Ensuring backwards compatibility: traditional genotyping efforts in the era of whole genome sequencing

S. Bletz[1], A. Mellmann[1], J. Rothgänger[2] and D. Harmsen[3]

1) *Institute of Hygiene, University Hospital Münster,* 2) *Ridom GmbH and* 3) *Department of Periodontology, University Hospital Münster, Münster, Germany*

Abstract

When using next-generation whole genome sequencing (WGS), extraction of spa types from WGS data is essential for backwards compatibility with Sanger sequencing-based spa typing of methicillin-resistant *Staphylococcus aureus* (MRSA). We evaluated WGS-based spa typing with a 2×250 bp protocol in a diverse collection of 423 MRSA isolates using two pipelines that executed sequence quality-trimming and *de novo* assembly before spa typing. The SeqSphere$^+$ pipeline correctly typed 419 isolates (99.1%) whereas the CLCbio pipeline succeeded in 249 isolates (58.9%). In summary, WGS combined with an optimized *de novo* assembly enables nearly full compatibility with Sanger sequencing-based spa typing data.

Keywords: Backwards compatibility, methicillin-resistant *Staphylococcus aureus*, spa typing, *Staphylococcus aureus*, whole genome sequencing

Original Submission: 9 September 2014; **Revised Submission:** 10 October 2014; **Accepted:** 13 October 2014

Editor: D. Raoult

Article published online: 25 November 2014

Corresponding author: A. Mellmann, Institut für Hygiene, Universitätsklinikum Münster, Robert-Koch-Strasse 41, 48149 Münster, Germany

E-mail: mellmann@uni-muenster.de

S. Bletz and A. Mellmann equally contributing to this work

Methicillin-resistant *Staphylococcus aureus* (MRSA) is one of the major multi-resistant nosocomial pathogens worldwide [1,2]. To prevent transmission, genotypic fingerprinting is essential to detect and subsequently terminate transmission routes [3]. In the last decade, Sanger sequencing-based typing of the repeat region of the *S. aureus* protein A gene (spa typing) emerged as a rapid and discriminatory tool to detect outbreaks and to elucidate the molecular epidemiology of MRSA [4–6]. The global success of spa typing was fostered by the ease of synchronization with the SpaServer (http://www.spaserver.ridom.de) assuring a unique typing nomenclature [6]. Today, the SpaServer hosts over 300 000 strains and their spa type. However, Sanger sequencing is increasingly being replaced by the development of next-generation sequencing that enables whole genome sequencing (WGS) within days and delivers both genotypic fingerprinting on whole genome level and resistance and virulence profiling [7]. To enable backwards compatibility with the large amount of historical spa typing data, extraction of spa types from WGS data is essential. In principle, extraction of spa types is possible [8], however, there is no systematic study to evaluate WGS-based spa typing in comparison to traditional Sanger sequencing.

Here, we systematically determined spa types using WGS from all MRSA that were detected during a 6-month period at the University Hospital Münster, Germany, and compared the results with Sanger sequencing, the reference method of this study, which is routinely performed in our laboratory for infection control purposes [9].

In total, 423 MRSA were isolated from clinical samples or screening swabs during the study period (15 October 2013 to 15 April 2014) and immediately spa typed using Sanger sequencing as previously described [9]. Simultaneously, these isolates were subjected to WGS on a MiSeq instrument (Illumina, San Diego, CA, USA). A single colony was therefore inoculated in nutrient broth (Heipha, Eppelheim, Germany) and incubated overnight at 37°C. Genomic DNA was purified using the MagAttract HMW DNA Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions with addition of 120 U Lysostaphin (Sigma, Taufkirchen, Germany) to lyse cells. One nanogram of genomic DNA was introduced into library preparation with the Nextera XT DNA Sample preparation Kit (Illumina) and paired-end sequenced with the MiSeq Reagent Kit v2, 2×250 bp (Illumina) with an average insert size of 400 bp. Libraries were scaled to reach 100-fold sequencing coverage. Only genomes that contained ≥95% of the 1861 MLST$^+$ target genes, with a minimum of ten times local coverage set as the quality threshold [8], were analysed further; otherwise genome sequencing was repeated. Resulting FASTQ files were quality-trimmed and *de novo* assembled using two different analysis pipelines. One pipeline

http://dx.doi.org/10.1016/j.cmi.2014.11.005

("SeqSphere$^+$") used the VELVET assembler [10] integrated in the RIDOM SEQSPHERE$^+$ software version 2.1 (Ridom GmbH, Münster, Germany). Here, reads were trimmed at their 5′ and 3′ ends until an average base quality of 30 was reached in a window of 20 bases, and the assembly was performed with VELVET version 1.1.04 using optimized k-mer size and coverage cut-off values based on the average length of contigs with >1000 bp. The other pipeline ("CLCbio") was based on the CLCBIO GENOMICS WORKBENCH software version 6.5 (Qiagen) as recently described [8]. All resulting assembly files (ACE file format) were analysed in RIDOM SEQSPHERE$^+$ to determine the spa type and the multilocus sequence type (MLST) and clonal complex (CC). Raw reads are deposited at the European Nucleotide Archive under project number PRJEB7089.

Sanger sequencing of the 423 isolates resulted in 75 different spa types, showing a wide-ranging length distribution of 2–19 repeats (Fig. 1). Clustering of the 75 spa types using the BURP (based upon repeat pattern) algorithm with default values [11] further demonstrated the diversity of this collection: BURP clustering resulted in seven spa clonal complexes and 12 singletons (four spa-types with fewer than five repeats were excluded) (Fig. 2) making this data set an appropriate data set to evaluate WGS-based spa typing.

After WGS quality control, the averages of successfully extracted MLST$^+$ target genes were 98.2% for the SeqSphere$^+$ and 98.5% for the CLCbio pipeline. Extraction of the correct spa type was initially possible in 411 of the 423 isolates using the SeqSphere$^+$ pipeline whereas the CLCbio pipeline succeeded only in 249 isolates (58.9%). Detailed analyses were therefore only performed with data from the SeqSphere$^+$ pipeline, where we repeated Sanger sequencing from the same DNA that had been used for WGS. This yielded results that were identical to WGS-based spa typing in seven isolates (00383, 00408, 00701, 00812, 00902, 00963 and 01049, see Supporting information, Table S1), indicating a possible interchange of samples previously. As sample 00563 gave the identical result after repeated Sanger sequencing, we re-cultured this isolate from − 70°C before repeated Sanger sequencing from eight different colonies per isolate as cultures from one specimen can contain different clones [12]. Indeed, two of these eight colonies showed the identical spa type as the one from WGS (t011), corroborating this finding. These additional analyses finally led to 419 (99.1%) correct spa types. Only isolates 00913, 00970 and 01377 showed an incomplete assembly of the *spa* repeat region, and isolate 00774 was misassembled: three repeats (marked with xx) were missed in the repeat pattern 26-23-23-20-13-31-29-17-31-29-17-xx-xx-xx-25-17-25-16-28.

Our results demonstrate that even with similar assembly strategies—as both pipelines rely on de Bruin graph assemblers [13,14]—the choice of parameters is crucial for a high-quality assembly of repeat regions that are prone to misassembling [15]. We selected the 2×250 bp read length protocol to balance the generation of long-read data with a short instrument run time as timely typing information is crucial in the nosocomial setting to implement focused infection control measures [9]. Depending on the sequencing platform nowadays, even longer reads are achievable but they are frequently

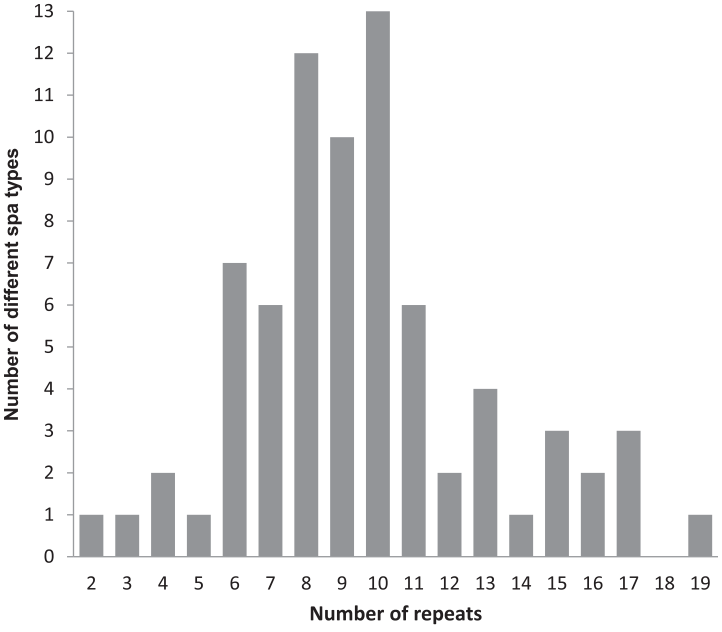

FIG. 1. Histogram of the repeat distribution and number of different spa types based on the 75 different spa types from the 423 isolates.