METHOD
Liou *et al.*, *Microbial Genomics* 2020;6
DOI 10.1099/mgen.0.000336

nanoMLST: accurate multilocus sequence typing using Oxford Nanopore Technologies MinION with a dual-barcode approach to multiplex large numbers of samples

Ci-Hong Liou[1]†, Han-Chieh Wu[1]†, Yu-Chieh Liao[2,*], Tsai-Ling Yang Lauderdale[1], I-Wen Huang[1] and Feng-Jui Chen[1,*]

Abstract

Multilocus sequence typing (MLST) is one of the most commonly used methods for studying microbial lineage worldwide. However, the traditional MLST process using Sanger sequencing is time-consuming and expensive. We have designed a workflow that simultaneously sequenced seven full-length housekeeping genes of 96 meticillin-resistant *Staphylococcus aureus* isolates with dual-barcode multiplexing using just a single flow cell of an Oxford Nanopore Technologies MinION system, and then we performed bioinformatic analysis for strain typing. Fifty-one of the isolates comprising 34 sequence types had been characterized using Sanger sequencing. We demonstrate that the allele assignments obtained by our nanopore workflow (nanoMLST, available at https://github.com/jade-nhri/nanoMLST) were identical to those obtained by Sanger sequencing (359/359, with 100% agreement rate). In addition, we estimate that our multiplex system is able to perform MLST for up to 1000 samples simultaneously; thus, providing a rapid and cost-effective solution for molecular typing.

DATA SUMMARY

The new allele pta_664 was submitted to BIGSdb: https://pubmlst.org/bigsdb?db=pubmlst_saureus_seqdef&page=alleleInfo&locus=pta&allele_id=664. The new allele glpF_732 was submitted to BIGSdb: https://pubmlst.org/bigsdb?db=pubmlst_saureus_seqdef&page=alleleInfo&locus=glpF&allele_id=732. The source code for nanoMLST is available at https://github.com/jade-nhri/nanoMLST. Supplementary material is available at Figshare: https://figshare.com/projects/nanoMLST/69026.

INTRODUCTION

Staphylococcus aureus, a Gram-positive bacterium, is an opportunistic pathogen that can cause community- and healthcare-associated infections including septicaemia, osteomyelitis, endocarditis, toxic shock syndrome, and skin and soft tissue infections [1, 2]. Meticillin resistance in *S. aureus* denotes resistance to virtually all available β-lactam agents; thus, meticillin-resistant *S. aureus* (MRSA) has been a major public-health problem around the world, and causes severe morbidity and mortality. Fast and accurate molecular typing methods are essential for effective surveillance and outbreak detection, as well as monitoring of the evolution and dynamics of MRSA clones.

Several molecular typing methods, including PFGE, staphylococcal protein A (*spa*) typing and multilocus sequence typing (MLST), are commonly used to study the epidemiology of *S. aureus* and support infection control measures [3–5]. PFGE, a technique used for the separation of large DNA on a gel, is a gold standard for genotyping. However, it is time-consuming, technically cumbersome and presents difficulties for comparing results from different laboratories [6]. *spa*-typing based on the sequence of a polymorphic variable number tandem repeat in the 3′ coding region X of the *spa* gene is a rapid and simple method [4]. Nevertheless, there are some non-typeable strains, due to deletions in the *spa* gene, and a reliable evolutionary history

Received 24 September 2019; Accepted 21 January 2020; Published 17 February 2020
Author affiliations: [1]National Institute of Infectious Diseases and Vaccinology, National Health Research Institutes, Zhunan, Taiwan, ROC; [2]Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan, ROC.
***Correspondence:** Feng-Jui Chen, frchen@nhri.edu.tw; Yu-Chieh Liao, jade@nhri.org.tw
Keywords: multilocus sequence typing (MLST); meticillin-resistant *Staphylococcus aureus* (MRSA); dual-barcode multiplexing; nanopore; MinION; molecular typing.
Abbreviations: CCS, circular consensus sequence; MLST, multilocus sequence typing; MRSA, meticillin-resistant *Staphylococcus aureus*; ONT, Oxford Nanopore Technologies; PacBio, Pacific Biosciences; ST, sequence type.
†These authors contributed equally to this work
Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Supplementary material is available with the online version of this article.

cannot be inferred from *spa* types of less than five repeats in length [7–9]. MLST, developed originally for *Neisseria meningitidis* [10], is now recognized as an unambiguous procedure for characterizing isolates of many bacterial and fungal species based on sequences of ~450 bp internal fragments of seven to ten housekeeping genes [5, 10–12]. MLST uses specific primer sets designed for PCR amplification of multiple housekeeping genes, followed by sequencing of the amplicons. Each sequence is then compared to the MLST database to obtain an allele number, with a single nucleotide difference being assigned a different allele number. The combination of the allele numbers determines the sequence type (ST) of the isolate. Such a conventional method is costly, time-consuming and labour-intensive. Next-generation sequencing technologies, including Roche 454 [13–15], Illumina [16] and Pacific Biosciences (PacBio) [17] sequencing, have been used for high-throughput and cost-effective MLST. Roche 454 and Illumina sequencing are so-called second-generation sequencing technologies and they can produce a massive number (over a million), short (400–500 bp for Roche 454) and accurate sequencing reads [18]. Roche 454 has been discontinued and the Illumina system has become the mainstream method in second-generation sequencing [19], but Illumina only can produce reads up to 2×300 bp, which restricts the size of amplicon sequences to 400–500 bp.

PacBio and Oxford Nanopore Technologies (ONT) systems are third-generation sequencing technologies that can produce long reads but with low accuracy (80–90 %). The high error rates suggested that MLST analysis would be challenging. Cao *et al*. tried to carry out MLST for *Klebsiella pneumoniae* using ONT MinION sequence data, but were unable to differentiate STs because the MLST profiles differ at even a single nucleotide in the seven housekeeping genes [20]. Thus, they proposed a novel gene-presence typing model for *K. pneumoniae*, *Escherichia coli* and *S. aureus* based on genome assemblies and the corresponding STs. However, this approach was restricted to hundreds of STs (107 for *S. aureus,* 125 for *K. pneumoniae* and 353 for *Escherichia coli*) and was only validated on several samples with whole-genome sequencing [20]. Although whole-genome sequencing of vancomycin-resistant *Enterococcus faecium* with MinION was also successfully used for MLST, the types of allele were only confirmed for two clinical isolates [21]. By contrast, amplicon sequencing allows large-scale MLST. To assuage the error of long reads, Chen *et al*. used PacBio circular consensus sequence (CCS) reads with higher accuracy due to multiple passes to perform MLST on *Cryptococcus neoformans* and *Cryptococcus gattii* [17]. Multiple full passes of reads limit the size of CCS reads to <2.5 kb, but the mean read length of continuous long reads of PacBio is over 10 kb [22]. In addition, the number of CCS reads is much less than all the output of PacBio data, for example, only 37 906 CCS reads (≥4 passes) were yielded from one SMRT cell of the PacBio RS II platform [17]. In comparison, ONT MinION is portable, requires lower amounts of DNA and is low cost ($1000 US dollars/£768, where £1=$1.30). The average MinION yield has increased from 500 Mbp to 20 Gbp in a single run (based on our experiences) since the MinION Access Program (MAP) started in 2014. With the advantage of high-throughput (millions of sequencing reads) and easy access to the MinION device, nanopore sequencing could be a potential candidate to replace Sanger sequencing and offer a relatively low-cost genotyping method.

Impact Statement

With the advantages of high-throughput and decreasing cost of next-generation sequencing, this technology has been introduced into molecular typing workflows. However, Illumina sequencing can only produce reads up to 2×300 bp, which restricts the size of amplicon sequences to 400–500 bp, and Oxford Nanopore Technologies (ONT) sequencing produces long reads but with low accuracy. Although highly accurate (>99.8 %) consensus sequences can be obtained by polishing ONT sequences, the homopolymer errors presented in the consensus sequences still impede accurate sequence typing. Here, we not only demonstrate the use of a dual-barcode approach to multiplex 96 meticillin-resistant *Staphylococcus aureus* isolates for sequencing seven housekeeping genes with lengths longer than 1000 bp, but also provide a bioinformatic workflow starting with read pre-processing, demultiplexing and consensus sequence generation to finally perform accurate multi-locus sequence typing (100%). This study paves the way for cost-effective and accurate sequencing of long amplicons.

In this study, to fully exploit the advantage of the MinION sequencer, a high-throughput of long reads, we designed a pipeline that utilizes a dual-barcoding system to sequence seven full-length housekeeping genes of *S. aureus* for MLST of 96 MRSA isolates in a single cell. We then polished the nanopore sequencing reads to generate consensus sequences and corrected homopolymer errors to obtain accurate MLST alleles.

THEOY AND IMPLEMENTATION

Bacterial isolates

The *S. aureus* isolates used in the present study were from the Taiwan Surveillance of Antimicrobial Resistance (TSAR), a national surveillance programme of inpatient and outpatient clinical isolates in Taiwan, ROC [23]. Species identification and antimicrobial-susceptibility testing of *S. aureus* isolates have been described previously [24]. A total of 96 MRSA isolates were used in this study, 51 of which comprised 34 sequencing types determined by Sanger sequencing were selected for validation.

Universal primer design and DNA amplification

In order to design primers for full-length PCR amplification, we analysed the flanking sequences of seven housekeeping genes used in conventional MRSA MLST [5], including genes encoding carbamate kinase (*arcC*), shikimate dehydrogenase (*aroE*), glycerol kinase (*glpF*), guanylate kinase (*gmk*), phosphate acetyltransferase (*pta*), triosephosphate isomerase (*tpiA*) and acetyl coenzyme A acetyltransferase (*yqiL*), from 10 *S. aureus* strains whose genome sequences have been deposited in the National Center for Biotechnology Information (NCBI). The GenBank accession numbers of the 10 *S. aureus* strains are as follows: NCTC8325 (CP000253.1), RF122 (AJ938182.1), ST398 (AM990992.1), N315 (BA000018.3), MW2 (BA000033.2), MRSA252 (BX571856.1), MSSA476 (BX571857.1), M013 (CP003166.1), LGA251 (FR821779.1) and Z172 (CP006838.1). The flanking sequences were aligned using Multalin [25] to obtain consensus regions for primer design. The seven sequences for theoretical PCR amplicons based on the *S. aureus* NCTC8325 genome are included in Supplementary file 1 (available with the online version of this article). The bacterial template DNA was prepared using DNAzol Direct (Molecular Research Center), following the manufacturer's instructions. Seven housekeeping genes of a 25 µl sample were separately amplified using the Thermo Scientific Phusion high-fidelity DNA polymerase (Thermo Fisher) for PCRs using different primers with a specified PCR barcode sequence as shown in Supplementary file 1. In this case, eight samples can be distinguished with eight PCR barcodes (PB01–PB08) (Fig. 1). Thermal cycling reactions consisted of an initial denaturation (98 °C for 30 s); 35 cycles of denaturation (98 °C for 10 s), annealing (55 °C for 30 s) and extension (72 °C for 1 min); and a single final extension (72 °C for 10 min). All PCR products were checked and quantified on an agarose gel.

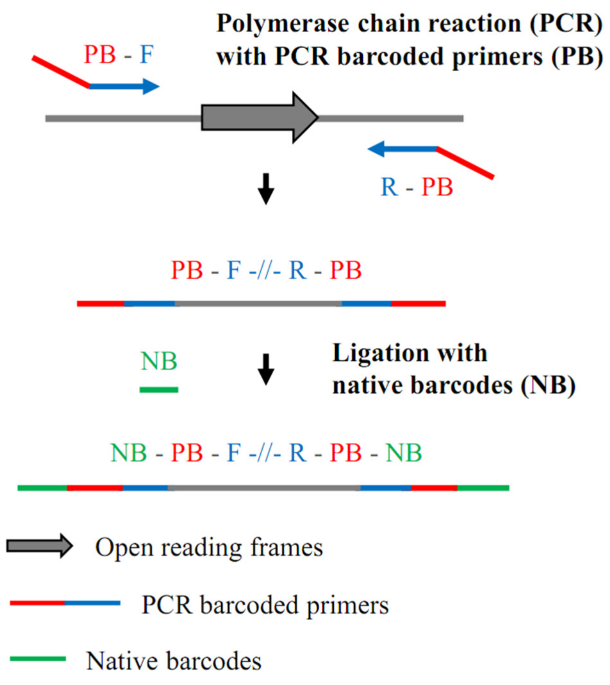
Polymerase chain reaction (PCR) with PCR barcoded primers (PB)
PB - F
R - PB
PB - F -//- R - PB
NB
Ligation with native barcodes (NB)
NB - PB - F -//- R - PB - NB
Open reading frames
PCR barcoded primers
Native barcodes

Fig. 1. Schematic illustration of library construction with dual-barcodes for multiplexing samples. In this study, eight PCR barcodes (PB01–PB08) and twelve native barcodes (NB01–NB12) were used for multiplexing 96 samples.

Library preparation and MinION sequencing

The seven amplicons from the same template were pooled together and every eight pooled amplicons with unique barcode sequences (PB01– PB08) were then pooled together. After 12 pools were collected for the 96 samples, DNA was purified using the Agencourt AMPure XP system (Beckman Coulter) for DNA library preparation. The DNA library was constructed according to the 1D native barcoding genomic DNA protocol (Oxford Nanopore Technologies) using the native barcoding expansion 1–12 (EXP-NBD103) and ligation sequencing kit 1D (SQK-LSK109). One MinION flow cell (FLO-MIN106, R9.4.1) was used in this study and sequencing was carried out on a MinION sequencer using the software program MinKNOW (v2.1) in the standard 48 h sequencing script without basecalling option.

Read preparation

After the sequencing run, Albacore (v2.3.1) was used for FASTQ extraction without demultiplexing. One megabase pair is an excessively high estimation to cover the size of the sequencing region (7×96×1300 bp) for amplicons of seven full-length housekeeping genes (~1300 bp) of 96 MRSA isolates. We filtered reads with the minimum quality value of 10 (mean_qscore_template) and the minimum length of 1000 bp (sequence_length_template) to obtain Q10L1000-filtered reads with a mean depth of 1000× (i.e. 1 Gbp) with a custom script getfastq.py. Information regarding the number of reads, total bases and mean length was calculated using fqstats.py. The code is available at https://github.com/jade-nhri/nanoMLST, and the detailed instructions are given in Supplementary file 1.

Demultiplexing with dual-barcodes

The ONT native barcoding kit was used to multiplex the 12 pools (NB01–NB12), each containing eight sets of pooled amplicons (PB01–PB08), resulting in 96 dual-barcodes (listed in Supplementary file 1) for the 96 samples. Please note that our PCR barcodes were designed based on the sequences of native barcodes. After initial examination of sequencing reads, we found that the native barcode kit attached the reverse complement barcode sequence along with a sequence of 'CAGCACCT' to the 5′ end of amplicons. The 96 dual-barcode sequences were generated using generatebcs.py [e.g. NB01PB01: rc(NB01)+CAGCACCT+(PB01)]. We mapped the sequencing reads to the dual-barcode sequences with 56 bp in length using Minimap2 (v2.11) [26] by -k7 -A1 -m42 -w1 options. With a pairwise read mapping format (PAF) file outputted by Minimap2, we generated 96 sequencing files (e.g. NB01PB01.fq) to include the corresponding sequencing reads for a specific dual-barcode using getbcfq.py.

Consensus sequence generation

With the reference sequences of seven housekeeping genes of *S. aureus* NCTC8325 (Supplementary file 1) and the demultiplexed reads, consensus sequences of each sample were generated by runcons.py. Minimap2 was used to map the demultiplexed reads to the reference sequences. Racon (v1.3.1) [27] and Nanopolish (v0.10.2) [28] were then used iteratively to polish the sequencing reads for consensus sequence generation. Since Nanopolish requires raw fast5 for signal-level analysis, before running runcons.py, we binned the fast5 files into 96 dual-barcode folders using binf5.py. In runcons.py, we firstly polished the reference sequences with Racon, and then polished the Racon-produced sequences with Nanopolish iteratively two times. If the two file sizes of Nanopolish-produced sequences were the same, the file of second-run Nanopolish was renamed, i.e. '_final' was suffixed to the file name. If the two file sizes were different, additional Racon was run twice based on the Racon-produced sequences, and then Nanopolish was run iteratively until the two sequential file sizes of Nanopolish-produced sequences were identical.

MLST typing

Seven consensus sequences for each sample were compared with the *S. aureus* MLST allele sequences to correct homopolynucleotide problems, to call alleles and to profile a ST using runtyping.py. In running runtyping.py, the seven allele sequences of housekeeping genes were downloaded automatically (e.g. *arcC* from https://pubmlst.org/data/alleles/saureus/arcC.tfa). The downloaded nucleotide sequences were translated to protein sequences to exclude the sequences having internal stop codons. The consensus sequences generated by runcons.py were aligned to the remaining allele sequences using BLAST [29] to correct homo-polynucleotide errors. A detailed description is provided in Supplementary file 1. After homopolymer correction, an allele number for each housekeeping gene was determined if the consensus sequence had an identical sequence to the allele sequence. For each sample, an allelic profile from the specific combination of seven alleles was obtained to finally call a ST based on the available profiles of *S. aureus* MLST from the PubMLST website (https://pubmlst.org/data/profiles/saureus.txt).

Simulation study

Based on our experiences with nanopore sequencing, we assumed that 80× coverage was sufficient to generate accurate consensus sequences [30]. We randomly sampled 560 sequencing reads (80× coverage of seven genes) three times. Sampling reads for each sample underwent the whole process to get a corresponding MLST type. MLST alleles obtained from sampling reads were compared with those obtained from the Q10L1000-filtered reads. To determine a reasonable sample size for running one MinION flow cell, we sampled reads based on the distribution (black bars in Fig. 2) using samplingreads.R. Briefly, the number of basecalled reads was divided by 1 to 50, then the numbers of rounding quotients were used for sampling reads of 96 samples. Read numbers for 96×1, 96×2, 96×3 samples, etc. were obtained accordingly. The number of samples with reads of less than 560 was determined. This simulation was conducted three times to get the mean percentage of samples having reads ≥560.

Workflow of nanopore sequencing in dual-barcode multiplexing

A dual-barcode multiplexing system was proposed in this study (Fig. 1). Seven primer pairs of *S. aureus* housekeeping genes coupled with PCR barcodes were used for PCR

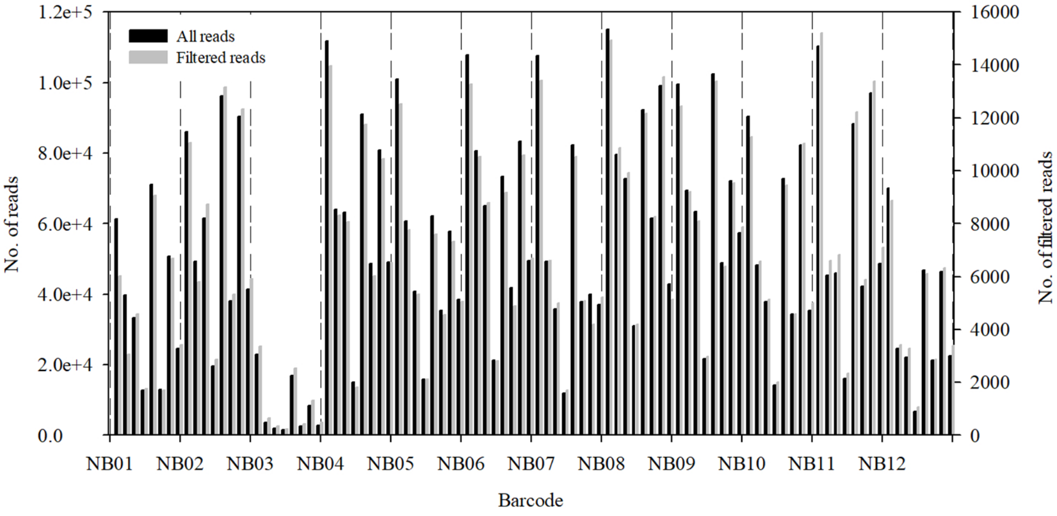

Fig. 2. Uneven read distributions of dual-barcode samples. Number of demultiplexed reads (black bars, left axis) and number of filtered and demultiplexed reads (grey bars, right axis) for 96 dual-barcode strains.

Table 1. Summary of the nanopore sequencing results

	Basecalled reads	Q10L1000-filtered reads*
No. of reads	6336419	675634
Total bases (bp)	8439547955	1000000903
Mean length (bp)	1332	1480
No. of demultiplexed reads	5008777	656979
No. of reads per dual-barcode		
Minimum	1501	239
Maximum	115048	15216
Mean	52175	6844
SD	30371	3926

*One gigabase pair reads are a subset of all filtered reads. Q10L1000-filtered reads (4.25 Gbp) are the reads with the minimum quality value of 10 (mean_qscore_template) and the minimum length of 1000 bp (sequence_length_template). The 675 634 filtered reads (reads.fastq) are available from Figshare (https://doi.org/10.6084/m9.figshare.9891455).

amplification (the primer sequences are listed in Supplementary file 1). All PCR products were checked and quantified by gel electrophoresis to get equal amounts of amplicons. Seven amplicons from the same template were pooled together and every eight pooled amplicons with PCR barcodes from PB01 to PB08 were then pooled together. For the 96 MRSA strains, 12 pools were collected and multiplexed with the native barcoding kit to include outer barcodes (NB01–NB12). Therefore, a total of 672 amplicons for 96 dual-barcode multiplexing MRSA strains were sequenced with MinION. A total of 7317286 raw sequencing files (fast5) were produced by the software program MinKNOW. The sequencing reads were basecalled using Albacore to produce a fastq file containing 6336419 passing-filter reads. In initial analysis of sequencing reads, the barcode sequences were found to be present in both ends (~100 bp) of sequencing reads. A dual-barcode sequence pattern included a reverse complement sequence of native barcode (24 bp), followed by a padding sequence of 'CAGCACCT' and a sequence of PCR barcode (24 bp). With the designed dual-barcode sequences of 56 bp in length (Supplementary file 1), sequencing reads were demultiplexed.

Nanopore data analysis

As shown in Table 1, over six million reads were basecalled and passed filtering by Albacore, resulting in the sum of 8.44 Gbp of sequencing reads with a mean length of 1332 bp. By mapping reads to the 96 dual-barcode sequences using Minimap2, we successfully demultiplexed 5008777 reads into 96 dual-barcodes with a mean read number of 52175 (Table 1, Fig. 2, black bars), but with an uneven distribution. Sequences were selected initially by length (≥1000 bp) and quality (≥10) with 1000× coverage to obtain 675634 Q10L1000-filtered reads with a mean length of 1480 bp. Please note that around half the amount of sequencing reads (4.25 Gbp) can be selected on the criteria of length (≥1000 bp) and quality (≥10), but only 1 Gbp was used for further analysis because the total sequencing amount is too high for 96 samples. Among the 675364 Q10L1000-filtered reads, 656979 reads were successfully demultiplexed into 96 dual-barcoded files (Table 1, Fig. 2, grey bars). As expected, the mean read number of Q10L1000-filtered reads per dual-barcode is close to 7000 (1000× coverage of seven genes), while on the contrary the minimum read number is as few as 239 for NB03PB04. Such a small number of reads makes it difficult for subsequent consensus sequence generation. Therefore, we increased the depth option to 3000× in getfq.py to get more sequencing reads for NB03. In this circumstance, the number of reads for NB03PB04 was increased to 579. Filtered and demultiplexed reads for each strain were subsequently mapped to the seven reference genes (Supplementary file 1) using Minimap2, and finally used to generate seven consensus sequences for each strain with Racon and Nanopolish.

MLST analysis

Through the above-mentioned processes, basecalling, demultiplexing and consensus sequence generation, 96 sequence files with '_final' as the file name suffix were produced. Although Racon and Nanopolish were run iteratively to improve the sequencing errors produced by ONT MinION, there were inevitable homopolymer errors present in the consensus sequences. Therefore, we used the allele sequences of the seven *S. aureus* housekeeping genes in PubMLST to correct homo-polynucleotide errors and then to call an allele number for each housekeeping gene. Among the 672 (7 genes×96 samples) alleles, only two genes (NB05PB07's *pta* and NB05PB08's *glpF*) were not assigned with an allele number because their sequences did not match the allele sequences downloaded from PubMLST, which suggests they are new alleles. The ST for each strain based on the specific combination of seven alleles was acquired.

The 96 MRSA isolates were assigned to 41 STs (Fig. 3, Supplementary file 2), including 5 new STs for 6 strains: new allele profile 1 for NB01PB02–151, 36, 321, 34, 256, 261, 323; new allele profile 2 for NB05PB07–2, 3, 513, 1, new *pta*, 4, 3; new allele profile 3 for NB05PB08 – 19, 23, new *glpF*, 2, 19, 20, 15; new allele profile 4 for NB06PB08 and NB12PB08–19, 23, 15, 2, 19, 20, 3; new allele profile 5 for NB10PB08–19, 23, 15, 2, 19, 20, 10. The two alleles (new *pta* and new *glpF*) were later Sanger sequenced (Supplementary file 2) and their alignments showed 100% identities to our nanopore consensus sequences (Supplementary file 1). These two allele sequences were submitted to BIGSdb [12] and assigned the allele numbers 664 and 732 for *pta* and *glpF*, respectively. The nanopore sequencing results were compared to the 51 MRSA strains that were Sanger sequenced to include a total of 357 alleles (Supplementary file 2). In this study, we found that allele assignments using nanopore and Sanger sequencing technologies were all consistent [(357+2)/(357+2), with 100% agreement rate). These results indicated that MinION

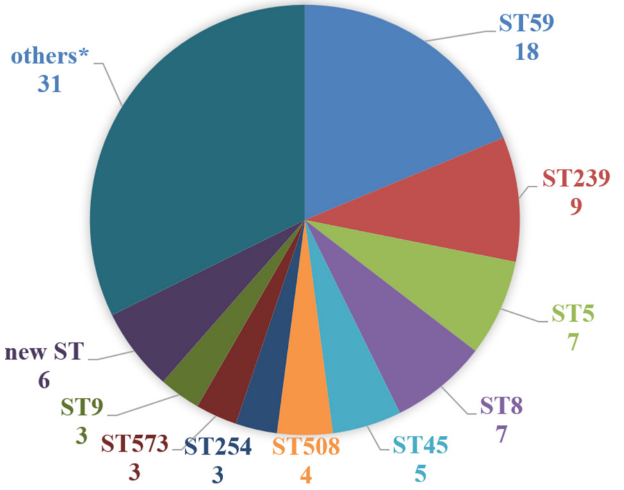

Fig. 3. MLST profile of nanopore sequencing. The number of samples is shown under the ST. *STs whose number of samples is less than three are grouped as 'others', including ST1, ST6, ST7, ST12, ST30, ST78, ST89, ST97, ST158, ST188, ST241, ST338, ST398, ST444, ST777, ST900, ST2123, ST2339, ST2340, ST2341, ST2342, ST2343, ST2344, ST2345, ST2346, ST2347 and ST2348.

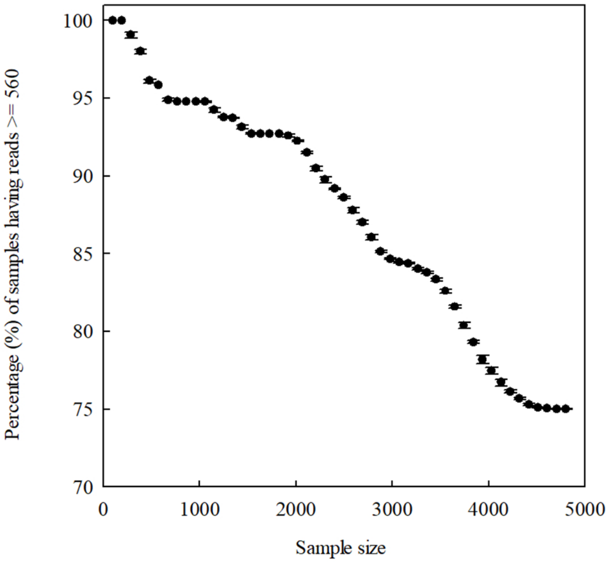

Fig. 4. Relationship between the percentage of samples possessing 560 reads or more and sample size.

sequencing along with our proposed analysis is able to perform MLST accurately.

Simulation study analysis

To demonstrate the feasibility of the application of MinION sequencing for MLST, Q10L1000-filtered reads were initially selected and analysed. Next, we randomly sampled a specific number of reads to go through the analysis. Although the number of sequencing reads for NB03PB04 was as few as 579, it was accurately typed as ST1 (identical to Sanger sequencing). Therefore, we sampled 560 sequencing reads (80× coverage of seven genes) from the basecalled and demultiplexed reads for each barcode three times. Each barcoded sample contained 560 sequencing reads, and these reads were used for consensus sequence generation and homopolymer correction to finally get assignment of alleles. To compare with the allele assignments acquired from analysing Q10L1000-filtered reads, barely 8, 11 and 12 (out of 672) discrepant alleles were found in the three simulations, which suggests that only 560 sequencing reads were sufficient for accurate consensus sequence generation (over 98% accuracy) of seven housekeeping genes. It appears that the throughput of MinION sequencing (over six million reads) was extremely high for performing MLST for 96 samples. Thus, we sampled reads based on the uneven distribution (black bars in Fig. 2) to see how many samples possess more than 560 sequencing reads when sample size increases in a MinION flow cell. As can be seen in Fig. 4, it still maintains approximately 95% of samples having more than 560 reads as sample size increases to over 1000. Therefore, the evidence points to the probability that the sample size for this application can be increased to 1000. With our proposed dual-barcode system, thousands of samples can be indexed effectively in place of tagging thousands of single barcodes. It may be reasonable to suppose that our system can generate close to 7000 accurate allele assignments using a MinION flow cell.

DISCUSSION

MLST is a portable, reproducible and unambiguous molecular typing technology applied to many pathogenic microorganisms, including prokaryotic bacteria and eukaryotic species [31]. The conventional Sanger sequencing method is costly and time-consuming. Thus, next-generation sequencing platforms such as Roche 454 [13–15], Illumina [16] and PacBio [17] have been used to perform high-throughput and cost-effective MLST. However, these studies either utilized massive but short reads for sequencing amplicons with a length of 400–500 bp or utilized a small number of long reads with higher accuracy (CCS) to barely perform MLST for 96 isolates (10.7% of alleles had less than three reads). In this study, we have successfully produced 96×7 MLST alleles using MinION and validated 359 alleles entirely consistent with Sanger sequencing. In addition, we have found that the sequencing reads produced in this study were sufficient for the molecular typing of 1000 samples. Recently, Srivathsan *et al.* also proposed a dual-index approach and stated that up to 1000 barcodes could be generated [32]. However, instead of using 1D template reads (as we used in this study) they used $1D^2$ and 2D reads for their analyses. Like CCS in PacBio, $1D^2$ and 2D are more accurate than 1D reads, but their sequences are short in length and their throughputs are lower than 1D reads in ONT; in addition, 2D reads are no longer available. Therefore, our approach is more practical for future use. ONT has released a new type of flow cell called 'Flongle', which can deliver up to 1.8 Gbp of sequencing data. Although Flongle is yet to be tested, we expect its capability to perform MLST for 96 isolates.

Illumina MiSeq is the only system that can produce 2×300 bp reads for sequencing 450–550 bp amplicons. It produces a maximum of 50 million reads per run (https://www.illumina.com/systems/sequencing-platforms.html) and, thus, is believed to be capable of typing thousands of samples. However, the length of amplicons is limited to 550 bp. In this study, we have totally sequenced 672 full-length genes with a mean length of 1300 bp using MinION for 96 samples, which could provide extra phylogenetic information for subtyping. By using the Q10L1000-filtered reads, we obtained plenty of reads (mean 6844) for each sample. The nanopore reads (with high error rate) were polished iteratively with Racon and Nanopolish to generate consensus sequences, the accuracy of the consensus sequences was higher than 99.8% [30], but we found that there were homopolymer errors in consensus sequences notwithstanding. Similar patterns were also observed elsewhere [21], i.e. a few deleted bases in repeat sequences. We assumed that frameshift mutations in housekeeping genes encoding an abnormal function of a protein would be rare. We downloaded the *S. aureus* MLST allele sequences from PubMLST and excluded the sequences having internal stop codons. Among the 4148 allele sequences downloaded in August 2018, as few as 45 sequences were ruled out. By assuming the homopolymer differences between the allele sequences and our consensus sequences were ONT sequencing's intrinsic errors, we corrected consensus sequences accordingly. Although this correction resulted in the perfect agreement of alleles, it is worth noting that the homopolymer correction is on the basis of allele sequences (450–500 bp) deposited in PubMLST; sequences are not corrected for the full length. In addition, a true frameshift mutation may be overlooked by using this approach. Nevertheless, to the best of our knowledge, this is the first study to perform high-throughput MLST accurately using ONT MinION.

Conclusion

We have designed primer sequences incorporating PCR barcodes (Supplementary file 1) ligated with native barcodes to form a dual-barcode system for sample multiplexing. However, this approach used 8×7×2=112 primers. As presented in the knowledge exchange platform of the ONT community (13th November 2018), a dual-barcoding system was proposed by combining native barcodes (EXP-NBD104) with PCR barcodes (EXP-PBC096) to expand multiplexing capabilities up to 2304 combinations. In this circumstance, only 7×2=14 primers would be required for *S. aureus* MLST. Future work should, therefore, include effective primer design. ONT provides qcat (https://github.com/nanoporetech/qcat) for demultiplexing dual-barcoding datasets. We used qcat to demultiplex the 6 336 419 basecalled reads, but only 37194 reads were demultiplexed, which suggests that our dual-barcode sequence patterns may not be identical to those of ONT. Nevertheless, in this study, we have implemented dual-barcode demultiplexing effectively using Minimap2 and our custom scripts. We have also demonstrated that the demultiplexed reads for each sample were successfully polished and corrected to produce accurate STs, which would make our MLST approach a rapid and cost-effective way for molecular typing.

Funding information
This work was supported by intramural grants from the National Health Research Institutes, Taiwan, ROC (IV-107-PP-07 to F.-J.C. and PH-107-PP-05 to Y.-C.L.), and the Ministry of Science and Technology, Taiwan, ROC (MOST 106–2923-B-400–001-MY3).

Acknowledgements
We would like to thank the National Core Facility for Biopharmaceuticals (NCFB, MOST 106–2319-B-492–002) and the National Center for High-performance Computing (NCHC) of the National Applied Research Laboratories (NARLabs) in Taiwan for providing computational resources.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Data Bibliography

Liou C-H, pta_664 allele information, https://pubmlst.org/bigsdb?db=pubmlst_saureus_seqdef&page=alleleInfo&locus=pta&allele_id=664 (2019)

Liou C-H, glpF_732 allele information, https://pubmlst.org/bigsdb?db=pubmlst_saureus_seqdef&page=alleleInfo&locus=glpF&allele_id=732 (2019)

Liao Y-C, GitHub of nanoMLST, https://github.com/jade-nhri/nanoMLST (2019)

Liao Y-C, figshare of nanoMLST, https://figshare.com/projects/nanoMLST/69026 (2019)

References

1. Lowy FD. *Staphylococcus aureus* infections. *N Engl J Med* 1998;339:520–532.
2. Tong SYC, Davis JS, Eichenberger E, Holland TL, Fowler VG. *Staphylococcus aureus* infections: epidemiology, pathophysiology, clinical manifestations, and management. *Clin Microbiol Rev* 2015;28:603–661.
3. Tenover FC, Arbeit R, Archer G, Biddle J, Byrne S *et al*. Comparison of traditional and molecular methods of typing isolates of *Staphylococcus aureus*. *J Clin Microbiol* 1994;32:407–415.
4. Shopsin B, Gomez M, Montgomery SO, Smith DH, Waddington M *et al*. Evaluation of protein A gene polymorphic region DNA sequencing for typing of *Staphylococcus aureus* strains. *J Clin Microbiol* 1999;37:3556–3563.
5. Enright MC, Day NPJ, Davies CE, Peacock SJ, Spratt BG. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. *J Clin Microbiol* 2000;38:1008–1015.
6. Murchan S, Kaufmann ME, Deplano A, de Ryck R, Struelens M *et al*. Harmonization of pulsed-field gel electrophoresis protocols for epidemiological typing of strains of methicillin-resistant *Staphylococcus aureus*: a single approach developed by consensus in 10 European laboratories and its application for tracing the spread of related strains. *J Clin Microbiol* 2003;41:1574–1585.
7. Lee GH, Pang S, Coombs GW. Misidentification of *Staphylococcus aureus* by the Cepheid Xpert MRSA/SA BC assay due to deletions in the *spa* gene. *J Clin Microbiol* 2018;56:e00530-18.
8. Baum C, Haslinger-Loffler B, Westh H, Boye K, Peters G *et al*. Non-*spa*-typeable clinical *Staphylococcus aureus* strains are naturally occurring protein A mutants. *J Clin Microbiol* 2009;47:3624–3629.
9. Hudson LO, Murphy CR, Spratt BG, Enright MC, Terpstra L *et al*. Differences in methicillin-resistant *Staphylococcus aureus* strains isolated from pediatric and adult patients from hospitals in a large county in California. *J Clin Microbiol* 2012;50:573–579.

10. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE *et al*. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA* 1998;95:3140–3145.
11. Muñoz M, Camargo M, Ramírez JD. Estimating the intra-taxa diversity, population genetic structure, and evolutionary pathways of *Cryptococcus neoformans* and *Cryptococcus gattii*. *Front Genet* 2018;9:148.
12. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* 2018;3:124.
13. Boers SA, van der Reijden WA, Jansen R. High-throughput multilocus sequence typing: bringing molecular typing to the next level. *PLoS One* 2012;7:e39630.
14. Singh P, Foley SL, Nayak R, Kwon YM. Multilocus sequence typing of *Salmonella* strains by high-throughput sequencing of selectively amplified target genes. *J Microbiol Methods* 2012;88:127–133.
15. Takahashi H, Iwakawa A, Ohshima C, Kyoui D, Kumano S *et al*. A rapid typing method for *Listeria monocytogenes* based on high-throughput multilocus sequence typing (Hi-MLST). *Int J Food Microbiol* 2017;243:84–89.
16. Zhang N, Wheeler D, Truglio M, Lazzarini C, Upritchard J *et al*. Multi-locus next-generation sequence typing of DNA extracted from pooled colonies detects multiple unrelated *Candida albicans* strains in a significant proportion of patient samples. *Front Microbiol* 2018;9:1179.
17. Chen Y, Frazzitta AE, Litvintseva AP, Fang C, Mitchell TG *et al*. Next generation multilocus sequence typing (NGMLST) and the analytical software program MLSTEZ enable efficient, cost-effective, high-throughput, multilocus sequencing typing. *Fungal Genet Biol* 2015;75:64–71.
18. Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. *Genomics* 2016;107:1–8.
19. Pérez-Losada M, Arenas M, Castro-Nallar E. Microbial sequence typing in the genomic era. *Infect Genet Evol* 2018;63:346–359.
20. Cao MD, Ganesamoorthy D, Elliott AG, Zhang H, Cooper MA *et al*. Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time MinION(TM) sequencing. *Gigascience* 2016;5:32.
21. Tarumoto N, Sakai J, Sujino K, Yamaguchi T, Ohta M *et al*. Use of the Oxford Nanopore MinION sequencer for MLST genotyping of vancomycin-resistant enterococci. *J Hosp Infect* 2017;96:296–298.
22. Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res* 2018;46:2159–2168.
23. Ho M, McDonald LC, Lauderdale TL, Yeh LL, Chen PC *et al*. Surveillance of antibiotic resistance in Taiwan, 1998. *J Microbiol Immunol Infect* 1999;32:239–249.
24. Chen FJ, Huang IW, Wang CH, Chen PC, Wang HY *et al*. *mecA*-positive *Staphylococcus aureus* with low-level oxacillin MIC in Taiwan. *J Clin Microbiol* 2012;50:1679–1683.
25. Corpet F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* 1988;16:10881–10890.
26. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–3100.
27. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res* 2017;27:737–746.
28. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* 2015;12:733–735.
29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
30. Liao Y-C, Cheng H-W, Wu H-C, Kuo S-C, Lauderdale T-LY *et al*. Completing circular bacterial genomes with assembly complexity by using a sampling strategy from a single MinION run with barcoding. *Front Microbiol* 2019;10:2068.
31. Urwin R, Maiden MCJ. Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol* 2003;11:479–487.
32. Srivathsan A, Baloğlu B, Wang W, Tan WX, Bertrand D *et al*. A MinION™-based pipeline for fast and cost-effective DNA barcoding. *Mol Ecol Resour* 2018;18:1035–1049.