

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



LSHTM Research Online

Raven, Kathy E; Blane, Beth; Leek, Danielle; Churcher, Carol; Kokko-Gonzales, Paula; Pugazhendhi, Dhamayanthi; Fraser, Louise; Betley, Jason; Parkhill, Julian; Peacock, Sharon; (2019) Methodology for whole genome sequencing of MRSA in a routine hospital microbiology laboratory. *Journal of Clinical Microbiology*. ISSN 0095-1137 DOI: <https://doi.org/10.1128/JCM.00180-19>

Downloaded from: <http://researchonline.lshtm.ac.uk/4652213/>

DOI: <https://doi.org/10.1128/JCM.00180-19>

Usage Guidelines:

Please refer to usage guidelines at <http://researchonline.lshtm.ac.uk/policies.html> or alternatively contact researchonline@lshtm.ac.uk.

Available under license: <http://creativecommons.org/licenses/by/2.5/>

<https://researchonline.lshtm.ac.uk>

1 **Methodology for whole genome sequencing of MRSA in a routine hospital microbiology**
2 **laboratory**

3 **Running title: Laboratory methods for routine MRSA sequencing**

4

5 Kathy E. Raven*^{a#}, Beth Blane*^a, Danielle Leek^a, Carol Churcher^a, Paula Kokko-Gonzales^b,
6 Dhamayanthi Pugazhendhi^b, Louise Fraser^b, Jason Betley^b, Julian Parkhill^c, Sharon J.
7 Peacock^{a,c,d}

8

9 ^aDepartment of Medicine, University of Cambridge, Box 157 Addenbrooke's Hospital, Hills
10 Road, Cambridge, CB2 0QQ, UK

11 ^bIllumina, Inc., Great Abington, Cambridge, UK

12 ^cWellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA,
13 UK

14 ^dLondon School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

15

16 *Joint first authors

17 Corresponding author: Kathy Raven (ker37@medschl.cam.ac.uk)

18 **Abstract**

19 There is growing evidence for the value of bacterial whole genome sequencing in hospital
20 outbreak investigation. Our aim was to develop methods that support efficient and accurate
21 low throughput clinical sequencing of methicillin-resistant *Staphylococcus aureus* (MRSA).
22 Using a test panel of 25 MRSA isolates associated previously with outbreak investigations,
23 we devised modifications to library preparation that reduced processing time by 1 hour. We
24 determined the maximum number of isolates that could be sequenced per run using an Illumina
25 MiniSeq and a 13 hour (overnight) run time, which equated to 21 MRSA isolates and 3
26 controls (no template, positive and negative). Repeatability and reproducibility assays based
27 on this sequencing methodology demonstrated 100% accuracy in assigning species and
28 sequence type (ST) and detecting *mecA*. Established genetic relatedness between isolates was
29 recapitulated. Quality control (QC) metrics were evaluated over nine sequencing runs.
30 168/173 (97%) test panel MRSA genomes passed QC metrics based on the correct species
31 assigned, detection of *mecA* and ST, and depth/coverage metrics. An evaluation of
32 contamination in these 9 runs showed that positive and negative controls and test MRSA
33 sequence files contained <0.14% and <0.48% of fragments matching another species,
34 respectively. Deliberate contamination experiments confirmed that this was insufficient to
35 impact on data interpretation. These methods support reliable and reproducible clinical MRSA
36 sequencing with a turnaround time (from DNA extraction to availability of data files) of 24
37 hours.

38 **Introduction**

39 There is growing evidence of the value of bacterial whole genome sequencing (WGS) in
40 hospital infection control practice and outbreak investigation (1). Numerous retrospective
41 studies have shown that bacterial sequencing provides the discrimination required to
42 distinguish between isolates of the same lineage, overcoming this limitation of previous typing
43 methods (2-7). There is also strong published support for its use to investigate carriage,
44 transmission and suspected outbreaks in high-risk areas such as intensive care units (2, 6).
45 Used early, this could lead to action that limits the size of an outbreak (6, 8). Furthermore,
46 sequencing can exclude outbreaks where a cluster of patients positive for the same pathogenic
47 species has arisen by chance (9), saving unnecessary infection control interventions and
48 outbreak investigations.

49

50 The benefit gained from using WGS during outbreak detection is likely to be greatest if the
51 technology is embedded within healthcare institutions and performed with a rapid turnaround
52 time. This has become increasingly feasible through technical advances in sequencing
53 instruments and the availability of commercial kits and liquid handling robots that simplify
54 DNA extraction and library preparation. The laboratory processing aspects of WGS are now
55 within the capabilities of larger diagnostic laboratories. The technical feasibility of sequencing
56 in real-time has been demonstrated previously at a tertiary care hospital in Germany, but the
57 turnaround time was 4.4-5.3 days with a cost of ~£170 (10). Reducing this turn-around time
58 to results and the cost of sequencing will be key to implementing sequencing in the clinical
59 setting and having an impact on infection control. In our clinical microbiology laboratory at
60 Addenbrooke's Hospital in Cambridge, United Kingdom, we are developing the methods and
61 processes to introduce routine WGS of targeted nosocomial pathogens in close to real-time to
62 enhance our infection control practice. Here, we describe the development of laboratory

63 processing methodology for low throughput clinical sequencing of methicillin-resistant
64 *Staphylococcus aureus* (MRSA).

65

66 **Materials and Methods**

67 ***Test panel isolates***

68 Twenty-nine bacterial isolates (27 *S. aureus* and 2 *E. coli*) were assembled into a test panel
69 for the study (Table 1). The majority of *S. aureus* (n=25) were MRSA from two evaluations
70 of sequencing at the Cambridge University Hospital NHS Foundation Trust hospital (CUH)
71 (6,7). Twenty-one MRSA were selected from a 12-month study of MRSA-positive patients
72 (7) to provide representation of the dominant clonal complexes in our setting (CC22, CC30
73 and CC5), combined with a range of genetic relatedness. A further 4 MRSA (all sequence type
74 (ST) 22) were from an outbreak in a special care baby unit (6). Also included were 4 reference
75 isolates: MRSA HO 5096 0412, methicillin-susceptible *S. aureus* NCTC 6571, *E. coli* NCTC
76 12241 and *E. coli* NCTC 10418. For sequencing, isolates were cultured from frozen stocks
77 onto Columbia Blood Agar (CBA, Oxoid), incubated in air at 37°C overnight, and single
78 colonies picked for DNA extraction and further processing. Table 1 indicates the isolates used
79 in each sequencing run.

80

81 ***Positive and negative controls***

82 Three controls were included in every sequencing run to monitor the ongoing performance of
83 the entire testing process. These were a no template control, a positive control (MRSA
84 MPROS0386) that is 115 core genome SNPs different from the MRSA HO 5096 0412
85 mapping reference, and a negative control (*E. coli* NCTC12241). The no template control
86 contained all assay components except for DNA and was used to verify the lack of
87 contamination across reagents and samples. The positive control was used to control the entire

88 assay process and analytical accuracy. The negative control was used to assess cross-
89 contamination during processing and represented the non-target DNA sample to verify
90 analytical specificity. In the first two runs an alternative *E. coli* control (NCTC10418) was
91 used, but this had a low match to *E. coli* in Kraken (~22%) and was replaced by NCTC12241
92 (>50% match). Fresh stocks of molecular grade water and phosphate-buffered-saline were
93 opened each week. Other 'reuse' reagents were checked for bacterial contamination weekly
94 by sub-culturing using a 1µl loop onto CBA and incubating overnight in air at 37°C.

95

96 ***DNA extraction, library preparation and sequencing***

97 DNA was extracted using the QIAgen DNA mini extraction kit
98 ([https://www.qiagen.com/gb/shop/sample-technologies/dna/genomic-dna/qiaamp-dna-mini-](https://www.qiagen.com/gb/shop/sample-technologies/dna/genomic-dna/qiaamp-dna-mini-kit/#resources)
99 [kit/#resources](https://www.qiagen.com/gb/shop/sample-technologies/dna/genomic-dna/qiaamp-dna-mini-kit/#resources)) following 'Appendix D: Protocols for Bacteria' 'Isolation of genomic DNA
100 from Gram-positive bacteria' with the following amendments: the incubation with proteinase
101 K was performed at 56°C for 30 minutes; and in the final elutions, 50ul distilled water was
102 added with the full 5 minutes incubation. DNA was quantified using a Qubit fluorometer.
103 Sequencing libraries were made using the Illumina Nextera DNA flex kit based on the
104 manufacturer's instructions (11), with several modifications to reduce processing time (see
105 results). In the first 3 runs, the input DNA to library preparation was normalized to ~100ng,
106 but thereafter we used a range of up to 500ng DNA. Libraries were sequenced on an Illumina
107 MiniSeq with a run time of 13 hours (overnight) using the high output 150 cycle MiniSeq
108 cartridge and the Generate Fastq workflow. Genomes were demultiplexed using the Generate
109 Fastq workflow and the data transferred to an external 1TB USB-connected hard drive for
110 further analysis. Ten sequencing runs were performed during this evaluation; the objective of
111 each run is summarized in Table S1.

112

113 ***Sequence data analysis***

114 Multilocus sequence types (ST) of the MRSA isolates were identified using ARIBA version
115 2.12.1 as described at [https://github.com/sanger-pathogens/ariba/wiki/MLST-calling-with-](https://github.com/sanger-pathogens/ariba/wiki/MLST-calling-with-ARIBA)
116 [ARIBA](https://github.com/sanger-pathogens/ariba/wiki/MLST-calling-with-ARIBA). Species were determined using Kraken version 1
117 (<https://ccb.jhu.edu/software/kraken/>) with the miniKraken database available at
118 https://ccb.jhu.edu/software/kraken/dl/minikraken_20171019_8GB.tgz. The presence of
119 *mecA* (accession number HE681097, position 2790560-2792566) was determined using
120 ARIBA, with a minimum percentage identity of 70% required based on Ito *et al.* (12) , and a
121 minimum of 90% of the gene length covered. All isolates were mapped to the MRSA HO 5096
122 0412 CC22 reference (accession number HE681097) using SMALT
123 (<https://www.sanger.ac.uk/science/tools/smalt-0>) with mapping and base calling performed as
124 described previously (13) with the following modifications: kmer size 13, step size 6. The
125 depth and percentage coverage of the mapping reference were determined using the script
126 available at [https://github.com/sanger-pathogens/vr-](https://github.com/sanger-pathogens/vr-codebase/blob/master/modules/VertRes/Pipelines/Mapping.pm)
127 [codebase/blob/master/modules/VertRes/Pipelines/Mapping.pm](https://github.com/sanger-pathogens/vr-codebase/blob/master/modules/VertRes/Pipelines/Mapping.pm).

128

129 ***Sequence metrics for controls***

130 Controls were required to pass the following quality metrics. MRSA positive control: highest
131 match to *S. aureus* using Kraken, assigned to ST22, *mecA* detected, minimum mean sequence
132 depth of 20x and minimum 80% coverage of the mapping MRSA reference genome (HO 5096
133 0412). *E. coli* negative control: highest species match to *E. coli* in Kraken, *mecA* not detected,
134 no *S. aureus* ST assigned. No template control: contamination from any bacterial DNA of less
135 than 95,000 fragments in Kraken. MRSA isolates from the test panel were required to pass the
136 following metrics: highest match to *S. aureus* using Kraken, assigned to the correct ST, *mecA*

137 detected, minimum sequence depth of 20x and minimum 80% coverage of the mapping MRSA
138 reference genome (HO 5096 0412).

139

140 ***Optimizing the number of isolates per sequencing run***

141 We estimated that the maximum number of MRSA isolates in a single sequencing run was 24
142 based on an expected total data output of 3.3-3.8Gb, an average MRSA genome size of 2.8
143 MB

144 ([https://www.ncbi.nlm.nih.gov/genome/?term=Staphylococcus%20aureus\[Organism\]&cmd=](https://www.ncbi.nlm.nih.gov/genome/?term=Staphylococcus%20aureus[Organism]&cmd=DetailsSearch)
145 [DetailsSearch](https://www.ncbi.nlm.nih.gov/genome/?term=Staphylococcus%20aureus[Organism]&cmd=DetailsSearch)) and a target of ~50x coverage (24 isolates would provide ~49x coverage). We
146 estimated that 21 test MRSA isolates and three controls (*E. coli*, MRSA and no template)
147 could be included per sequence run. This was evaluated by performing sequencing runs that
148 contained either 14, 18 or 21 test MRSA isolates from the study panel plus the 3 controls. One
149 MRSA isolate from the 21-test isolate run failed to produce sufficient DNA during extraction
150 and the *E. coli* control was included twice.

151

152 ***Repeatability and reproducibility***

153 Repeatability was evaluated by sequencing six MRSA isolates (HO 5096 0412, MPROS0386,
154 SASCBU17, SASCBU18, SASCBU25 and SASCBU35) in triplicate in a single sequencing
155 run. For each isolate, frozen stock was sub-cultured onto CBA, incubated in air at 37°C
156 overnight, and three separate colonies taken forward for individual DNA extraction, library
157 preparation and sequencing. Reproducibility was evaluated by sequencing 21 MRSA isolates
158 from the test panel in three independent runs. Each isolate was sub-cultured onto CBA and
159 incubated in air at 37°C overnight, after which three individual colonies were taken forward
160 for DNA extraction, library preparation and sequencing, one for each sequence run. The entire
161 process was performed by different laboratory staff on three different days. The resulting fastq

162 files were analysed as above. Isolates that failed QC metrics were excluded from further
163 analysis (3/18 and 1/63 test isolates failed the repeatability and reproducibility assays,
164 respectively, based on low depth/coverage.

165

166 Definition of a correct result was based on species identification, ST assignment, detection of
167 *mecA*, and identification of genetic relatedness based on the detection of single nucleotide
168 polymorphisms (SNPs) in the core genome compared to the original sequence and the within-
169 run or between run-replicates. Genetic relatedness was determined based on mapping to a
170 clonal complex (CC)-specific references, excluding positions denoted as ‘N’ because of failure
171 to call a base. Each repeat and the original sequence data were mapped to a CC-specific
172 mapping reference using SMALT (MRSA HO 5096 0412 (CC22) for ST22 and ST2371;
173 MRSA252 (CC30, BX571856) for ST30; and N315 (CC5, BA000033) for ST5). Mobile
174 genetic elements were removed using the files available at
175 https://figshare.com/authors/Francesc_Coll/5727779 and the script available at
176 https://github.com/sanger-pathogens/remove_blocks_from_aln. Single nucleotide
177 polymorphisms (SNPs) were identified using the script available at [https://github.com/sanger-](https://github.com/sanger-pathogens/snp-sites)
178 [pathogens/snp-sites](https://github.com/sanger-pathogens/snp-sites). SNPs were identified based on the following parameters: minimum
179 number of reads matching the SNP = 4; minimum number of reads matching the SNP per
180 strand = 2; ratio of SNP base to alternative base >0.75; variant quality >50; mapping quality
181 >30.

182

183 Diagnostic sensitivity and specificity were calculated, using the following definitions: true
184 positives, the number of genetically related isolates based on the original data that cluster
185 together based on the test data; false negatives, the number of genetically related isolates based
186 on the original data that do not cluster together in the test data; true negatives, the number of

187 genetically unrelated isolates based on the original data that do not cluster together in the test
188 data; and false positives, the number of genetically distant isolates based on the original data
189 that cluster together based on the test data (14). Clustering was defined based on three SNP
190 classifications: (i) Recent transmission highly likely, 0-10 SNPs different (based on a median
191 within host diversity of 6 SNPs over a year (7) and an estimated mutation rate of 4 SNPs/core
192 genome/year (15), (ii) Recent transmission likely, 11-25 SNPs, and (iii) Recent transmission
193 possible, 26-50 SNPs different (based on the definition of a cluster described by Coll et al.
194 (7)). Isolates >50 SNPs different were classified as genetically unrelated.

195

196 ***Analysis of contamination***

197 The impact on quality metrics from varying levels of DNA contamination during clinical
198 MRSA sequencing was evaluated using intentional spiking experiments. One MRSA isolate
199 from the test panel (MPROS1839 (ST22)) and *E. coli* NCTC 12241 were cultured and DNA
200 extracted and quantified as described above. Donor DNA was inoculated into the recipient
201 sample to achieve a final spiked concentration of 0%, 0.1%, 1%, 10% or 20% (see results for
202 details of donor and recipient). Contamination with the spike was defined based on the number
203 and proportion of fragments matching to *S. aureus* or *E. coli* based on Kraken. The effect of
204 contamination was evaluated using this metric together with the proportion of the *S. aureus*
205 CC22 reference covered during mapping, depth of coverage of the mapping reference, and
206 *mecA* and ST detected by Ariba. Unintentional contamination from internal controls or
207 external sources was evaluated based on the number and proportion of reads matching to other
208 species in Kraken.

209

210 ***Data availability***

211 Sequence data generated during this study are available from the European Nucleotide Archive
212 (<https://www.ebi.ac.uk/ena>) under the accession numbers listed in Table 1.

213

214 **Results**

215 Our aim was to develop methods that would support efficient and accurate low throughput
216 MRSA sequencing in a routine microbiology laboratory in less than 24 hours (from DNA
217 extraction to availability of sequence data). Key goals were to maximize the number of isolates
218 sequenced per run, reduce processing time of DNA preparation, and evaluate quality controls,
219 precision (reproducibility and repeatability), and contamination.

220

221 Maximizing the number of isolates per sequencing run was evaluated by performing
222 sequencing runs that contained either 14, 18 or 21 test MRSA isolates from the study panel
223 plus the 3 controls, which were sequenced using the Illumina MiniSeq with a run time of 13
224 hours. Median (range) sequence depth for the test MRSA isolates was 92x (33-247x), 63x (45-
225 77x) and 65x (18-107x), respectively, with a minimum of 87% of the genome covered (Table
226 S2). One isolate in the 21 test MRSA run failed the QC metrics based on depth of coverage
227 (17.9x), which on further evaluation could be explained by low input DNA (Table S2). Based
228 on this, we used 21 test isolates plus 3 controls per run during the remainder of the study.

229

230 We sought modifications to the manufacturer's protocol for library preparation (Illumina
231 Nextera DNA flex kit) that would reduce processing time while maintaining performance. We
232 proposed that two steps could be changed: (i) the tagmentation (TAG program) and
233 tagmentation stop (TSB incubation) steps each require 15 minutes incubation, which were
234 reduced to 5 minutes each; (ii) Pooling of libraries is recommended after bead clean-up and
235 size selection, but we pooled libraries after PCR and before the bead-cleanup and size

236 selection. Two sequencing runs of 21 test panel MRSA + 3 controls were compared, one of
237 which used the original protocol and the other made both changes to the protocol. Data were
238 compared for quantity of DNA added to the library preparation versus the size of the resulting
239 fastq files and depth of coverage, as surrogates for the individual DNA quantity outputs from
240 library preparation, which are unavailable with the modified protocol. Detailed results are
241 provided in Table S3. In summary, comparison of original versus modified protocol showed
242 negligible difference. The median (range) fastq size for the original versus modified protocols
243 were 171MB (77-208MB) following 174-480ng DNA input versus 112MB (90-133MB)
244 following 90-384ng DNA input. The median (range) depth of coverage for the original versus
245 modified protocols were 87x (37-99x) versus 56x (43-70x). Together, these resulted in a
246 reduction in processing time from 3.5 to 2.5 hours for library preparation, taking the combined
247 time for DNA preparation and library preparation to 4.5 hours. Subsequent runs used these
248 modifications.

249

250 Repeatability was based on concordance of assay results and quality metrics for six MRSA
251 isolates sequenced in triplicate in a single sequencing run. This demonstrated 100%
252 concordance in assigning species, ST and detecting *mecA*. Four of the six isolates were drawn
253 from a study that investigated a single outbreak on an intensive care unit (6) and were
254 previously identified as being 0 SNPs different (SASCBU17 and SACBU18), 5 SNPs different
255 (SASCBU25), or unrelated to the outbreak (SASCBU35, >1,500 SNPs different from other
256 isolates). The remaining two isolates were MRSA HO 5096 0412 and the positive MRSA
257 control (MPROS0386). Zero SNPs were identified between the within-run replicates for all
258 isolates, equating to a repeatability per replicate of 100%. Using the original published
259 sequence mapped to the CC22 reference (HO 5096 0412) as the gold standard, all 6 isolates
260 in triplicate had identical base calls to the original sequence (excluding positions denoted as

261 'N' because of failure to call a base), equating to a repeatability per replicate and per base pair
262 of 100%.

263

264 Reproducibility was evaluated by sequencing 21 test panel MRSA isolates in three
265 independent runs. This demonstrated 100% accuracy in assigning species, ST and detecting
266 *mecA*. Eighteen of the 21 isolates represented six distinct outbreaks encompassing four
267 different STs (ST22, ST30, ST5 and ST2371) identified during 12 months of genomic
268 surveillance (n=15) (7) or a single outbreak in an intensive care unit (n=3) (6) . Of the
269 remainder, 2 isolates were not involved in these outbreaks based on low relatedness, and 1
270 isolate was the mapping reference MRSA HO 5096 0412. There were 0 SNPs identified
271 between between-run replicates, providing a reproducibility per replicate of 100%. Using the
272 original published sequence when mapped to the CC22 reference as the gold standard, 18
273 isolates were identical to the original sequence across replicates. The remaining three isolates
274 showed a difference in SNPs compared with the original sequence: MPROS0292 (ST22) had
275 1-2 SNPs different, one of which was reproduced in all three repeats and the other was
276 reproduced in two repeats with an N base call in the remaining repeat. MPROS1125 (ST22)
277 had 1 SNP different in one repeat with an N base call in the same position in the remaining
278 two repeats. MPROS2335 was identical for two replicates but the third replicate had 10 SNPs.
279 In comparison to the original sequence this provides an assay accuracy of 92.3% (60/65
280 repeats), although the true accuracy is likely be higher as the majority of SNPs may be genuine
281 based on their presence among repeats.

282

283 We next sought to determine the diagnostic sensitivity and specificity for outbreak detection
284 in each of the three reproducibility runs, using the genetic relatedness established previously
285 (6,7) as the gold standard (Table S4). All test isolate pairs within each run were in the same

286 genetic relatedness category (0-10 SNPs, 11-25 SNPs, 26-50 SNPs, >50 SNPs) as isolate pairs
287 in the original data. This was reproducible across all three runs and represents a diagnostic
288 sensitivity and specificity for outbreak detection of 100%, which was retained across a range
289 of definitions for genetic relatedness. The majority of isolate pairs were within 1 SNP of the
290 expected SNP difference based on the gold standard. The exceptions were Cluster 3 (2 SNPs
291 different between MPROS0046 and MPROS1125 in two runs relating to failure to call a base
292 at one position, and a SNP in a region that was absent across the replicates but present at low
293 coverage in the original sequence); and Cluster 4 (MPROS0688 and MPROS2335; 2 SNPs
294 different in two runs due to two positions at which a base failed to be called; and 6 SNPs
295 different in the final run. The isolate sequence MPROS2335 was genetically identical to a
296 second isolate from the same patient sequenced by Coll *et al.* (7) but not included here. From
297 this, we suspect within-host diversity of MRSA in this case, and sequencing of different
298 colonies of the same lineage.

299

300 Quality control metrics were evaluated for the assay controls and MRSA isolates from the test
301 panel over nine sequencing runs (Table S5 provides further details). All three controls in each
302 sequence run passed the required QC metrics. Of 173 *S. aureus* test panel MRSA isolates
303 sequenced, 168 (97%) passed the QC metrics. The five failures were based on insufficient
304 depth/coverage associated with low input DNA (n=2) or potential loss of DNA during library
305 preparation (n=3). Excluding these 5 failed isolates and the control isolates, *S. aureus* was the
306 top match in Kraken in all cases (median (range) 85.8% (77.2-89.3%), the median (range)
307 depth was 59x (21–247x), and the median (range) proportion of the reference genome covered
308 was 94.6% (86.3-100%).

309

310 We then undertook deliberate contamination experiments to allow us to estimate the impact
311 of varying levels of DNA contamination from internal controls or external sources on quality
312 metrics. Details of the donor and recipient DNA, the concentrations of spiked DNA and our
313 findings are summarized in Table 2. Contamination of the no template control with increasing
314 concentrations of MRSA DNA did not lead to the control erroneously passing the QC metrics
315 for MRSA until the final spiked concentration reached 10% or greater. This indicates that
316 contamination of the no template control at 1% (which equated to 96,671 fragments matching
317 *S. aureus* in Kraken) can be tolerated. Contaminating the positive MRSA control with
318 increasing concentrations of *E. coli* DNA demonstrated that this could tolerate up to 10%
319 contamination (which equated to 4.04% fragments matching *E. coli* in Kraken) before the
320 MRSA QC metrics were not achieved.

321

322 We also evaluated unintentional contamination in nine runs (excluding the deliberate
323 contamination assay). All *E. coli* and MRSA control sequence data files contained less than
324 0.14% of fragments matching another species (Table S6). For the test MRSA sequence files,
325 matches to other staphylococcal species were identified in over half of samples (109/173,
326 median 0.05%, range 0.01-0.48% of fragments). Very low-level matches (0.01-0.13%) to
327 other species were also identified in specific files (Table S6). All isolates had less than 0.2%
328 of fragments matching to another species, with the exception of a single reference isolate of
329 MSSA that had a match of up to 0.48% to *Staphylococcus nepalensis*. Based on the number
330 of fragments in Kraken for the no template controls, and the proportion of fragments in Kraken
331 for the remaining sequences, this demonstrates that, with the exception of the isolate above,
332 all controls and test isolates had levels of contamination below 1% (0.4% of fragments) across
333 the nine sequencing runs (Table S6).

334

335 **Discussion**

336 Our aim was to develop and describe methods for low throughput clinical sequencing of
337 MRSA using commercial kits and manual methods. Our rationale was that this could support
338 wider uptake in smaller diagnostic laboratories that are not equipped to undertake high volume
339 sequencing using automated robots. Whilst liquid handling robots are essential for high
340 volume sequencing such as that increasingly performed by public health reference
341 laboratories, the majority of routine clinical laboratories have yet to invest in sequencing
342 pipelines with their associated capital and maintenance costs.

343

344 An important objective was to enable a 24-hour turnaround time from DNA extraction to
345 availability of sequence data. The combined time for DNA preparation and library preparation
346 is 4.5 hours, followed by a 13-hour (overnight) sequencing run on the Illumina MiniSeq. This
347 would support a pipeline of clinical sequencing in which relevant cultures were identified in
348 a routine laboratory and processed including sequencing within a day. The methods described
349 here are based on a single colony, which when implemented in routine practice could be
350 obtained from the original diagnostic clinical plate. This turnaround time, in combination with
351 a rapid automated analysis pipeline, would allow infection control to determine whether
352 patients were involved in an outbreak or not the day after a positive culture. This could allow
353 rapid instigation of enhanced infection control procedures when an outbreak is detected to
354 prevent further spread of the outbreak, and prevention of infection control actions such as ward
355 closures if a suspected outbreak could be refuted.

356

357 We also maximised the number of MRSA per sequencing run to minimize the cost per isolate.
358 Based on 21 clinical isolates per run with three controls, the price per clinical isolate is
359 currently £70 for DNA extraction, library preparation and sequencing. Whilst individual

360 hospitals are unlikely to frequently reach 21 clinical MRSA isolates suspected to be involved
361 in an outbreak, we suggest a paradigm shift whereby all patients identified as MRSA positive
362 have an isolate sequenced, and whole-genome sequencing leads infection control actions. This
363 would reduce the turnaround time for action since current outbreak detection relies on multiple
364 time-consuming steps including manually identifying patients that have been in the same ward
365 at the same time. Using whole-genome sequencing combined with automated analysis would
366 rapidly pinpoint which patients are involved in an outbreak or not, defining the cases that
367 infection control need to act on, and those that require no action. The combination of
368 turnaround and cost are critical measures for clinical translation. Alternative sequencing
369 instruments such as the Oxford Nanopore Technologies provide the option for further
370 reductions in sequencing time (16), and over time the cost and turnaround time of sequencing
371 will undergo further reductions. As costs fall, lower-throughput technologies such as the
372 Illumina iSeq 100 may become viable for routine clinical laboratories with smaller sample
373 numbers.

374

375 We described the use and evaluation of assay controls, examined the impact of contamination
376 on data interpretation and determined the extent to which we inadvertently contaminated the
377 assay. All three controls passed the required QC metrics in every run, together with 97% of
378 test panel MRSA isolates sequenced. High levels of contamination were required before the
379 controls failed QC metrics, and levels of inadvertent contamination were low. Evaluation of
380 precision showed 100% repeatability and reproducibility in assigning species and ST and
381 detecting *mecA*. SNP detection was 100% repeatable, but reproducibility was less than 100%
382 because of the detection of a small number of SNPs that were not present in the original
383 sequence. These can be explained by minor heterogeneity in colonies prepared for independent
384 sequencing, with similar findings reported previously based on sequencing of a range of

385 bacterial species (14). Importantly, diagnostic sensitivity and specificity for outbreak detection
386 were 100%, indicating that the data generated accurately determined MRSA relatedness,
387 which supports use of this assay during outbreak investigation. The parameters evaluated in
388 this study were in line with the workflow for validation of whole-genome sequencing in
389 clinical laboratories described previously, and obtained comparable results (14).

390

391 Our findings indicate that the methods evaluated here can provide high quality data. The single
392 largest impediment to clinical sequencing is lack of fully automated data interpretation
393 software that has a rapid turn-around time and is suitable for use by non-experts. This will
394 need to be addressed for routine clinical sequencing to become viable, and is currently being
395 investigated by numerous groups and investigators.

396

397 **Acknowledgements**

398 This publication presents independent research supported by the Health Innovation Challenge
399 Fund (WT098600, HICF-T5-342), a parallel funding partnership between the Department of
400 Health and Wellcome Trust. The views expressed in this publication are those of the author(s)
401 and not necessarily those of the Department of Health or Wellcome Trust. This project was
402 also funded by a grant awarded to the Wellcome Trust Sanger Institute (098051).

403

404 **Conflict of interest**

405 SJP and JP are consultants to Specific Technologies and Next Gen Diagnostics. PK-G, DP,
406 LF and JB are employees of and hold stock in Illumina, Inc. The other authors have no
407 conflicts of interest.

408

409

410 **References**

- 411 1. Peacock SJ, Parkhill J, Brown NM. 2018. Changing the paradigm for hospital
412 outbreak detection by leading with genomic surveillance of nosocomial pathogens.
413 *Microbiol* 164:1213–1219.
- 414 2. Gorrie CL, Mirc Eta M, Wick RR, Edwards DJ, Thomson NR, Strugnell RA, Pratt
415 NF, Garlick JS, Watson KM, Pilcher D V., McGloughlin SA, Spelman DW, Jenney
416 AWJ, Holt KE. 2017. Gastrointestinal carriage is a major reservoir of *Klebsiella*
417 *pneumoniae* infection in intensive care patients. *Clin Infect Dis* 65:208–215.
- 418 3. Raven KE, Gouliouris T, Brodrick H, Coll F, Brown NM, Reynolds R, Reuter S,
419 Török ME, Parkhill J, Peacock SJ. 2017. Complex routes of nosocomial
420 vancomycin-resistant *Enterococcus faecium* transmission revealed by genome
421 sequencing. *Clin Infect Dis* 64:15–17.
- 422 4. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, Ip CLC,
423 Golubchik T, Batty EM, Finney JM, Wyllie DH, Didelot X, Piazza P, Bowden R,
424 Dingle KE, Harding RM, Crook DW, Wilcox MH, Tim EA, Walker AS. 2014.
425 Diverse sources of *C. difficile* infection identified on whole- genome sequencing. *N*
426 *Engl J Med* 369.
- 427 5. Walker TM, Kohl TA, Omar S V., Hedge J, Del Ojo Elias C, Bradley P, Iqbal Z,
428 Feuerriegel S, Niehaus KE, Wilson DJ, Clifton DA, Kapatai G, Ip CLC, Bowden R,
429 Drobniowski FA, Allix-Béguet C, Gaudin C, Parkhill J, Diel R, Supply P, Crook
430 DW, Smith EG, Walker AS, Ismail N, Niemann S, Peto TEA, Davies J, Crichton C,
431 Acharya M, Madrid-Marquez L, Eyre D, Wyllie D, Golubchik T, Munang M. 2015.
432 Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug
433 susceptibility and resistance: A retrospective cohort study. *Lancet Infect Dis*
434 15:1193–1202.

- 435 6. Harris SR, Cartwright EJP, Török ME, Holden MTG, Brown NM, Ogilvy-Stuart AL,
436 Ellington MJ, Quail MA, Bentley SD, Parkhill J, Peacock SJ. 2013. Whole-genome
437 sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*:
438 A descriptive study. *Lancet Infect Dis* 13:130–136.
- 439 7. Coll F, Harrison EM, Toleman MS, Reuter S, Raven KE, Blane B, Palmer B,
440 Kappeler ARM, Brown NM, Török ME, Parkhill J, Peacock SJ. 2017. Longitudinal
441 genomic surveillance of MRSA in the UK reveals transmission patterns in hospitals
442 and the community. *Sci Transl Med* 9:eaak9745.
- 443 8. Toleman MS, Reuter S, Coll F, Harrison EM, Peacock SJ. 2016. Local persistence of
444 novel MRSA lineage after hospital ward outbreak, Cambridge, UK, 2011–2013.
445 *Emerg Infect Dis* 22:1658–1659.
- 446 9. Török ME, Harris SR, Cartwright EJP, Raven KE, Brown NM, Allison MED,
447 Greaves D, Quail MA, Limmathurotsakul D, Holden MTG, Parkhill J, Peacock SJ.
448 2014. Zero tolerance for healthcare-associated MRSA bacteraemia: Is it realistic? *J*
449 *Antimicrob Chemother* 69:2238–2245.
- 450 10. Mellman A, Bletz S, Boking T, Kipp F, Becker K, Schultes A, Prior K, Harmsen D.
451 2016. Real-time genome sequencing of resistant bacteria provides precision infection
452 control in an institutional setting. *J Clin Microbiol* 54: 2874-2881.
- 453 11. Nextera DNA Flex Library Prep protocol, Illumina, Document number
454 1000000031471 v01, available from
455 [https://support.illumina.com/content/dam/illumina-](https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera_dna_flex/nextera-dna-flex-library-prep-reference-guide-1000000025416-04.pdf)
456 [support/documents/documentation/chemistry_documentation/samplepreps_nextera/n](https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera_dna_flex/nextera-dna-flex-library-prep-reference-guide-1000000025416-04.pdf)
457 [extera_dna_flex/nextera-dna-flex-library-prep-reference-guide-1000000025416-](https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera_dna_flex/nextera-dna-flex-library-prep-reference-guide-1000000025416-04.pdf)
458 [04.pdf](https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera_dna_flex/nextera-dna-flex-library-prep-reference-guide-1000000025416-04.pdf). Accessed 04/02/2019.

- 459 12. Ito T, Hiramatsu K, Tomasz A, De Lencastre H, Perreten V, Holden MTG, Coleman
460 DC, Goering R, Giffard PM, Skov RL, Zhang K, Westh H, O'Brien F, Tenover FC,
461 Oliveira DC, Boyle-Vavra S, Laurent F, Kearns AM, Kreiswirth B, Ko KS,
462 Grundmann H, Sollid JE, John JF, Daum R, Soderquist B, Buistx G. 2012.
463 Guidelines for reporting novel *mecA* gene homologues. *Antimicrob Agents*
464 *Chemother* 56:4997–4999.
- 465 13. Klemm EJ, Shakoor S, Page AJ, Qamar N, Judge K, Saeed DK, Wong VK.
466 Emergence of an extensively drug-resistant *Salmonella enterica* serovar typhi clone
467 harboring a promiscuous plasmid encoding resistance to fluoroquinolones and third-
468 generation cephalosporins 1–10.
- 469 14. Kozyreva VK, Truong C_H, Greninger A, Crandall J, Mukhopadhyay R C V. 2017.
470 Validation and implementation of clinical laboratory improvements act-compliant
471 whole-genome sequencing in the public health microbiology laboratory. *J Clin*
472 *Microbiol* 55:2502–2520.
- 473 15. Holden, M TG. Hsu, LY. Kurt, K. Weinert, LA. Mather, AE. Harris S, Strommenger,
474 B. Layer F, Witte, W. Lencastre, HD. Skov R, Westh, H. Edgeworth, J. Gould I,
475 Gant, V. Cooke, J. Edwards G, Mcadam, PR. Templeton K, Al E. 2013. A genomic
476 portrait of the emergence, evolution, and global spread of a methicillin-resistant
477 *Staphylococcus aureus* pandemic. *Genome Res* 23:653–664.
- 478 16. Tyler AD, Mataseje L, Urfano CJ, Schmidt L, Antonation KS, Mulvey MR, Corbett
479 CR. 2018. Evaluation of Oxford Nanopore's MinION sequencing device for
480 microbial whole genome sequencing applications. *Sci Rep* 8:1–12.
481

482 **Table 1. Panel of bacterial isolates used in the study.**

Sample name	Accession number	Control or Test isolate	Species	ST	Original study	Transmission clusters	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
SASCBU35	ERR131801	Test isolate	<i>Staphylococcus aureus</i>	22	Harris et al.	Unrelated to Cluster 1	x	x		x	x	x	x	x	x	
SASCBU17	ERR72246	Test isolate	<i>Staphylococcus aureus</i>	2371	Harris et al.	Cluster 1	x	x	x	x	x	x	x	x	x	
SASCBU18	ERR72247	Test isolate	<i>Staphylococcus aureus</i>	2371	Harris et al.	Cluster 1	x	x	x	x	x	x	x	x	x	
SASCBU25	ERR108054	Test isolate	<i>Staphylococcus aureus</i>	2371	Harris et al.	Cluster 1	x	x	x	x	x	x	x	x	x	
MPROS0386	ERR212946	Control isolate	<i>Staphylococcus aureus</i>	22	Coll et al.	Unrelated to Cluster 2	x	x	x (2)	x	x	x	x	x	x	
MPROS1839	ERR715142	Test isolate	<i>Staphylococcus aureus</i>	22	Coll et al.	Cluster 2	x	x	x	x	x		x	x	x	x
MPROS2508	ERR715397	Test isolate	<i>Staphylococcus aureus</i>	22	Coll et al.	Cluster 2	x	x	x	x	x		x	x	x	
MPROS2264	ERR715156	Test isolate	<i>Staphylococcus aureus</i>	22	Coll et al.	Cluster 2	x	x	x	x	x		x	x	x	
MPROS2239	ERR715240	Test isolate	<i>Staphylococcus aureus</i>	22	Coll et al.	Cluster 2			x	x	x		x	x	x	
MPROS0292	ERR212846	Test isolate	<i>Staphylococcus aureus</i>	22	Coll et al.	Cluster 2	x	x		x	x		x	x	x	
MPROS2066	ERR702160	Test isolate	<i>Staphylococcus aureus</i>	30	Coll et al.	Cluster 3	x	x	x	x	x		x	x	x	
MPROS1560	ERR737278	Test isolate	<i>Staphylococcus aureus</i>	30	Coll et al.	Cluster 3			x		x		x	x	x	
MPROS0947	ERR714803	Test isolate	<i>Staphylococcus aureus</i>	30	Coll et al.	Cluster 3	x	x		x (2)	x		x	x	x	
MPROS2402	ERR715316	Test isolate	<i>Staphylococcus aureus</i>	30	Coll et al.	Unrelated to Cluster 3	x	x	x		x		x	x	x	
MPROS0541	ERR702114	Test isolate	<i>Staphylococcus aureus</i>	5	Coll et al.	Cluster 4	x		x	x	x		x	x	x	
MPROS1125	ERR737419	Test isolate	<i>Staphylococcus aureus</i>	5	Coll et al.	Cluster 4	x		x	x	x		x	x	x	
MPROS0046	ERR212783	Test isolate	<i>Staphylococcus aureus</i>	5	Coll et al.	Cluster 4			x	x	x		x	x	x	
MPROS0238	ERR204190	Test isolate	<i>Staphylococcus aureus</i>	5	Coll et al.	Cluster 4			x							
MPROS2412	ERR715326	Test isolate	<i>Staphylococcus aureus</i>	5	Coll et al.	Cluster 4	x		x							
MPROS0158	ERR211966	Test isolate	<i>Staphylococcus aureus</i>	5	Coll et al.	Unrelated to Cluster 4	x		x (2)							
MPROS0688	ERR701921	Test isolate	<i>Staphylococcus aureus</i>	22	Coll et al.	Cluster 5				x	x		x	x	x	
MPROS2335	ERR736981	Test isolate	<i>Staphylococcus aureus</i>	22	Coll et al.	Cluster 5				x	x		x	x	x	
MPROS0659	ERR701905	Test isolate	<i>Staphylococcus aureus</i>	22	Coll et al.	Cluster 6				x	x		x	x	x	
MPROS2044	ERR702173	Test isolate	<i>Staphylococcus aureus</i>	22	Coll et al.	Cluster 6				x	x		x	x	x	
MPROS1689	ERR737479	Test isolate	<i>Staphylococcus aureus</i>	22	Coll et al.	Not applicable	x	x	x	x						
H050960412	HE681097	Test isolate	<i>Staphylococcus aureus</i>	22	Reference strain	Not applicable	x	x	x	x	x	x	x	x	x	
NCTC 6571	ERR1100774	Test isolate	<i>Staphylococcus aureus</i>	Not available	Reference strain	Not applicable	x	x								
NCTC 12241	ERR718772	Control isolate	<i>Escherichia coli</i>	Not applicable	Reference strain	Not applicable			x (2)	x	x	x	x	x	x	x
NCTC 10418	ERS523599	Control isolate	<i>Escherichia coli</i>	Not applicable	Reference strain	Not applicable	x	x								

483

484

485

486

487 **Table 2. Deliberate contamination of controls and MRSA test isolates**

Question	Recipient	Contaminant	Evaluation of impact	Interpretation
Determine the effect of contaminating the no template control with increasing concentrations of MRSA DNA	No template control	Spiked with MRSA MPROS1839 DNA at a final concentration of 0, 0.01 or 0.1%	<ul style="list-style-type: none"> - Number of fragments matching <i>S. aureus</i> were 6, 6 & 20, respectively - Coverage of mapping reference 36.8-46.7% - Average depth 0x - Did not pass QC metric for MRSA 	No template control can tolerate up to 1% contamination with MRSA DNA
	No template control	Spiked with MRSA MPROS1839 DNA at final concentration of 1%	<ul style="list-style-type: none"> - Number of fragments matching <i>S. aureus</i> were 96,671 - Coverage of mapping reference 99.3% - Average depth 7.5x - Did not pass QC metric for MRSA 	
	No template control	Spiked with MRSA MPROS1839 DNA at final concentration of 10% or 20%	<ul style="list-style-type: none"> - Number of fragments matching <i>S. aureus</i> were 363,031 and 623,855, respectively. - Coverage of mapping reference 99.3% (at both concentrations) - Average depth 28.3x and 48.6x, respectively - Both passed QC metrics for MRSA based on depth/coverage, species identification, assignment to ST22 and detection of <i>mecA</i> 	
Determine the effect of contaminating the MRSA control with increasing concentrations of <i>E. coli</i> DNA	MRSA control	Spiked with serial <i>E. coli</i> NCTC12241 DNA at final concentration of 0, 0.01, 0.1, 1, 10 or 20%	<ul style="list-style-type: none"> - MRSA control passed QC metrics at all spikes except 20%, when the proportion of <i>S. aureus</i> genome covered fell from 84.6-91.6% (0-10% contamination) to 77.8% (20% contamination) - Proportion of fragments matching <i>E. coli</i> was 0.44, 4.02 and 8.19 at 1%, 10% & 20%, respectively 	Positive control can tolerate up to 10% contamination with <i>E. coli</i> DNA

488