods 8,
563 785-786.
564 Rasmussen, M., 2016. Aerococcus: an increasingly acknowledged human
565 pathogen. Clin Microbiol Infect 22, 22-27.

566 Rodriguez-Ortega, M.J., Norais, N., Bensi, G., Liberatori, S., Capo, S., Mora, M.,
567 Scarselli, M., Doro, F., Ferrari, G., Garaguso, I., Maggi, T., Neumann, A., Covre, A.,
568 Telford, J.L., Grandi, G., 2006. Characterization and identification of vaccine
569 candidate proteins through analysis of the group A Streptococcus surface
570 proteome. Nat Biotechnol 24, 191-197.
571 Schneewind, O., Missiakas, D., 2014. Sec-secretion and sortase-mediated
572 anchoring of proteins in Gram-positive bacteria. Biochim Biophys Acta 1843,
573 1687-1697.
574 Schuur, P.M., Kasteren, M.E., Sabbe, L., Vos, M.C., Janssens, M.M., Buiting, A.G.,
575 1997. Urinary tract infections with Aerococcus urinae in the south of The
576 Netherlands. Eur J Clin Microbiol Infect Dis 16, 871-875.
577 Senneby, E., Goransson, L., Weiber, S., Rasmussen, M., 2016. A population-based
578 study of aerococcal bacteraemia in the MALDI-TOF MS-era. Eur J Clin Microbiol
579 Infect Dis 35, 755-762.
580 Senneby, E., Petersson, A.C., Rasmussen, M., 2012. Clinical and microbiological
581 features of bacteraemia with Aerococcus urinae. Clin Microbiol Infect 18, 546-
582 550.
583 Senneby, E., Petersson, A.C., Rasmussen, M., 2015. Epidemiology and antibiotic
584 susceptibility of aerococci in urinary cultures. Diagn Microbiol Infect Dis 81, 149-
585 151.
586 Severin, A., Nickbarg, E., Wooters, J., Quazi, S.A., Matsuka, Y.V., Murphy, E.,
587 Moutsatsos, I.K., Zagursky, R.J., Olmsted, S.B., 2007. Proteomic analysis and
588 identification of Streptococcus pyogenes surface-associated proteins. J Bacteriol
589 189, 1514-1522.
590 Shannon, O., Morgelin, M., Rasmussen, M., 2010. Platelet activation and biofilm
591 formation by Aerococcus urinae, an endocarditis-causing pathogen. Infect
592 Immun 78, 4268-4275.
593 Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N.,
594 Schwikowski, B., Ideker, T., 2003. Cytoscape: a software environment for
595 integrated models of biomolecular interaction networks. Genome Res 13, 2498-
596 2504.
597 Smeesters, P.R., McMillan, D.J., Sriprakash, K.S., 2010. The streptococcal M
598 protein: a highly versatile molecule. Trends Microbiol 18, 275-282.

599 Solis, N., Larsen, M.R., Cordwell, S.J., 2010. Improved accuracy of cell surface
600 shaving proteomics in Staphylococcus aureus using a false-positive control.
601 Proteomics 10, 2037-2049.
602 Sunnerhagen, T., Nilson, B., Olaison, L., Rasmussen, M., 2016. Clinical and
603 microbiological features of infective endocarditis caused by aerococci. Infection
604 44, 167-173.
605 Teleman, J., Hauri, S., Malmstrom, J., 2017. Improvements in Mass Spectrometry
606 Assay Library Generation for Targeted Proteomics. J Proteome Res 16, 2384-
607 2392.
608 Teleman, J., Rost, H.L., Rosenberger, G., Schmitt, U., Malmstrom, L., Malmstrom, J.,
609 Levander, F., 2015. DIANA--algorithmic improvements for analysis of data-
610 independent acquisition MS data. Bioinformatics 31, 555-562.
611 Wastfelt, M., Stalhammar-Carlemalm, M., Delisse, A.M., Cabezon, T., Lindahl, G.,
612 1996. Identification of a family of streptococcal surface proteins with extremely
613 repetitive structure. J Biol Chem 271, 18892-18897.
614 Wattam, A.R., Davis, J.J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N.,
615 Dietrich, E.M., Disz, T., Gabbard, J.L., Gerdes, S., Henry, C.S., Kenyon, R.W., Machi,
616 D., Mao, C., Nordberg, E.K., Olsen, G.J., Murphy-Olson, D.E., Olson, R., Overbeek, R.,
617 Parrello, B., Pusch, G.D., Shukla, M., Vonstein, V., Warren, A., Xia, F., Yoo, H.,
618 Stevens, R.L., 2017. Improvements to PATRIC, the all-bacterial Bioinformatics
619 Database and Analysis Resource Center. Nucleic Acids Res 45, D535-D542.
620 Wood, D.E., Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence
621 classification using exact alignments. Genome Biol 15, R46.
622 Zhang, Z., Schwartz, S., Wagner, L., Miller, W., 2000. A greedy algorithm for
623 aligning DNA sequences. J Comput Biol 7, 203-214.
624
625
626
627
628
629
630
631

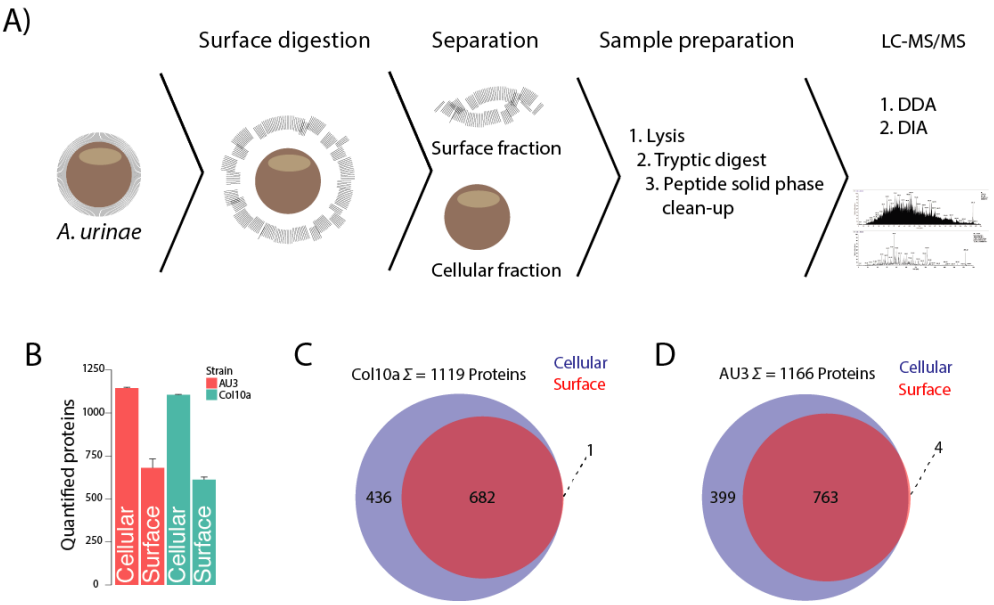


Figure 1. Proteomics of *A. urine* cellular and surface fractions. (a) Summary of the proteomics workflow for *A. urine* cellular and surface fractions. Surface proteins were released from two different *A. urine* strains using a short incubation with trypsin and the fractions separated by centrifugation. Following the sample preparation, the derived peptide samples were analysed with LC-MS/MS using both shotgun- and SWATH-MS. (b) Average numbers ± standard deviations of quantified proteins per respective strain and fractions using three replicates per condition. (c, d) Number of proteins common between fractions and per respective strain.

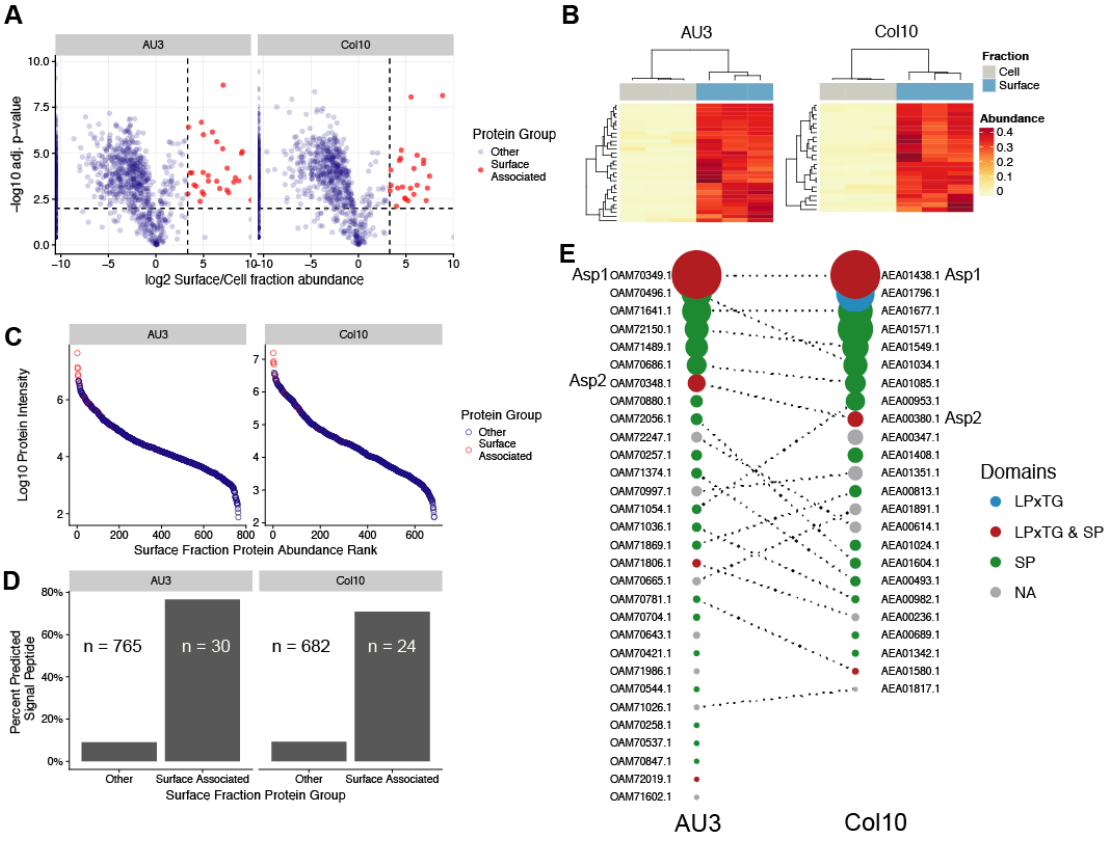
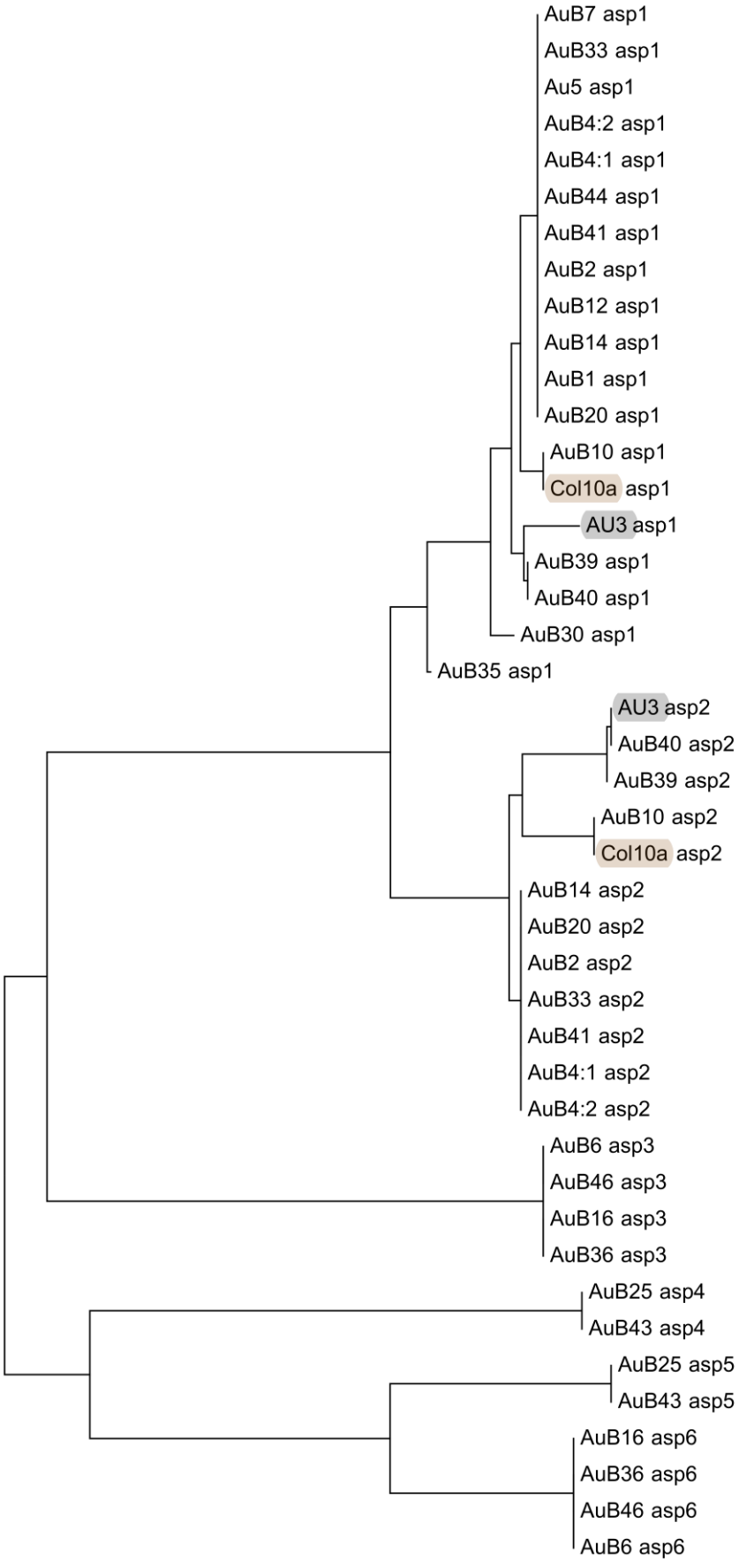


Figure 2. Proteomic analysis reveals *A. urine* surface associated proteins. (a) Volcano plots showing Hochberg adjusted –log10 p-values correlated to log2 fold changes in surface vs cell fractions of respective strain. Proteins in red represent the surface associated proteins with a fold-change ≥ 10 (vertical dotted lines) and adjusted p-value < 0.01 (horizontal dotted lines). (b) Heat map and unsupervised hierarchical clustering of surface associated proteins (red dots in (a)) across individual replicates and intensities normalized on total ion current (TIC) normalization (c). Average abundance distribution of surface fraction proteins of respective strain with the surface associated proteins marked with red. (d) The proportion of proteins with a predicted signal peptide of Surface Fraction proteins divided into Surface Associated and other proteins groups. The number (n) of proteins per protein group and strain are indicated in the plot. (e) Shows the ortholog pairs indicated as dotted lines (edges) between the surface associated proteins (nodes named with GenBank protein accession number) of the two strains. The node size is proportional to relative protein abundance per strain. Node colour represents predicted presence of protein domains SP: signal peptide and LPXTG: Gram-positive LPXTG cell wall anchor.


AuB7 asp1
AuB33 asp1
Au5 asp1
AuB4:2 asp1
AuB4:1 asp1
AuB44 asp1
AuB41 asp1
AuB2 asp1
AuB12 asp1
AuB14 asp1
AuB1 asp1
AuB20 asp1
AuB10 asp1
Col10a asp1
AU3 asp1
AuB39 asp1
AuB40 asp1
AuB30 asp1
AuB35 asp1
AU3 asp2
AuB40 asp2
AuB39 asp2
AuB10 asp2
Col10a asp2
AuB14 asp2
AuB20 asp2
AuB2 asp2
AuB33 asp2
AuB41 asp2
AuB4:1 asp2
AuB4:2 asp2
AuB6 asp3
AuB46 asp3
AuB16 asp3
AuB36 asp3
AuB25 asp4
AuB43 asp4
AuB25 asp5
AuB43 asp5
AuB16 asp6
AuB36 asp6
AuB46 asp6
AuB6 asp6
0.10

660

Figure 3.

Mid-point rooted maximum likelihood tree visualizing the clustering of the *asp* genes in the 25 isolates. The *asp* genes of Col10a (brown) and AU3 (grey) are highlighted in the tree. The scale length indicates 0.1 base substitutions per base.

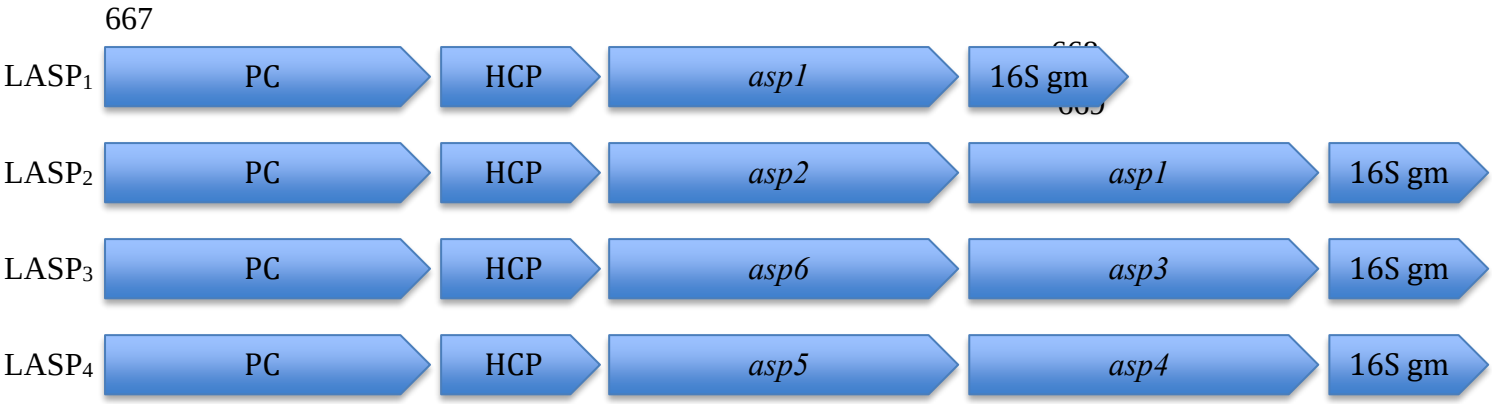
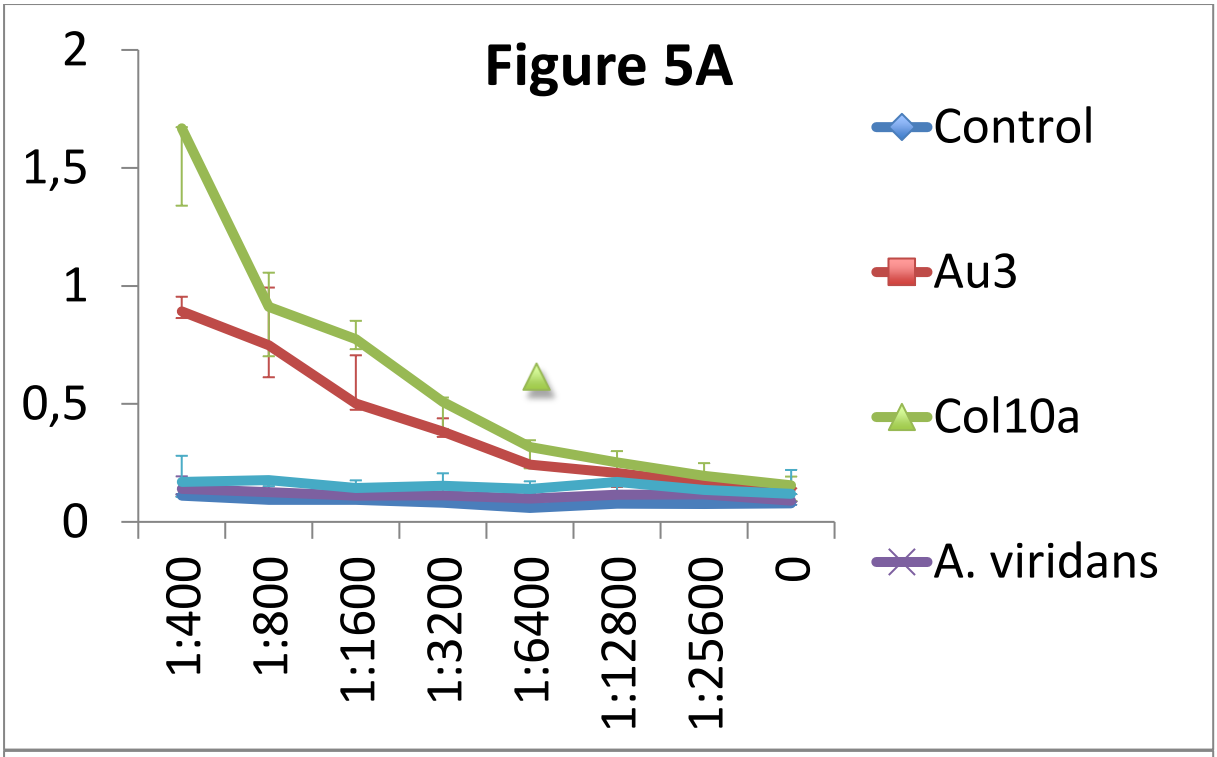


Figure 4. Schematic figure of different types of the Locus encoding Aerococcal Surface Proteins (LASP).

The LASP includes three genes enclosing the *asp* gene(s). Four different types of the LASP were identified in the *A. urinae* isolates, denoted by us as $LASP_{1-4}$. *PC*; Pyruvate carboxylase gene, *HCP*; Hypothetical cytosolic protein gene, *asp;* aerococcal surface protein gene, *16S gm;* 16S rRNA guanine-methyltransferase gene. The distance between genes and length of genes are not according to scale.



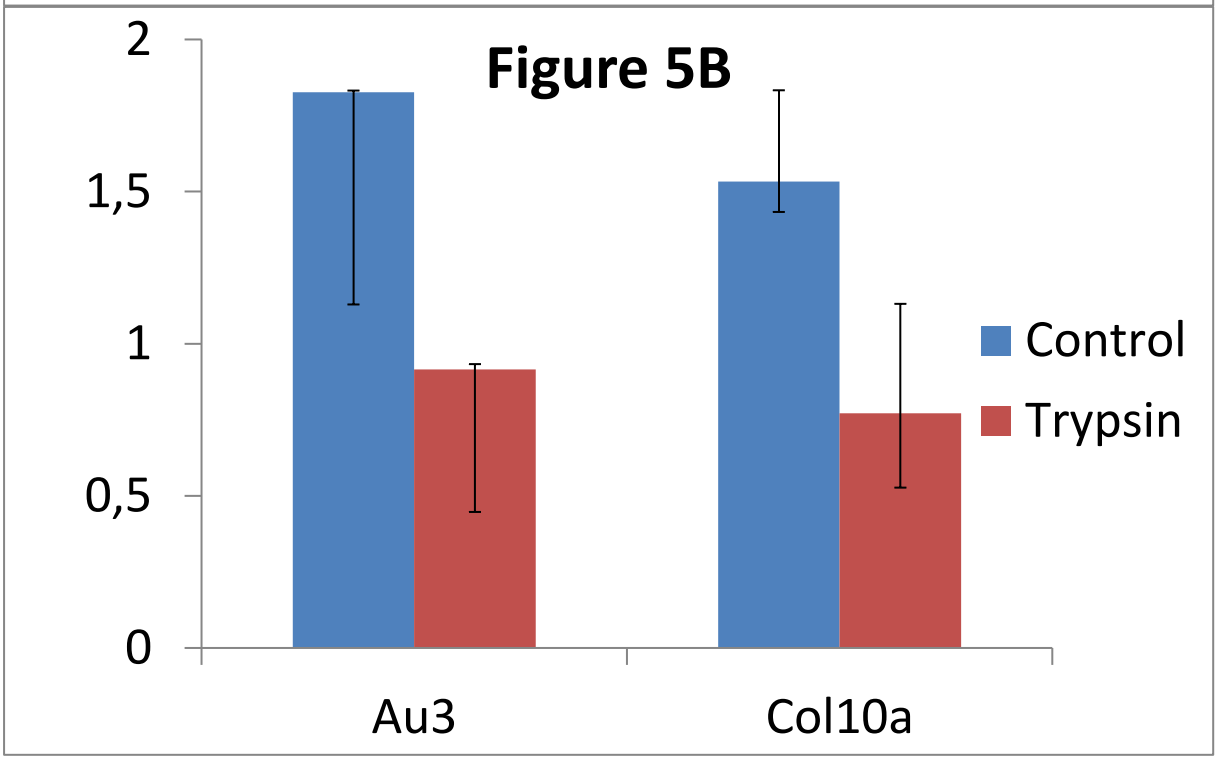


Figure 5.

Data shown as median with whiskers showing range, experiments performed in triplicate. A) Binding of serum from rabbits immunized with Asp1 was studied with ELISA. Serum from the same rabbit taken from before the immunization was used as

686 a negative control. The plate was coated with AU3, Col10a, or *A. viridans*. Dilutions
687 of rabbit serum are shown on the x axis and range from 1:400 to no serum at all. B)
688 AU3 and Col10a were treated with trypsin to reduce the amount of Asp1 on the
689 surface. AU3 and Col10a with trypsin inhibitor added were used as controls.
690

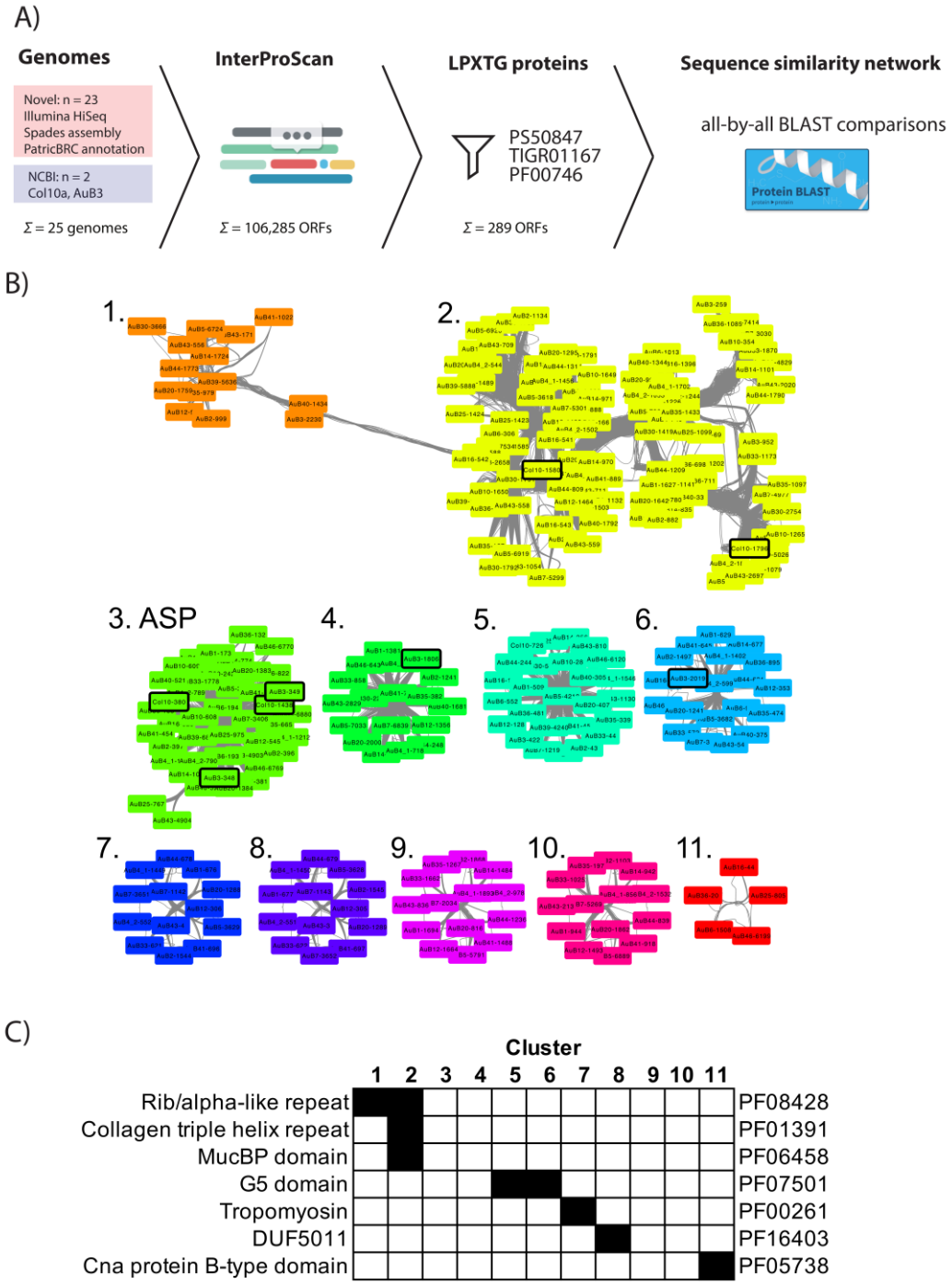


Figure 6. Sequence similarity network of *A. urinae* LPXTG proteins.

A) All *A. urinae* ORFs (n=106,285) from 25 genomes were analyzed with InterProScan to obtain predictions of functional domains. ORFs containing LPXTG motifs (n=289) were selected with the criteria of InterPro signature accessions PS50847 and/or TIGR01167 and/or PF00746. For sequence similarity analysis the

289 selected ORFs were analyzed all-by-all BLAST comparisons. B) Sequence similarity network of the 289 LPXTG ORFs with 9,647 edges each representing a BLAST hit. Eleven clusters of LPXTG ORFs are colored accordingly. Proteins that were previously defined as being surface associated in Col10a or AU3 were outlined in black. C) Predicted functional domains of the LPXTG-ORFs and in which clusters they were detected in.

Direction	Gene	
	asp1	*asp2*
Forward	GAGGAATAATAAATGGCAGACGCA TTTGTAACACCAGTA	GAFFAATAATAATAAATGGCAGTT GCTAAAGCTGAAATGTTT
Reverse	AGTAGCTGGTAATTTAGCGTTAGCG TTAGCTTTTTGAGCTTTGC	TAATTTAGCATTAGCTTTTTCAGC TTTAGCTTCTTTATC

Table 1. Primers used for expression cloning

The primers used to construct the pGEX-6p-1 plasmids containing the *asp1* and *asp2* genes were designed with an RBS.

Asp	1	2	3	4	5	6
1	72-100 %	40-53%	20-21%	12-13%	10-11%	11-12%
2	-	90-100 %	16-17%	14-15%	13%	13%
3	-	-	100 %	16%	14%	15%
4	-	-	-	100 %	12%	13%
5	-	-	-	-	100 %	52%
6	-	-	-	-	-	100 %

Table 2.

Identity scores (%) for Asp1-6 amino acid sequences.