Characteristics of Publicly-Available *Staphylococcus aureus* Genomes

Nir Gilad[1], Michal Ziv-Ukelson[1], Vered Chalifa-Caspi[1] and Jacob Moran-Gilad[1,2,3]

[1]Ben-Gurion University of the Negev, Beer Sheva, Israel; [2]Public Health Services, Ministry of Health, Jerusalem, Israel; [3]ESCMID Study Group for Molecular Diagnostics

Introduction

- Genome depositions of important pathogens such as *S. aureus* are rapidly accumulating in public domain databases such as NCBI and ENA.
- Publicly available genomes could prove useful for development and evaluation of bioinformatics tools as well as cross-sectional microbiological studies.
- Concerns have been raised regarding the quality and usability of publicly deposited genomic data for many microorganisms.
- We thus sought to determine the characteristics of publicly available *S. aureus* genomes.

Methods

- *S. aureus* genome assemblies were downloaded from NCBI and a local database was created.
- First, genomes were checked for quality using QUAST, and subsequently analysed using an in house *S. aureus* pipeline (**Fig. 1**).
- Genus and species identification (ID), inferred antimicrobial susceptibility testing (AST) results, and presence of virulence factors were determined by performing BLAST analyses as well as *in silico* PCR.
- The target sequences for resistance and virulence were derived from published literature as well as public databases such as The Comprehensive Antibiotic Resistance Database (CARD), consisting of >3,000 genes in total.
- Typing was performed by extracting the *spa* locus as well as the 7 MLST gene loci and querying against the PubMLST and *spa* databases.
- cgMLST analysis and construction of minimum spanning trees was performed using SeqSphere v.3.0.1 (Ridom GmbH, Munster, Germany).

Results

- The initial analysis is summarised in **Fig. 1**.
- Overall, 4,262 genomes were downloaded, of which 32 (0.76%) samples were excluded after QC due to low N50 values, high L50 values and contig size.
- The remaining genomes all appeared to belong to the *Staphylococcus* genus.
- Analysis using the *nuc* gene marker for *S. aureus* revealed that 20 (0.47%) samples belonged to species other than *S. aureus*.
- The majority (3,876, 92%) of the remaining genomes were methicillin-resistant *S. aureus* (MRSA).
- SCC*mec* typing could be performed in 3,530 (91.9%) of MRSA samples, yielding types I, II, III and IV in 3.99%, 55.38%, 0.99% and 30.88% of typeable samples, respectively.
- Only 31 (0.8%) of the MRSA genomes were tagged as MRSA when deposited, of which 2 (6.45%) appeared *mecA* and SCC*mec* negative.
- All 8 samples tagged as MSSA were *mecA* negative.
- Of only 9 samples tagged with MLST results, the ST designation was correct in 100%.

Contact details:
Prof. Jacob Moran-Gilad
giladko@post.bgu.ac.il

Fig 1. *S. aureus* pipeline outputs

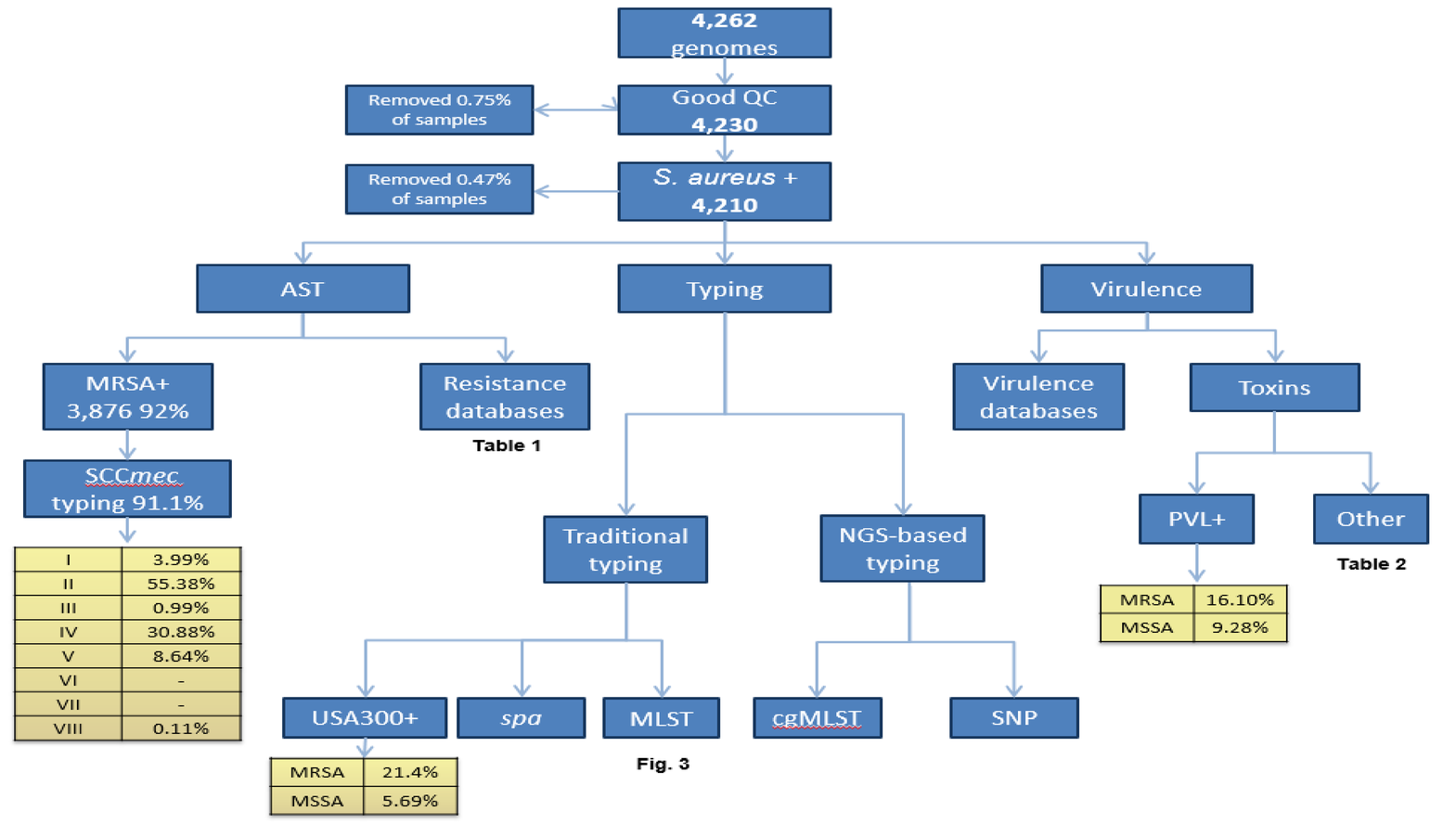


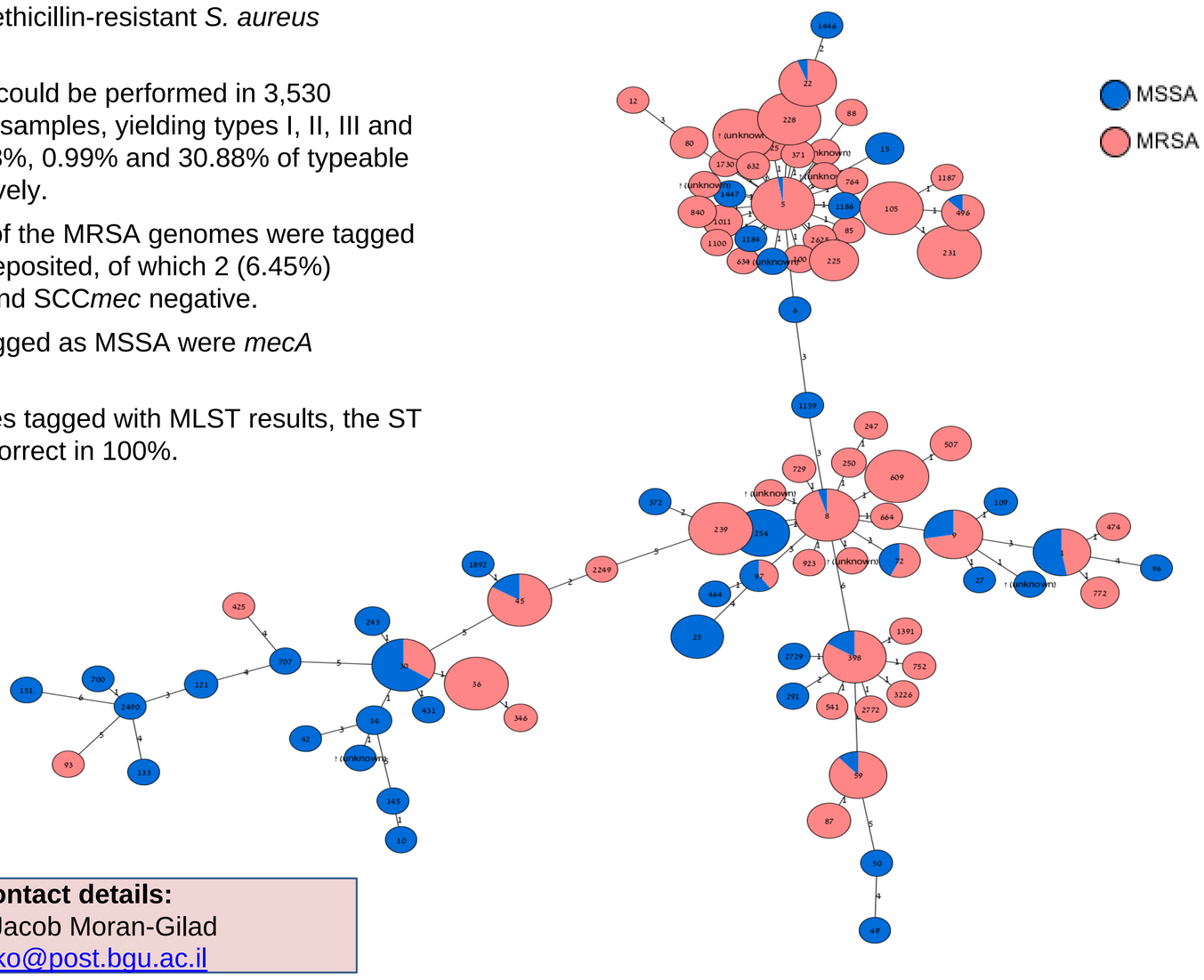
Table 1. Inferred AST results

Resistance determinant	MRSA	MSSA
ermA	60.24%	8.38%
ermC	5.91%	0.00%
ermB	0.98%	0.00%
blaZ	53.30%	0.60%
APH(3')-IIIa	25.95%	0.60%
AAC(6')-Ie-APH(2'')-Ia	13.49%	0.30%
aad(6)	3.72%	-
ANT(6)-Ia	0.28%	-
tetK	8.77%	-
qacA	1.57%	0.60%
cat	0.13%	1.20%

Table 2. Inferred toxins

Virulence determinant	MRSA	MSSA
Enterotoxin type A	5.57%	14.67%
Enterotoxin type B	4.72%	5.69%
Enterotoxin type C 1	0.72%	1.80%
Enterotoxin type C 2	1.03%	2.99%
Enterotoxin type C 3	0.72%	2.10%
Enterotoxin type D	11.48%	2.99%
Enterotoxin type E	-	-
Enterotoxin type H	0.31%	5.09%
Enterotoxin J	0.31%	-
Toxic shock syndrome toxin 1	1.47%	8.08%
Exfoliative toxin A	0.03%	1.50%
Exfoliative toxin B	0.00%	-
Exfoliative toxin D 1	0.13%	4.49%
Exfoliative toxin D 2	0.00%	0.30%
Arginine ornithine antiporter	16.69%	-
Arginine deaminase	16.59%	-
Accessory gene regulator B	23.35%	24.55%
Accessory gene regulator C	37.36%	60.48%
Accessory gene regulator D	0.34%	2.69%

Fig 3. Phylogeny per MLST



Conclusions

- Analysis of over 4,000 *S. aureus* genome assemblies available through the public domain showed an overall good quality.
- The corresponding rate of erroneous genus, species and MRSA classifications was low.
- This suggests that publicly available *S. aureus* datasets could be safely used for developing and validating *S. aureus* bioinformatics pipelines for clinical and public health usage.
- Systematic analysis of important microbiological characteristics of *S. aureus* revealed that publicly available datasets are significantly enriched with samples of MRSA and PVL-positive samples compared to natural epidemiology.
- This selection bias most probably reflects the medical and research focus in this field.
- The public database should further be assessed with NGS-based typing approaches.