JCM
Journals.ASM.org

Comparing Whole-Genome Sequencing with Sanger Sequencing for *spa* Typing of Methicillin-Resistant *Staphylococcus aureus*

Mette Damkjær Bartels,[a,c] Andreas Petersen,[b] Peder Worning,[c] Jesper Boye Nielsen,[a,c] Hanna Larner-Svensson,[c]* Helle Krogh Johansen,[d,e] Leif Percival Andersen,[f] Jens Otto Jarløv,[g] Kit Boye,[a] Anders Rhod Larsen,[b] Henrik Westh[a,c,h]

Department of Clinical Microbiology, Hvidovre Hospital, University of Copenhagen, Hvidovre, Denmark[a]; Microbiology and Infection Control, Statens Serum Institut, Copenhagen, Denmark[b]; MRSA Knowledge Center, Hvidovre Hospital, University of Copenhagen, Hvidovre, Denmark[c]; Department of Clinical Microbiology, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark[d]; Novo Nordisk Center for Biosustainability, The Technical University of Copenhagen, Hørsholm, Denmark[e]; Department of Infection Control, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark[f]; Department of Clinical Microbiology, Herlev Hospital, University of Copenhagen, Herlev, Denmark[g]; Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark[h]

***spa* typing of methicillin-resistant *Staphylococcus aureus* (MRSA) has traditionally been done by PCR amplification and Sanger sequencing of the *spa* repeat region. At Hvidovre Hospital, Denmark, whole-genome sequencing (WGS) of all MRSA isolates has been performed routinely since January 2013, and an in-house analysis pipeline determines the *spa* types. Due to national surveillance, all MRSA isolates are sent to Statens Serum Institut, where the *spa* type is determined by PCR and Sanger sequencing. The purpose of this study was to evaluate the reliability of the *spa* types obtained by 150-bp paired-end Illumina WGS. MRSA isolates from new MRSA patients in 2013 ($n = 699$) in the capital region of Denmark were included. We found a 97% agreement between *spa* types obtained by the two methods. All isolates achieved a *spa* type by both methods. Nineteen isolates differed in *spa* types by the two methods, in most cases due to the lack of 24-bp repeats in the whole-genome-sequenced isolates. These related but incorrect *spa* types should have no consequence in outbreak investigations, since all epidemiologically linked isolates, regardless of *spa* type, will be included in the single nucleotide polymorphism (SNP) analysis. This will reveal the close relatedness of the *spa* types. In conclusion, our data show that WGS is a reliable method to determine the *spa* type of MRSA.**

Methicillin-resistant *Staphylococcus aureus* (MRSA) remains endemic in many hospitals (1), and the prevalence in the community has been increasing worldwide for the last decades (2). *spa* typing has been widely used for typing of MRSA isolates to investigate outbreaks and local epidemiology. The introduction of the Ridom StaphType software program in 2003 (3) has made *spa* typing an excellent tool to compare MRSA types across regions and countries, since it has proven to be a typing scheme with high interlaboratory reproducibility (4). *spa* typing is the sequencing of the 24-bp (21 to 30 bp) repeat region of the *Staphylococcus* protein A gene (*spa*). Repeats are assigned a numerical code, and the *spa* type is determined from the order of these repeats. By 9 July 2014, 647 repeats and 13,857 different *spa* types were registered in the SpaServer database (http://spaserver.ridom.de). The traditional method of *spa* typing is by PCR amplification of the *spa* repeat region followed by Sanger sequencing (3).

At the Department of Clinical Microbiology at Hvidovre Hospital, Hvidovre, Denmark, whole-genome sequencing (WGS) has been performed since January 2013 on all MRSA isolates from the Capital Region of Denmark. WGS was introduced to improve outbreak investigations of MRSA and to gain a better and more thorough understanding of the spread and evolution of MRSA. MRSA genome sequences are analyzed for *spa* types, multilocus sequence types (MLST), and the presence or absence of PVL genes (Panton-Valentine leukocidin), and the findings are used as a common nomenclature to share with clinicians in patient records. For infection control, single nucleotide polymorphism (SNP) analysis is used routinely to compare MRSA isolates with identical or related *spa* types and isolates belonging to the same clonal complex (CC) from patients suspected to be part of an MRSA outbreak.

In Denmark, MRSA from infections and carriage has been notifiable since 2006. MRSA isolates from the departments of clinical microbiology in Denmark are sent for national surveillance at the Staphylococcal Laboratory at Statens Serum Institut, where isolates are *spa* typed by PCR amplification and Sanger sequencing.

The purpose of the present study was to evaluate the reliability of MRSA *spa* typing by WGS. This was done by comparing the *spa* typing results of WGS with Sanger sequencing results. As far as we know, this is the first study comparing *spa* typing by Sanger sequencing and WGS.

MATERIALS AND METHODS

The Capital Region of Denmark has three departments of clinical microbiology that perform microbiology analysis for 12 hospitals and the general-practice health care services of the region's 1.72 million inhabitants. One MRSA isolate per patient was confirmed, with an in-house multiplex real-time PCR assay detecting the presence of *nuc*, *femA*, *mecA*, and *mecC* (data not shown). All confirmed MRSA isolates from the Capital Region were whole genome sequenced on a MiSeq instrument (Illumina, San Diego, CA, USA) at the Department of Clinical Microbiology at Hvidovre

Received 11 July 2014 **Returned for modification** 11 August 2014
Accepted 4 October 2014
Published ahead of print 8 October 2014
Editor: S. S. Richter
Address correspondence to Mette Damkjær Bartels, mette.damkjaer@dadlnet.dk.
* Present address: Hanna Larner-Svensson, Denmark National Biobank, Statens Serum Institut, Copenhagen, Denmark.

doi:10.1128/JCM.01979-14

TABLE 1 Isolates with different *spa* types by WGS and Sanger sequencing[a]

spa type/ST by WGS	*spa* type by Sanger sequencing	*spa* repeats by WGS	*spa* repeats by Sanger sequencing
t015/ST45	t026	08-16-02-16-34-13-17-34-16-34	08-16-34
t032/ST22	t1249	26-23-23-13-23-31-29-17-31-29-17-25-17-25-16-28	26-23-23-13-23-23-13-23-31-29-17-31-29-17-25-17-25-16-28
t086/ST88	t690	07-12-21-17-13-13-13-34-34-34-33-34	07-12-21-17-13-13-34-34-34-33-34
t186/ST88	t690	07-12-21-17-13-13-34-34-33-34	07-12-21-17-13-13-34-34-34-33-34
t304/ST6	t197	11-10-21-17-34-24-34-22-25	11-10-34-24-34-22-25
t355/ST152	t595	07-56-12-17-16-16-33-31-57-12	07-56-12-17-16-16-33-31-57-31-57-12
t359/ST97	t267	07-23-12-21-17-34-34-33-34	07-23-12-21-17-34-34-34-33-34
t630/NA	t304	08-16-02-16-34-17-34-16-34	11-10-21-17-34-24-34-22-25
t670/ST22	t5177	26-23-23-13-23-29-17-25-17-25-16-28	26-23-23-13-23-29-17-31-29-17-31-29-17-25-17-25-16-28
t718/ST22	t1249	26-23-23-23-13-23-31-29-17-31-29-17-25-17-25-16-28	26-23-23-13-23-23-13-23-31-29-17-31-29-17-25-17-25-16-28
t728/ST45[b]	t015	08-16-34-16-34	08-16-02-16-34-13-17-34-16-34
t790/ST22	t022	26-23-13-23-31-29-17-25-17-25-16-28	26-23-13-23-31-29-17-31-29-17-25-17-25-16-28
t934/ST80	t1198	07-23-12-34-34-34-34-33-34	07-23-12-34-34-34-34-34-33-34
t1028/ST78	t237	07-34-33-34	07-34-34-33-34
t4699/ST93	t3949	11-17-16-16-25	11-17-23-17-17-17-16-16-25
t5090/ST130	t843	04-82-17-16-17	04-82-17-25-17-25-25-16-17
t5608/ST5	t002	26-23-17-34-17-13-17-20-17-12-17-16	26-23-17-34-17-20-17-12-17-16
t3119/ST8	t1774	11-19-19-19-12-05-17-34-24-34-22-25	11-19-19-12-05-17-34-24-34-22-25

[a] ST, sequence type; NA, not available.
[b] Two isolates of t728. Both were t015 by Sanger sequencing.

Hospital. The current workload was a 4-day setup. DNA concentrations were normalized using a Qubit fluorometer (Invitrogen, United Kingdom). Libraries were made using a Nextera XT DNA sample preparation kit (Illumina, USA), and genomes were multiplexed to 24 isolates per run and sequenced using 2 × 150-bp paired-end reads. Genomes were assembled using the software programs Velvet v1.0.11 (5) and VelvetOptimiser v2.1.7, with hash (kmer) size and coverage parameters optimized to give the highest N_{50} value. Variants were called using the SAMtools v0.1.12 mpileup command with options -M0-Q30 -q30-o40 -e20-h100 -m2 -D -S (6). An in-house analysis pipeline has been developed to analyze MRSA genomes for the *mecA*, *mecC*, *nuc*, and *ccr* genes, *spa* type, MLST, *dru* types, PVL, and ACME (the arginine catabolic mobile element). *spa* types were called from the assembled contigs by comparison to the published types on the SpaServer (http://spa.ridom.de/index.shtml). In cases where a *spa* type could not be determined, the WGS was routinely repeated. MLST was called from the assembled contigs by comparison to the published sequence types on http://saureus.mlst.net.

At Statens Serum Institut, all isolates were confirmed to be MRSA using a multiplex PCR detecting the presence of *mecA*, *mecC*, PVL, and *spa* (7). The *spa* type was determined by direct Sanger sequencing of the amplicons. The *spa* types were assigned using the software program BioNumerics v6.1 (Applied Maths, Sint-Martens-Latem, Belgium). In cases where the *spa* amplicons were undetected, alternative primers were used: spa_239f, 5′-ACTAGGTGTAGGTATTGCATCTGT-3′; spa_1717r, 5′-TCCAGCTAATAACGCTGCACCTAA-3′; spa_1084f, 5′-ACAACGTAACGGCTTCATCC-3′; spa_1618r, 5′-TTAGCATCTGCATGGTTTGC-3′.

The MRSA databases from Hvidovre Hospital and Statens Serum Institut were compared, resulting in a common database of 699 MRSA isolates from new patients in 2013.

We also evaluated whether the N_{50}, an indicator of assembly quality, could be used as a quality control for the likelihood of a correct *spa* type. To find N_{50} of an assembly that contains N nucleotides, the contigs are sorted by size and the nucleotides in the sorted contigs are counted from one end. N_{50} is the size of the contig containing nucleotide number $N/2$.

RESULTS

Of the 699 MRSA isolates, 680 (97%) had identical *spa* types by Sanger sequencing and WGS. These 680 isolates included 136 different *spa* types, the most common being t002 ($n = 74$), t008 ($n = 63$), t019 ($n = 62$), and t304 ($n = 62$). The number of *spa* repeats ranged from 3 to 17. Nineteen isolates had different *spa* types by Sanger sequencing and WGS.

The *spa* types of the 19 isolates with *spa* types that were different by the two methods are presented in Table 1. In 18 cases the *spa* type difference was caused either by the lack of 24-bp repeats (13 isolates) or by additional 24-bp repeats (5 isolates) by WGS. In one case, the *spa* types found by the two methods were unrelated based on *spa* repeats (t630/t304). In this case, the assembly was of low quality ($N_{50} = 2{,}248$ bp) and only partial *mecA* and *nuc* genes were identified. Furthermore, the genome size was 3.4 Mbp, indicating an isolate contamination. The patient was related to a person with t304. We therefore resequenced the isolate and found both t304 and t630 in the same sample and a genome size of 3.35 Mbp.

In Table 2, the isolates are divided into two groups, either with an N_{50} below or above 20,000. The number of isolates that obtained a *spa* type and the number of isolates with a correct *spa* type, using Sanger *spa* types as gold standards, are presented in the table, as is the percentage of a correct *spa* type in each group.

Forty isolates (5.7%) did not obtain a *spa* type by the first WGS due to either a poor sequencing result ($n = 27$) or the *spa* repeats being assembled on more than one contig ($n = 13$). These 40 isolates were resequenced, resulting in a *spa* type in all isolates (Table 2). A number of isolates lacked the *spa* amplicons using the multiplex PCR prior to Sanger sequencing, and this was resolved by the use of alternative *spa* primers. At Statens Serum Institut, repetition of Sanger sequencing due to an initial poor sequencing result is done for less than 5% of isolates.

DISCUSSION

With the rapidly increasing number of clinical bacterial isolates being whole genome sequenced, backward compatibility to sequence-based typing methods, such as *spa* typing and MLST, are very important. In this study, we focused on the commonly used *S. aureus* sequence-based typing method, *spa* typing, represented by more than 298.000 isolates in the SpaServer database. The *spa* gene

TABLE 2 N_{50} values and *spa* typing results[a]

WGS run	N_{50} category	No. of isolates	No. of isolates with a *spa* type	No. of isolates with a correct *spa* type	% correct *spa* types (95% confidence interval)
First	Below 20,000	70	47	44	94 (82.4–98.7)
	Above 20,000	629	612	600	98 (96.9–99.1)
Second	Below 20,000	7	6	5	83
	Above 20,000	33	32	29	91
Third	Above 20,000	2	2	2	100

[a] Resequencing was performed only for isolates that lacked a *spa* type.

is composed of a variable number of highly similar repeats and therefore should be more challenging for WGS, since repeated sequences can be misassembled (8). We used Illumina 150-bp paired-end sequencing, and thus *spa* types longer than six repeats are always on more than one read. When these reads are assembled, it can lead to a misinterpretation if identical repeats are in a series. The use of 250-bp paired-end sequencing might improve the *spa* typing reliability. However, with an agreement in this study of 97%, it would probably not give an add-on value compensating for increased expenses and longer run time.

In the 19 cases where the *spa* types differed by the two methods, 18 were caused by *spa* types differing by one or more repeats. In most cases, series of repeats were missed by the WGS assembly (Table 1). However, most of these *spa* types were generally so closely related that the BURP (based upon repeat pattern) algorithm (9) would have grouped them into the same *spa* complex. In one case, the *spa* types obtained by the two methods seemed unrelated (t630/t304). However, a further look at the WGS data revealed a low N_{50}, a larger genome size than expected, and only partial sequences of the *mecA* and *nuc* genes. Repeating the WGS resulted in the finding of both t304 and t630 in the same sample and again a genome size of 3.35 Mbp. This example illustrates the importance of having a thorough look at the WGS data before accepting a *spa* type. The reporting of incorrect *spa* types should not have any consequence in the case of a suspected MRSA outbreak, if MRSA isolates from all outbreak-related patients, irrespective of *spa* types, are included in the SNP analysis. The SNP analysis will reveal the close relatedness of the outbreak isolates and would therefore still confirm a connection between patients.

In our study, for 13 isolates, the *spa* repeats were located on 2 or 3 contigs after the first WGS, resulting in a nontypeable *spa* type. These isolates were all resequenced, and all isolates obtained a *spa* type by WGS after a second run (Table 2). The 13 isolates had *spa* types with between 4 and 16 repeats, and the assembly difficulties could therefore not be explained by long *spa* repeat regions. Furthermore, all isolates had different *spa* types, so the assembly difficulties could not be correlated to specific *spa* types.

No defined quality control for microbial WGS results exists. We look at the N_{50}, the genome size, and the complete detection of the *mecA* and *nuc* genes when evaluating the WGS data. We evaluated whether a low N_{50} could be used as an indicator of when to repeat the WGS. As expected, a smaller number of isolates obtain a *spa* type when the N_{50} is below 20,000 (Table 2). Although there is a higher percentage of correct *spa* types when the N_{50} is above 20,000 than with an N_{50} below 20,000 (94% versus 98%), the 95% confidence intervals overlap. Therefore, the N_{50} is not a parameter for when to trust a *spa* type in our study. Our routine is therefore only to repeat the WGS when no *spa* type is obtained or when the genome size and/or the *mecA* and *nuc* genes are not of the expected length.

We believe that *spa* types are still important to obtain, since this information can easily be exchanged with and communicated to clinicians and distributed in the public domain. Furthermore, the *spa* typing scheme is well known and will be a natural part of an international WGS database. The advantage of WGS is that other genes of interest can easily be analyzed, including additional genes used for typing, for example, MLST, virulence genes, and resistance genes (10). In addition, the WGS data are stored and whenever needed can be reanalyzed for genes of interest or compared to other whole-genome-sequenced MRSA isolates. This is especially interesting for community isolates, where collection of epidemiological data can be difficult to obtain. The cost of WGS of one isolate is at the moment approximately €100, but obtaining all the data included in our analysis pipeline by Sanger sequencing would be more expensive and time-consuming.

In conclusion, our data show that *spa* typing by WGS can reliably replace Sanger sequencing. Virtually total agreement was found between the two methods. The few divergences had no true significance because in outbreak investigations, these isolates would cluster together with related *spa* types by SNP analysis. We believe that the advantage of the additional genomic information gained by WGS is higher than the disadvantage of a small number of inexact *spa* types.

ACKNOWLEDGMENTS

We are grateful to Susanne M. Rohde, Louise B. Christensen, and Dilek Ozdemur (all at Hvidovre University Hospital) for WGS of all isolates and to Alexandra Medina, Stine Frese-Madsen, and Lone Ryste Kildevang Hansen (Statens Serum Institut) for performance of Sanger sequencing.

H.K.J. was supported by a clinical research stipend from the Novo Nordisk Foundation. We thank Toyotafonden, Denmark, for a grant to purchase the Illumina MiSeq instrument.

We have no conflicts of interest to declare.

REFERENCES

1. **Grundmann H, Aires-de-Sousa M, Boyce J, Tiemersma E.** 2006. Emergence and resurgence of meticillin-resistant *Staphylococcus aureus* as a public-health threat. Lancet **368:**874–885. http://dx.doi.org/10.1016/S0140-6736(06)68853-3.
2. **Otto M.** 2013. Community-associated MRSA: what makes them special? Int. J. Med. Microbiol. **303:**324–330. http://dx.doi.org/10.1016/j.ijmm.2013.02.007.
3. **Harmsen D, Claus H, Witte W, Rothganger J, Claus H, Turnwald D, Vogel U.** 2003. Typing of methicillin-resistant *Staphylococcus aureus* in a university hospital setting by using novel software for *spa* repeat determination and database management. J. Clin. Microbiol. **41:**5442–5448. http://dx.doi.org/10.1128/JCM.41.12.5442-5448.2003.

4. **Aires-de-Sousa M, Boye K, de Lencastre H, Deplano A, Enright MC, Etienne J, Friedrich A, Harmsen D, Holmes A, Huijsdens XW, Kearns AM, Mellmann A, Meugnier H, Rasheed JK, Spalburg E, Strommenger B, Struelens MJ, Tenover FC, Thomas J, Vogel U, Westh H, Xu J, Witte W.** 2006. High interlaboratory reproducibility of DNA sequence-based typing of bacteria in a multicenter study. J. Clin. Microbiol. **44:**619–621. http://dx.doi.org/10.1128/JCM.44.2.619-621.2006.
5. **Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW.** 2012. Transforming clinical microbiology with bacterial genome sequencing. Nat. Rev. Genet. **13:**601–612. http://dx.doi.org/10.1038/nrg3226.
6. **Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, Ip CL, Wilson DJ, Didelot X, O'Connor L, Lay R, Buck D, Kearns AM, Shaw A, Paul J, Wilcox MH, Donnelly PJ, Peto TE, Walker AS, Crook DW.** 2012. A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. BMJ Open **2:**e001124. http://dx.doi.org/10.1136/bmjopen-2012-001124.
7. **Stegger M, Andersen PS, Kearns A, Pichon B, Holmes MA, Edwards G, Laurent F, Teale C, Skov R, Larsen AR.** 2012. Rapid detection, differentiation and typing of methicillin-resistant *Staphylococcus aureus* harbouring either *mecA* or the new *mecA* homologue *mecA*(LGA251). Clin. Microbiol. Infect. **18:**395–400. http://dx.doi.org/10.1111/j.1469-0691.2011.03715.x.
8. **Salzberg SL, Yorke JA.** 2005. Beware of mis-assembled genomes. Bioinformatics **21:**4320–4321. http://dx.doi.org/10.1093/bioinformatics/bti769.
9. **Mellmann A, Weniger T, Berssenbrugge C, Rothganger J, Sammeth M, Stoye J, Harmsen D.** 2007. Based Upon Repeat Pattern (BURP): an algorithm to characterize the long-term evolution of *Staphylococcus aureus* populations based on *spa* polymorphisms. BMC Microbiol. **7:**98. http://dx.doi.org/10.1186/1471-2180-7-98.
10. **Leopold SR, Goering RV, Witten A, Harmsen D, Mellmann A.** 2014. Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. J. Clin. Microbiol. **52:**2365–2370. http://dx.doi.org/10.1128/JCM.00262-14.