**In vivo expression of Staphylococcus aureus virulence genes in human skin and soft tissue infections revealed by dual RNA-sequencing**

Shicheng Guo1, Michael Pulia2, Zhan Ye3, Jennifer Anderson, Thao Le, Thomas R. Fritsch, Jake Patitucci, Bradley Sullivan Matthew Hall, Sanjay K. Shukla

1.Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, WI 54449, USA

2.University of Wisconsin, School of Medicine and Public Health, Madison, WI , USA

**Introduction**: *S. aureus* is a commensal of humans and could be an opportunistic pathogen in humans and many animals. It is capable of causing a variety of diseases from a variety of skin and soft infections to severe diseases such as pneumonia, osteomyelitis, to bacteremia. A number of virulence genes including PVL have been implicated in playing a significant role in Skin and Soft Tissue infections (SSTI). However, SSTI is a complicated disease and involves a number of steps leading up to the infection and disease. The goal of this study was to determine which of the common virulence *S. aureus* genes were expressed in S. aureus recovered from the SSTI pus material or wound swabs. We hypothesize that that virulence genes that play a role during SSTI is likely to be expressed during the infection process. Here we show that a large number of genes were expressed during SSTI showing that it takes more than just toxin genes to cause SSTI.

**Methods:** Fifty two subjects, with a variety of SSTI were enrolled in the study. Patients came from Marshfield Clinic campus in Marshfield and emergency department of University of Wisconsin hospital in Madison, WI. The study was approved by the Institutional Review Boards of both Marshfield Clinic and UW Madison, WI. Each subject had given informed written consent for the study. Two bioinformatics tools, BWA (0.7.13) and Bowtie (2.2.7) were used to assemble the sequences and map the raw reads to S. aureus genomes. We focused on known virulence genes.

**Result**: Out of 53 clinical samples isolated from swab, 30 grew S. aureus of which 20 were MRSA as confirmed by mecA PCR and 10 were MSSA. cDNAs were from 19 MRSA and 8 MSSA. The average number of cDNA reads per sample was 5,049,463 and the range was from 4,464,533 to 5,717,924. Although we blasted the sequences against 12 S. aureus genomes so that we do not miss any particular gene, our data analysis here is presented from one S. aureus reference genome, the USA300. To further simplify (or expedite) the process, we focused on a select number (n=42) of well described S. aureus virulence genes first. Of the 42 gene interrogated, we detected and quantified cDNA count of 39 S. aureus genes. These were: sea, seg, sei, sek, sel, sem, seo (enterotoxin genes); tst, (toxic shock syndrome toxin gene), spa (staphylococcal protein A), fnbA and fnbB (fibronecting binding genes A and B), coa (coagulase), cna (collagen), icaA (intracellular adhesion gene A), sdrC, sdrD, sdrE (serien-aspartate repeat genes), clfA and clfB (clumping factor A and B), lukSF, lukD, lukE, (leukocidin genes), hla, hlb, hld, hlgA, hlgB, hlgC ((hemolysins), ear, sel2, sek, sec4, set16, and lpl10 (putative exotoxins). As expected, not all genes were expressed in all samples. Pus samples on average had higher cDNA counts than swab samples, most likely due to higher S. aureus burden in those samples. cDNA for following genes were not detected: seb, sec, sed, see, seh, sej, eta, etb, seg2 as these genes were not likely to be present in the strains that caused SSTI in our samples. For example, eta and etb of S. aureus are exfoliative toxin genes and are associated with scalded skin syndrome. We have summarized our finding in three tables.

Conclusion:

**Running title:**

**Background**

Complex microbial communities are an area of growing interest in biology. Metatranscriptomics allows researchers to quantify microbial gene expression in an environmental sample via high-throughput sequencing. Metatranscriptomic experiments are computationally intensive because the experiments generate a large volume of sequence data and each sequence must be compared with reference sequences from thousands of organisms.

In analysis of microbiomes, complex microbial communities that occur in many different environments, there has been a recent shift towards high-throughput methods that capture both cultured and uncultured microbial species. Although 16S ribosomal profiling is still most commonly used, there is an increasing shift towards using deeper sequencing methods, such as metagenomics and metatranscriptomics. Of these high-throughput omics methods, metatranscriptomics is advantageous in that it captures both organism abundance and species or strain level functional expression data, giving the activity of the microbiome.

**Methods**

RNA extraction and cDNA synthesis from Thao…

Total RNA was extracted with TRIzol reagent. The swab was transferred to a 2ml centrifuge tube containing 500 ul of Ramel MH broth and 1 mL RNAprotect bacterial reagent. The sample was incubated at room temperature for 5 minutes and centrifuged at maximum speed for 5 minutes to pellet the cells. The supernatant was discarded and 1ml TRIzol added to the cell pellet. The TRIzol mixture was transferred to 2ml lysing matrix tube (MP cat# 6911-050) and homogenized for 20 seconds with a homogenizer. 200 ul of Chloroform was added to the tube, and the homogenate was centrifuged at high speed for 10 minutes to separate into a clear aqueous upper layer containing RNA, an interphase, and a red lower organic layer containing DNA and proteins. RNA from the aqueous layer was precipitated with 500 ul of isopropanol and purified using Qiagen’s RNeasy protocol (QIAGEN cat #74104). Dnase treatment was then performed using Ambion’s Dnase treatment kit (Ambion cat# AM2238). RNA was converted to cDNA using the Life Technologies High-capacity cDNA reverse transcription kit (Life Technologies cat# 4368814).

From Jennifer Anderson…

Total cDNA samples were purified and concentrated using the QIAquick PCR Purification kit (QIAGEN, Valencia, CA). Libraries were prepared for sequencing using the KAPA HyperPlus Library Preparation Kit (KAPA Biosystems, Inc, Wilmington, MA) as described below. Depending on the quantity and quality of cDNA available for each sample, 1-50 ng of starting material was enzymatically fragmented for 2 min at 37 ºC. The ends of each fragment were enzymatically repaired, with 5’ phosphorylation and 3’ a-tailing. Illumina® (San Diego, CA) TruSeq universal adapters (IDT, Coralville, IA) were duplexed in solution by annealing together at a concentration of 3 µM and ligated to the end-repaired fragments. Ligation reaction products were cleaned up using Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, IN). Libraries were amplified using iTru5 and iTru7 primers (Glenn et al, 2016), to incorporate dual 8 bp indexes. PCR products were cleaned up using size selection beads as described above. The final cDNA libraries were sized using the TapeStation (Agilent Technologies, Santa Clara, CA) and quantitated by qPCR using a KAPA library quantitation kit (KAPA Biosystems). The libraries were pooled in equimolar ratios, diluted to 10 pM, and sequenced on a MiSeq instrument using the V3 150 cycle reagent kit (Illumina®).

From Shicheng Guo

Paired-end fastq files from read 1 and 2 were assembled with BWA1,2 or BOWTIE3 whose version is depended on different wrapped tools, such as DIAMOND4 and Rock-hopper5. Fastqc6 was applied to check the reads quality and Trimmomatic7 was applied for low-qualify reads remove and adaptor remove. FastQ Screen was applied for the data quality and reads mapping distribution among different genomes from human to E.coli8. PEAR(version 0.9.11)9 was applied to merge paired-end sequencing reads to single-end reads for the further analysis. SortMeRNA (version 2.1) was used to split sequencing data to rRNA reads and non-rRNA reads10. [DIAMOND](https://www.nature.com/articles/nmeth.3176) (version 0.9.30)4 was applied to alignment to meta-transcriptomes and estimated the diversity of the microbial4. Rockhopper (version 2.0.3)5 was applied for gene expression estimation which was measured as modified reads per kilobase per million mapped reads (RPKM) normalized by the upper quartile of gene expression5. Bayesian generalized linear model in R (version 3.6.0) was applied to test the association between severity and gene expression of virulence genes and P<0.05 was considered as significant association since we are discovery study with limited sample size in which multiple-test correction will discard interesting association signals.

**Results**

A total of 22 cDNA samples were sequenced using 4 MiSeq V3 flow cells (Illumina®). Total sequencing reads per sample ranged from 3,867,530 – 7,552,309 (avg 5,316,860).

We identified xx

We also tried to detected mecA positive, gender and (t008) related USA300 gene expressions. However, we did not identify any statistical significant genes.

**Discussion**

As a pilot study, we did not conducted 16S rRNA mapping and organism identification since16S rRNA is only able to provide a limited resolution for metagenome organism identification. In this study, our goal is to identify down to genus and/or species level resolution, which is where 16S rRNA struggles. As our experiences, it is unlikely to provide additional information to the meta-transcriptome activity profiles with the involvement of 16S rRNA.

**Author Contribution:**

SG conducted all the data analyses, figure and table preparation and draft preparation. JA and TL conducted next-generation sequencing. SS designed the study and supervisor the project.

**References**

1. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).

2. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).

3. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).

4. Cox, J.W. *et al.* A fast and robust protocol for metataxonomic analysis using RNAseq data. *Microbiome* **5**, 7 (2017).

5. Tjaden, B. De novo assembly of bacterial transcriptomes from RNA-seq data. *Genome Biol* **16**, 1 (2015).

6. Brown, J., Pirrung, M. & McCue, L.A. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* **33**, 3137-3139 (2017).

7. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-20 (2014).

8. Wingett, S.W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res* **7**, 1338 (2018).

9. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614-20 (2014).

10. Kopylova, E., Noe, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211-7 (2012).

**Figure legends:**

Figure 1. **Microbial composition of the samples based on strain transcriptome**. Stacked bar chart shows different composition among the samples with different clinical characteristics. Since samples are RNA, the proportion of mapped reads represents the confounded variable of microbe presence and microbial gene expression.

Figure 2. **Microbial composition of the samples based on functional annotation**. Stacked bar chart shows different composition among the samples with different clinical characteristics. Since samples are RNA, the proportion of mapped reads represents the confounded variable of microbe presence and microbial gene expression.

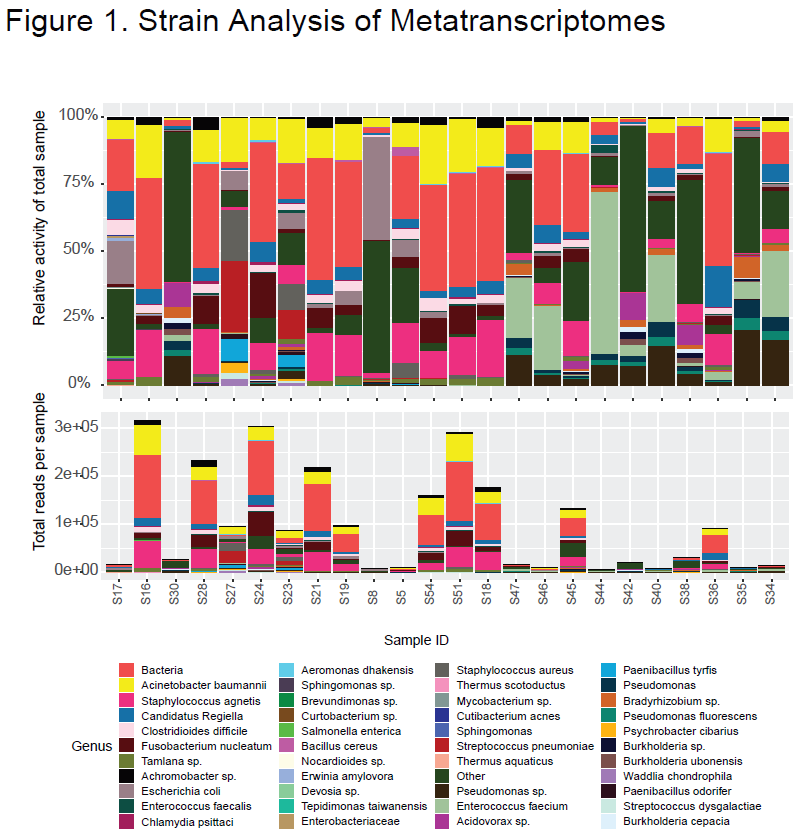
Figure 3. **Manhattan plot to show the association between USA300 gene expression and Swab or Pubs**. Swab or Pubs was recoded as binary phenotypes and then bayeslogistic was applied to detect the association with each gene transcripts measured as RPKM.

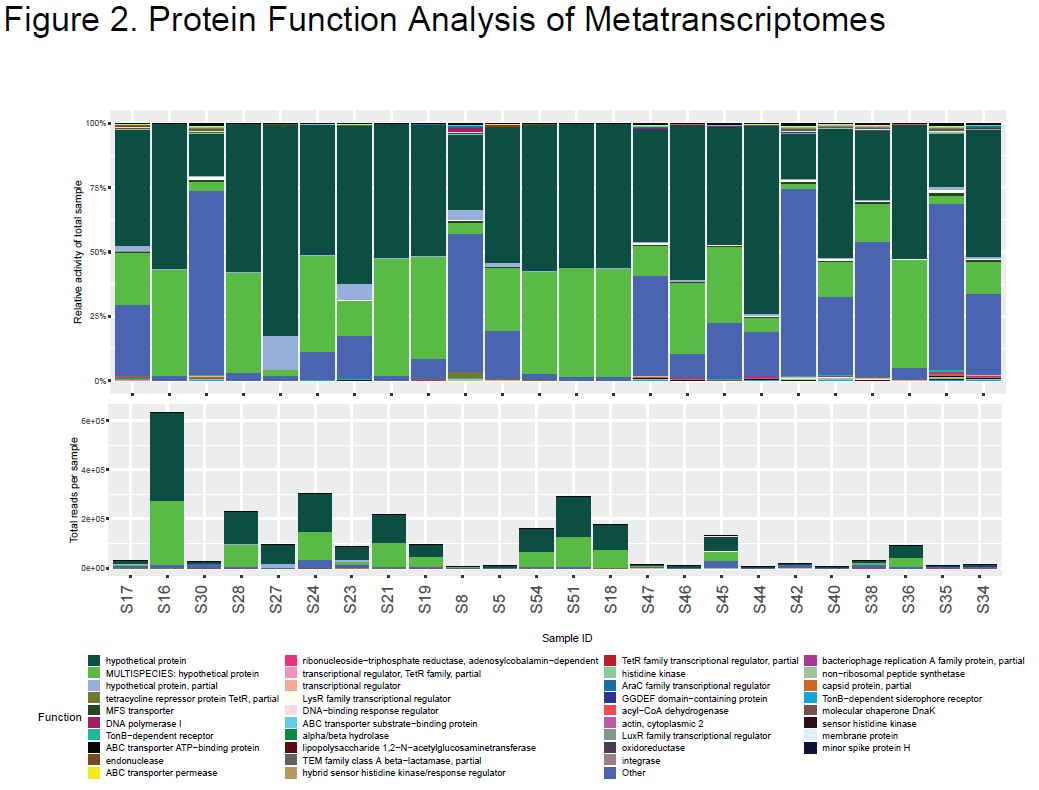
Figure 4. **Manhattan plot to show the association between USA300 gene expression and severity**. Severity was recoded as binary phenotypes and then bayeslogistic was applied to detect the association with each gene transcripts measured as RPKM.

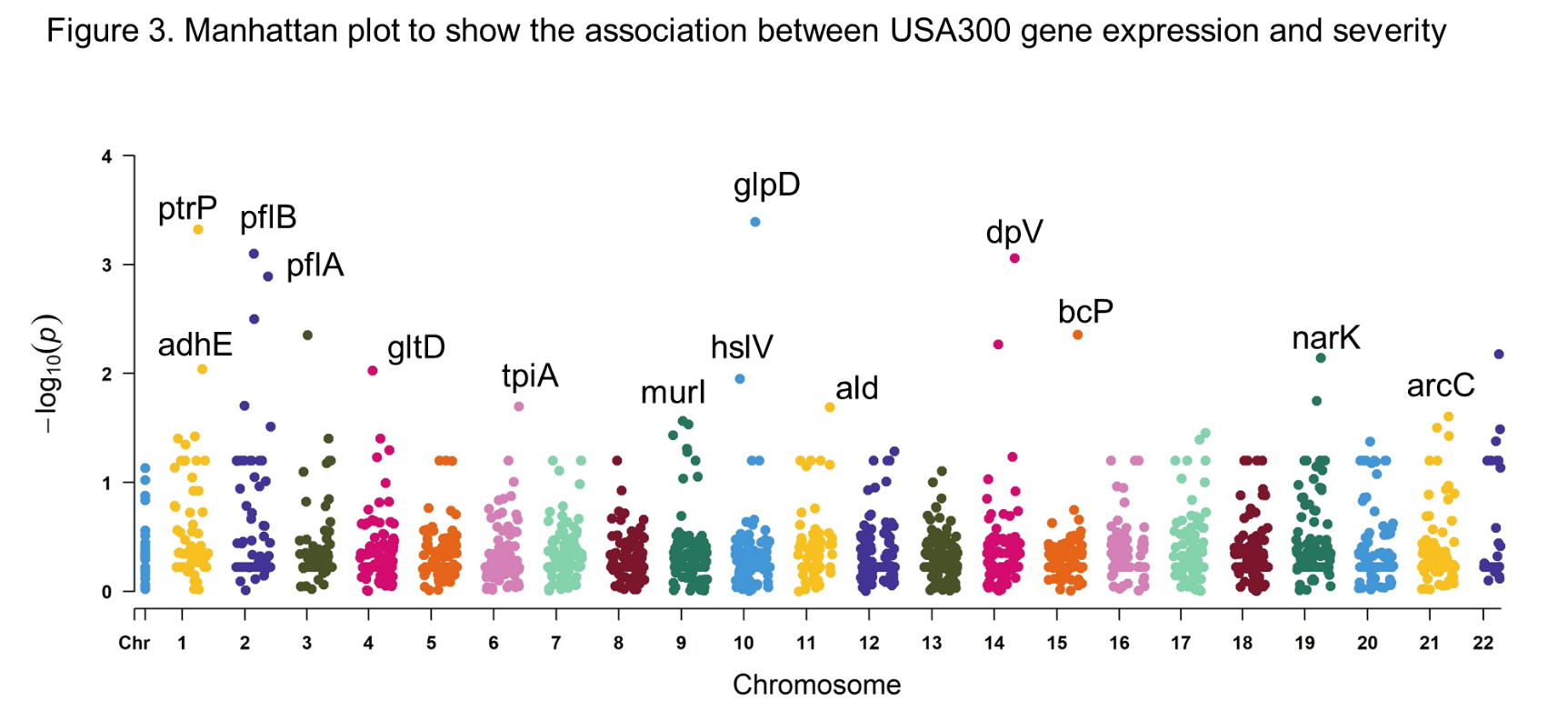
**Table legends:**

Table 1. Significant differential expressed genes between Swab and Pubs samples. Swab and pubs samples was recoded as binary phenotypes and then bayeslogistic was applied to detect the association with each gene transcripts measured as RPKM.

Table 2. Significant differential expressed genes associated with infection severity. Infection severity was recoded as binary phenotypes and then bayeslogistic was applied to detect the association with each gene transcripts measured as RPKM.







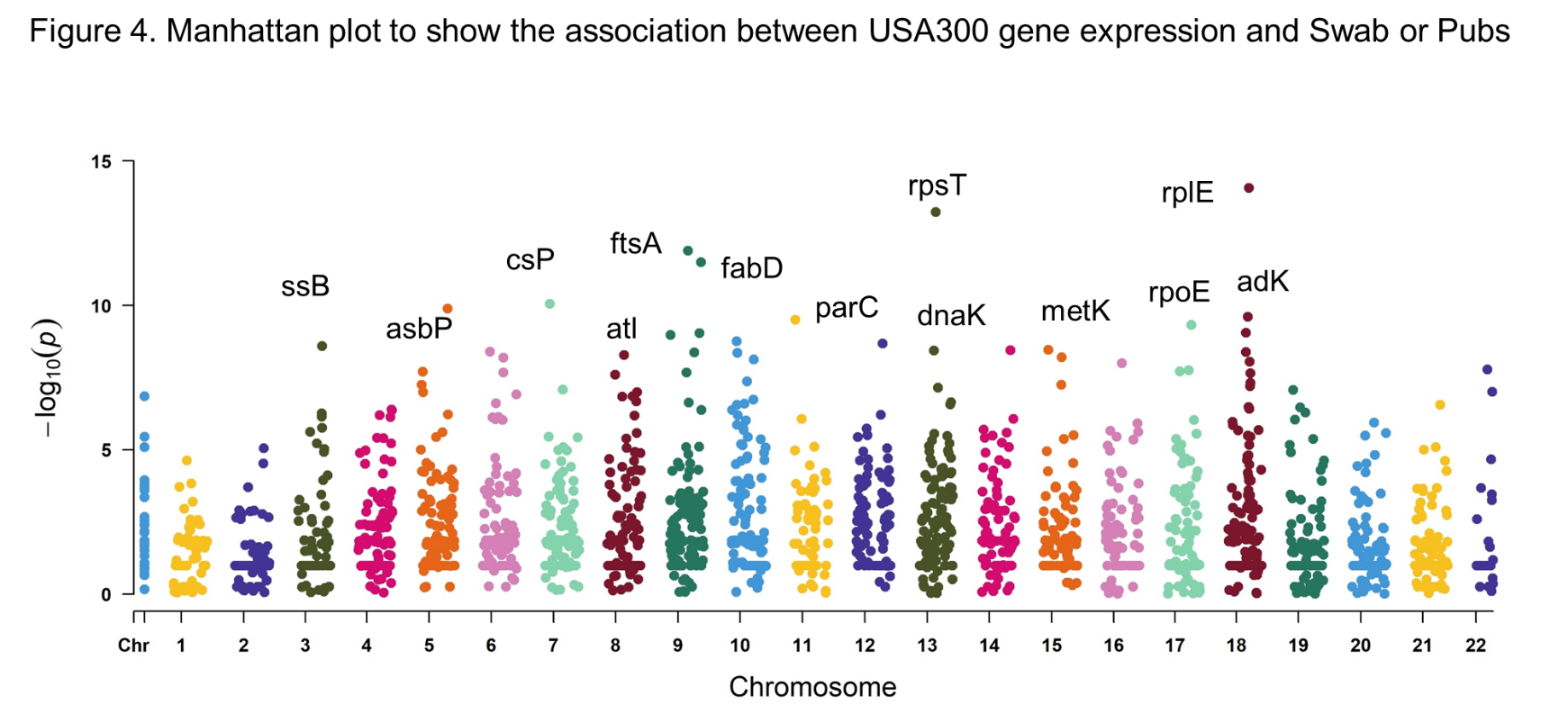


Table 1. Significant differential expressed genes between Swab and Pubs samples

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SNP | Position | P-value | Name | Protein |
| SAUSA300\_2192 | 2364918 | 8.52E-15 | rplE | 50S ribosomal protein L5 |
| SAUSA300\_1545 | 1697154 | 6.00E-14 | rpsT | 30S ribosomal protein S20 |
| SAUSA300\_1079 | 1181516 | 1.30E-12 | ftsA | cell division protein ftsA |
| SAUSA300\_1123 | 1230521 | 3.23E-12 | fabD | malonyl CoA-acyl carrier protein transacylase |
| SAUSA300\_0777 | 867458 | 8.91E-11 | - | cold shock protein |
| SAUSA300\_0618 | 691511 | 1.31E-10 | - | ABC transporter substrate-binding protein |
| SAUSA300\_2183 | 2360199 | 2.51E-10 | adk | adenylate kinase |
| SAUSA300\_1251 | 1376569 | 3.12E-10 | parC | DNA topoisomerase IV subunit A |
| SAUSA300\_2082 | 2245083 | 4.79E-10 | rpoE | DNA-directed RNA polymerase subunit delta |
| SAUSA300\_2172 | 2352953 | 8.91E-10 | rplM | 50S ribosomal protein L13 |
| SAUSA300\_1147 | 1259626 | 1.77E-09 | hslU | ATP-dependent protease ATP-binding subunit HslU |
| SAUSA300\_1448 | 1601372 | 2.12E-09 | - | FUR family transcriptional regulator |
| SAUSA300\_0367 | 419593 | 2.61E-09 | ssb | single-strand binding protein |
| SAUSA300\_1730 | 1913878 | 3.51E-09 | metK | S-adenosylmethionine synthetase |
| SAUSA300\_1700 | 1874455 | 3.63E-09 | - | polysaccharide biosynthesis protein |
| SAUSA300\_1540 | 1689411 | 3.67E-09 | dnaK | molecular chaperone DnaK |
| SAUSA300\_0672 | 746323 | 4.07E-09 | - | MarR family transcriptional regulator |
| SAUSA300\_2171 | 2352541 | 4.13E-09 | rpsI | 30S ribosomal protein S9 |
| SAUSA300\_1101 | 1204825 | 4.33E-09 | - | putative fibronectin/fibrinogen binding protein |
| SAUSA300\_1150 | 1262596 | 4.41E-09 | tsf | elongation factor Ts |
| SAUSA300\_0955 | 1044219 | 5.20E-09 | atl | autolysin |
| SAUSA300\_1783 | 1965116 | 6.30E-09 | hemE | uroporphyrinogen decarboxylase |
| SAUSA300\_0716 | 796239 | 6.41E-09 | - | ribonucleotide-diphosphate reductase subunit alpha |
| SAUSA300\_1201 | 1324118 | 7.57E-09 | glnA | glutamine synthetase- type I |
| SAUSA300\_2198 | 2367160 | 9.02E-09 | rpsC | 30S ribosomal protein S3 |
| SAUSA300\_2075 | 2236052 | 1.79E-08 | rho | transcription termination factor Rho |
| SAUSA300\_2037 | 2201245 | 1.92E-08 | - | ATP-dependent RNA helicase |
| SAUSA300\_0532 | 596508 | 2.02E-08 | fusA | elongation factor G |
| SAUSA300\_1075 | 1176055 | 2.10E-08 | pbpA | penicillin-binding protein 1 |
| SAUSA300\_0717 | 797328 | 2.12E-08 | nrdF | ribonucleotide-diphosphate reductase subunit beta |
| SAUSA300\_2204 | 2370355 | 2.27E-08 | rplC | 50S ribosomal protein L3 |
| SAUSA300\_0919 | 1010101 | 2.51E-08 | murE | UDP-N-acetylmuramoylalanyl-D-glutamate--L-lysine ligase |
| SAUSA300\_1179 | 1299149 | 4.37E-08 | - | phosphodiesterase |
| SAUSA300\_2202 | 2369430 | 4.90E-08 | rplW | 50S ribosomal protein L23 |
| SAUSA300\_1781 | 1962711 | 5.55E-08 | hemG | protoporphyrinogen oxidase |
| SAUSA300\_0528 | 592865 | 5.62E-08 | rpoC | DNA-directed RNA polymerase subunit beta- |
| SAUSA300\_2201 | 2368564 | 6.41E-08 | rplB | 50S ribosomal protein L2 |
| SAUSA300\_1556 | 1705048 | 7.14E-08 | - | GTP-binding protein YqeH |
| SAUSA300\_0835 | 915869 | 8.22E-08 | dltA | D-alanine--poly-phosphoribitol- ligase subunit 1 |
| SAUSA300\_2645 | 2868509 | 9.86E-08 | gidA | tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA |
| SAUSA300\_0996 | 1093418 | 1.01E-07 | lpdA | dihydrolipoamide dehydrogenase |
| SAUSA300\_0533 | 597909 | 1.02E-07 | tuf | elongation factor Tu |
| SAUSA300\_0757 | 845686 | 1.21E-07 | pgk | phosphoglycerate kinase |
| SAUSA300\_0016 | 22187 | 1.43E-07 | dnaB | replicative DNA helicase |

Table 2. Significant differential expressed genes associated with infection severity

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SNP | Position | P-value | Name | Protein |
| SAUSA300\_1193 | 1316578 | 0.000406 | glpD | aerobic glycerol-3-phosphate dehydrogenase |
| SAUSA300\_0139 | 159143 | 0.000476 | - | putative tetracycline resistance protein |
| SAUSA300\_0220 | 262828 | 0.000797 | pflB | formate acetyltransferase |
| SAUSA300\_1697 | 1871165 | 0.000878 | - | dipeptidase PepV |
| SAUSA300\_0264 | 315859 | 0.00129 | - | ribose transporter RbsU |
| SAUSA300\_0221 | 263606 | 0.003164 | pflA | pyruvate formate-lyase activating enzyme |
| SAUSA300\_1844 | 2004963 | 0.00441 | - | bacterioferritin comigratory protein |
| SAUSA300\_0308 | 359576 | 0.004436 | - | ABC transporter permease |
| SAUSA300\_1648 | 1809775 | 0.005404 | - | putative NADP-dependent malic enzyme |
| SAUSA300\_2641 | 2865639 | 0.006638 | - | hypothetical protein |
| SAUSA300\_2333 | 2508369 | 0.007186 | narK | nitrite extrusion protein |
| SAUSA300\_0151 | 173816 | 0.00916 | adhE | bifunctional acetaldehyde-CoA/alcohol dehydrogenase |
| SAUSA300\_0446 | 503382 | 0.009462 | gltD | glutamate synthase subunit beta |
| SAUSA300\_1146 | 1258157 | 0.011225 | hslV | ATP-dependent protease peptidase subunit |
| SAUSA300\_2319 | 2492790 | 0.017865 | - | pyridine nucleotide-disulfide oxidoreductase |
| SAUSA300\_0194 | 227381 | 0.019798 | - | sucrose-specific PTS transporter protein |
| SAUSA300\_0758 | 846569 | 0.020148 | tpiA | triosephosphate isomerase |
| SAUSA300\_1331 | 1492510 | 0.020431 | ald | alanine dehydrogenase |
| SAUSA300\_2567 | 2778192 | 0.024872 | arcC | carbamate kinase |
| SAUSA300\_1049 | 1149076 | 0.027343 | murI | glutamate racemase |
| SAUSA300\_1070 | 1169651 | 0.029487 | - | hypothetical protein |
| SAUSA300\_0274 | 326169 | 0.030768 | - | hypothetical protein |
| SAUSA300\_2534 | 2734899 | 0.031492 | panB | 3-methyl-2-oxobutanoate hydroxymethyltransferase |
| SAUSA300\_2646 | 2870453 | 0.032624 | trmE | tRNA modification GTPase TrmE |
| SAUSA300\_2106 | 2276386 | 0.035089 | - | putative transcriptional regulator |
| SAUSA300\_1014 | 1112387 | 0.036854 | pyc | pyruvate carboxylase |
| SAUSA300\_2569 | 2780671 | 0.037543 | arcB | ornithine carbamoyltransferase |
| SAUSA300\_0126 | 145681 | 0.037797 | - | hypothetical protein |
| SAUSA300\_0387 | 438318 | 0.039498 | pbuX | xanthine permease |
| SAUSA300\_0073 | 81947 | 0.03955 | - | peptide ABC transporter peptide-binding protein |
| SAUSA300\_0473 | 532532 | 0.039617 | purR | pur operon repressor |
| SAUSA300\_2090 | 2254266 | 0.04058 | deoC | deoxyribose-phosphate aldolase |
| SAUSA300\_2627 | 2854156 | 0.041829 | - | 2-oxoglutarate/malate translocator |
| SAUSA300\_2404 | 2588435 | 0.042137 | - | hypothetical protein |
| SAUSA300\_0099 | 109979 | 0.045048 | plc | 1-phosphatidylinositol phosphodiesterase |
| SAUSA300\_1063 | 1163534 | 0.049063 | arcC | carbamate kinase |