

Resolving a neonatal intensive care unit outbreak of methicillin-resistant *Staphylococcus aureus* to the SNV level using Oxford Nanopore simplex reads and HERRO error correction

5 Authors

Max Bloomfield^{1,2}, Sarah Bakker³, Megan Burton¹, M. Leticia Castro³, Kristin Dyet³, Alexandra Eustace³, Samantha Hutton¹, Donia Macartney-Coxson³, William Taylor⁴, Rhys T. White³

10 Affiliation(s)

¹Awanui Labs Wellington, Department of Microbiology and Molecular Pathology, Wellington 6021, New Zealand

²Te Whatu Ora/Health New Zealand, Infection Prevention and Control, Capital, Coast & Hutt Valley, Wellington 6021, New Zealand

15 ³Institute of Environmental Science and Research, Health Group, Porirua 5022, New Zealand

⁴Institute of Environmental Science and Research, Environment Group, Christchurch 8041, New Zealand

Corresponding author and email address

20 *Corresponding author: Rhys White, Institute of Environmental Science and Research, Porirua, New Zealand; Telephone: +64-4-914-0700; E-mail: rhys.white@esr.cri.nz

Keywords

25 Genomic surveillance; outbreak detection; phylogenetic analysis; infection control; single-nucleotide variants

Abstract

Objectives: Our laboratory began prospective genomic surveillance for healthcare-associated organisms in 2022 using Oxford Nanopore Technologies (ONT) sequencing as a standalone platform. This has permitted the early detection of outbreaks but has been insufficient for single-nucleotide variant (SNV)-level analysis due to lower read accuracy than Illumina sequencing. This study aimed to determine whether Haplotype-aware ERRor cORrection (HERRO) of ONT data could permit high-resolution comparison of outbreak isolates.

Methods: We used ONT simplex reads from isolates involved in a recent outbreak of methicillin-resistant *Staphylococcus aureus* (MRSA) in our neonatal unit. The raw sequence data were re-basecalled and adapter-trimmed using Dorado v0.7.0. The simplex reads then underwent HERRO correction. The resulting genome assemblies and phylogenies were compared with previous analyses (using Dorado v0.3.4, no HERRO correction and data generated by Illumina sequencing).

Results: Five of nine outbreak isolates were included in the analysis. The remaining four isolates had insufficient read lengths (N50 values <10,000 bp) and did not provide complete chromosome coverage after HERRO correction. The average chromosome sequencing depth for nanopore data was 147× (range: 44–220×) with an average read N50 of 12,215 bp (interquartile range (IQR): 11,439–12,711 bp). The median pairwise SNV distance between outbreak isolates from the original investigation was 51 SNVs (range: 40–68), which decreased to 3 SNVs (range: 1–15) with HERRO correction. Illumina sequencing generated a median SNV distance of 2 (range: 0–13). The resulting standalone ONT HERRO-corrected phylogeny was almost indistinguishable from the standalone Illumina-generated phylogeny.

Conclusions: The addition of HERRO correction meant isolates from this MRSA outbreak could be resolved to a level on par with Illumina sequencing. ONT data following HERRO correction represents a viable standalone option for high-resolution genomic analysis of hospital outbreaks, provided sufficient read lengths can be generated.

Introduction

Next Generation Sequencing (NGS) offers huge promise for advancing surveillance of hospital pathogens, but it is still predominantly performed in larger reference or research laboratories [1]. However, this is a rapidly changing field and Oxford Nanopore Technologies (ONT) sequencing offers practical advantages as an accessible sequencing option for front-line diagnostic microbiology laboratories, including minimal footprint, low up-front capital expenses, and rapid sequence data generation [2]. Since 2022, our small-to-medium-sized laboratory lacking NGS bioinformatics expertise, has performed real-time genomic surveillance of hospital organisms using ONT sequencing to support local Infection Prevention and Control (IPC). This has been integrated into our laboratory workflow at relatively low cost and person-time expense [2] and has employed a decentralised analysis model, whereby sequencing and basic bioinformatic analysis are performed in-house and if more granular analysis is required, data are transferred to the reference laboratory for analysis by specialist bioinformaticians. This surveillance program has permitted the early detection of outbreaks, thus limiting the impact on patients and hospital services [3, 4]. Utilising publicly available sequence data, standalone ONT sequencing has been sufficient to demonstrate the clustering of outbreak isolates and permit effective IPC interventions [3, 4]. However, to date, short-read data have been required for single-nucleotide variant (SNV)-level analysis. In a prior analysis, ONT-only data resolved isolates from an outbreak of methicillin-resistant *Staphylococcus aureus* (MRSA) on our Neonatal Intensive Care Unit (NICU) to a median pairwise SNV distance of 65 SNVs, whereas Illumina sequencing produced a median distance of two SNVs due to its higher read accuracy [4]. ONT accuracy continues to improve, with updates to chemistry, basecallers and basecalling models, and more recently the introduction of Haplotype-aware ERRor cORrection (HERRO) [5-7]. HERRO correction integrates read overlapping with artificial intelligence-based error

correction of long reads which can be incorporated into ONT basecalling. Using original raw sequence data from our prior outbreak investigation, this study aimed to assess whether the addition of HERRO correction to our real-time ONT genomic surveillance system would
80 enable SNV-level resolution of isolates.

Methods

For this analysis, we used the originally basecalled data from White *et al.* [4], which utilised Pod5 files basecalled with Dorado v0.3.4 (<https://github.com/nanoporetech/dorado>, accessed
85 05 June 2024). Additionally, we rebasecalled Pod5s using Dorado v0.7.0. The basecalling process utilised the ‘super accuracy’ model (dna_r10.4.1_e8.2_400bps_sup@v5.0.0), with other parameters set at default. Pod5s were also rebasecalled using Dorado v0.7.0 with single-read error correction using the HERRO algorithm (implemented by the ‘dorado correct’ function using model-v1) [5]. HERRO correction recommends read lengths of
90 >10,000 bp. Illumina sequencing data from the original investigation were also included for comparison.

NanoStat v1.6.0 and NanoQC v0.9.4 from NanoPack v1.6.0 [8] were used for initial quality assessment of the raw nanopore reads. NanoFilt v2.8.0, also from NanoPack, was used for
95 read trimming. Initially, 50 nucleotides were trimmed from the start and end of each read to remove low-quality regions. NanoFilt was used again to filter out reads with a quality score below Q7.

Trimmed and filtered reads were *de novo* assembled and polished using the original outbreak
100 investigation protocol for both nanopore and Illumina data [4]. However, the HERRO-corrected reads underwent *de novo* assembly using Flye v2.9.2 [9, 10] using a genome size

estimate of 2.8 Mb and three polishing iterations. The corresponding Illumina reads were compared against each chromosome using Snippy v4.6.0 (<https://github.com/tseemann/snippy>, accessed 05 June 2024) to identify potentially erroneous variants.

We included Illumina sequence data for 13 genomes with recent evolutionary ties to the *S. aureus* sequence type (ST)97 genomes from the mid-2023 NICU outbreak [4]. These included 12 genomes from New Zealand's national staphylococcal surveillance and one publicly available genome (Supplementary Materials, Table S1).

High-resolution analyses of genetic variants were performed using SPANDx v4.0 [11] as described previously [4]. SNV alignments were used to reconstruct phylogenies, and pairwise SNV distances were determined with snp-dist v0.6.3 (<https://github.com/tseemann/snp-dists>, accessed 03 July 2024). Maximum parsimony trees were reconstructed using the PAUP v4.0a heuristic search feature [12]. The resulting phylogenetic trees were visualised using FigTree v1.4.4 (<http://tree.bio.ed.ac.uk/software/figtree/>, accessed 03 July 2024).

Results

Nanopore sequence data quality metrics for the nine ST97 genomes from the original outbreak are presented in Table 1. Four genomes lacked complete chromosome coverage (lower bound of median chromosome depth range includes zero) after HERRO correction due to lower read lengths and were excluded from further analysis. Figure 1 shows maximum parsimony phylogenetic trees comprising the five outbreak genomes and the 13 comparator ST97 genomes. Median pairwise SNV distances were: 2 (range: 0–13) for Illumina, 51 (range: 40–68) for ONT-only with Dorado v0.3.4, 14 (range: 9–21) for Dorado v0.7.0 without

HERRO correction, and 3 (range: 1–15) for Dorado v0.7.0 with HERRO correction.

HERRO-corrected and Illumina alignments were almost identical, except for two additional

SNVs in the HERRO-corrected alignment and one in the Illumina alignment. Mapping

130 Illumina reads to each chromosome revealed minimal SNV differences, mainly in insertion sequences/repeat regions and the Staphyloferrin A transporter (*sfaA*) ([Supplementary Materials, Table S2](#)).

Discussion

135 This analysis shows that HERRO-corrected long-read nanopore data can resolve a NICU-based outbreak of MRSA to the SNV level. The assemblies and phylogenies were almost indistinguishable from those generated by Illumina sequencing. The synonymous mutation c.1045A>G in the Staphyloferrin A transporter gene (*sfaA*) in 23MR1431, 23MR1432, and 23MR1433 might be biased due to the higher chemical similarity within purines (A and G)
140 [13, 14]. HERRO correction yielded shorter phylogenetic tree branches compared to our previous analysis [4]. Here, the median pairwise SNV distance between outbreak isolates was three, well within the proposed cutoff for relatedness of MRSA isolates of 15 SNVs [15].

These results were obtained by reanalysing the same sequence data from our original

145 outbreak analysis, generated by our laboratory's routine day-to-day sequencing activities.

This demonstrates the viability of using ONT sequencing for real-time, high-resolution

outbreak detection in hospitals, making it suitable for front-line clinical microbiology

laboratories. ONT sequencing offers low up-front capital costs, a minimal physical footprint,

and long reads that enable complete *de novo* genome assembly, as well as the detection of

150 mobile genetic elements and plasmids, crucial for IPC and antimicrobial resistance gene detection.

A limitation of this analysis was that only a small number of isolates from a single species were included. To confirm these results, analysis of larger datasets and investigations of the effects of differing depths of genome coverage on HERRO-corrected assemblies are required. Of note, our lowest depth isolate had a median chromosome depth of 35×, with the lowest region being 12× (Table 1). Despite this, an Illumina-equivalent assembly was generated. However, the small number of available isolates was further reduced by excluding four genomes due to insufficient read lengths. This highlights a potential limitation of the implementation of the HERRO correction approach by ONT for their solo long reads, particularly for organisms that typically produce shorter reads, which has been our experience with pathogens such as *Clostridioides difficile* [2, 3].

Recent publications highlight advances in ONT sequencing accuracy through updated chemistry, duplex reads, and ‘methylation-aware’ basecallers, which have significantly improved ONT-only phylogenomic analyses and reduces the need for hybrid genome assembly with short read data [6, 7]. Our analysis further supports ONT sequencing as a viable standalone option for high-resolution surveillance of hospital pathogens, with HERRO correction resulting in highly accurate assemblies.

Author statements

Conflicts of interest: The authors declare that there are no conflicts of financial, general, or institutional competing interests.

Funding: This study was supported by internal departmental funds at Awanui Laboratories Wellington, the Institute of Environmental Science and Research (ESR), New Zealand

Ministry of Health/Manatū Hauora, and by Genomics Aotearoa through funding from the Ministry of Business Innovation and Employment (MBIE).

Acknowledgements: We acknowledge the facilities, and the scientific and technical assistance of staff at Awanui Labs, formerly known as Southern Community Laboratories (SCL) (Wellington, New Zealand). We are grateful to the diagnostic microbiology laboratories across New Zealand for contributing isolates and associated data to the surveillance programmes at the Institute of Environmental Science and Research (ESR). We thank the Next Generation Sequencing Team and the Antimicrobial Resistance Laboratory at ESR (Porirua, New Zealand) for their valuable scientific and technical assistance. We acknowledge and thank the following ESR staff for their valuable feedback: Xiaoyun Ren (Kenepuru Science Centre) and Vinko Besic (Kenepuru Science Centre). This research was made possible by the Computational Science Team and the High-Performance Compute (HPC) platform at ESR, with specific thanks to Russell Smithies and Shane Sturrock for HPC support and helpful discussions about software pipelines. Finally, we thank the New Zealand Ministry of Health/Manatū Hauora for funding this work.

Data availability statement: The study sequences are available in the National Center for Biotechnology Information (NCBI) BioProject accession number PRJNA1046639. The raw sequence read data from Illumina, and Oxford Nanopore Technologies (ONT) basecalled with Dorado v0.3.4, have been deposited to the NCBI Sequence Read Archive (SRA) under accession numbers SRR26992863 to SRR26992891. The re-basecalled (Dorado v0.7.0) ONT raw sequence read data generated in this study are available under accession numbers SRR29788345 to SRR29788353. The complete assembly for strain 23MR1425 has been deposited to GenBank under the accession numbers CP143800 and CP143801.

Contribution: *Writing – Original Draft:* M.Bl. and R.T.W.; *Writing – Review & Editing:* M.Bl., S.B., M.bu., M.L.C., K.D., A.E., S.H., D.MC., W.T., and R.T.W.; *Conceptualization:* M.Bl. and R.T.W.; *Investigation:* M.Bl. and R.T.W.; *Methodology:* S.B., M.bu., M.L.C., K.D., A.E., S.H., and R.T.W.; *Data Curation:* R.T.W.; *Visualization:* R.T.W.; *Formal Analysis:* R.T.W.; *Project Administration:* M.Bl. and R.T.W.; *Funding Acquisition:* M.Bl. and D.MC. All authors have read and agreed to the published version of the manuscript.

References

[1] Jhaveri TA, Weiss ZF, Winkler ML, Pyden AD, Basu SS, Pecora ND. A decade of clinical microbiology: top 10 advances in 10 years: what every infection preventionist and antimicrobial steward should know. *Antimicrob Steward Healthc Epidemiol.* 2024;4:1–11. DOI: <https://doi.org/10.1017/ash.2024.10>

[2] Bloomfield M, Hutton S, Velasco C, Burton M, Benton M, Storey M, et al. Oxford nanopore next generation sequencing in a front-line clinical microbiology laboratory without on-site bioinformaticians. *Pathology.* 2023;56:444–447. DOI: <https://doi.org/10.1016/j.pathol.2023.07.014>

[3] Bloomfield M, Hutton S, Burton M, Tarring C, Velasco C, Clissold C, et al. Early identification of a ward-based outbreak of *Clostridioides difficile* using prospective multilocus sequence type-based Oxford Nanopore genomic surveillance. *Infect Control Hosp Epidemiol.* 2024;6:1–7. DOI: <https://doi.org/10.1017/ice.2024.77>

[4] White RT, Bakker S, Burton M, Castro ML, Couldrey C, Dyet K, et al. Rapid identification and subsequent contextualization of an outbreak of methicillin-resistant *Staphylococcus aureus* in a neonatal intensive care unit using nanopore sequencing. *Microb Genom.* 2024;10. DOI: <https://doi.org/10.1099/mgen.0.001273>

[5] Stanojevic D, Lin D, Florez De Sessions P, Sikic M. Telomere-to-telomere phased genome assembly using error-corrected Simplex nanopore reads. *bioRxiv:* 2024.05.18.594796v1 [Preprint]. 2024. Available from: <https://www.biorxiv.org/content/10.1101/2024.05.18.594796v1> DOI: <https://doi.org/10.1101/2024.05.18.594796>

- [6] Bogaerts B, Van den Bossche A, Verhaegen B, Delbrassinne L, Mattheus W, Nouws S., et al. Closing the gap: Oxford Nanopore Technologies R10 sequencing allows comparable results to Illumina sequencing for SNP-based outbreak investigation of bacterial pathogens. J Clin Microbiol. 2024;62:e0157623. DOI: <https://doi.org/10.1128/jcm.01576-23>
- [7] Sanderson ND, Hopkins KM, Colpus M, Parker M, Lipworth S, Crook D, Stoesser N. Evaluation of the accuracy of bacterial genome reconstruction with Oxford Nanopore R10.4.1 long-read-only sequencing. Microb Genom. 2024;10. DOI: <https://doi.org/10.1099/mgen.0.001246>
- [8] De Coster W, D’hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34:2666–2669. DOI: <https://doi.org/10.1093/bioinformatics/bty149>
- [9] Lin Y, Yuan J, Kolmogorov M, Shen MW, Chaisson M, Pevzner PA. Assembly of long error-prone reads using de Bruijn graphs. Proc Natl Acad Sci USA. 2016;113:E8396–E8405. DOI: <https://doi.org/10.1073/pnas.1604560113>
- [10] Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–546. DOI: <https://doi.org/10.1038/s41587-019-0072-8>
- [11] Sarovich DS, Price EP. SPANDx: a genomics pipeline for comparative analysis of large haploid whole genome re-sequencing datasets. BMC Res Notes. 2014;7:618. DOI: <https://doi.org/10.1186/1756-0500-7-618>
- [12] Wilgenbusch, JC, Swofford D. Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics. 2003;6. DOI: <https://doi.org/10.1002/0471250953.bi0604s00>
- [13] Dohm JC, Peters P, Stralis-Pavese N, Himmelbauer H. Benchmarking of long-read correction methods. NAR Genomics and Bioinformatics. 2020;2. DOI: <https://doi.org/10.1093/nargab/lqaa037>
- [14] Delahaye C, Nicolas J. Sequencing DNA with nanopores: Troubles and biases. PLoS ONE. 2021;16:e0257521. DOI: <https://doi.org/10.1371/journal.pone.0257521>
- [15] Coll F, Raven KE, Knight GM, Blane B, Harrison EM, Leek D, et al. Definition of a genetic relatedness cutoff to exclude recent transmission of meticillin-resistant *Staphylococcus aureus*: a genomic epidemiology analysis. Lancet Microbe. 2020;1:e328–e335. DOI: [https://doi.org/10.1016/S2666-5247\(20\)30149-X](https://doi.org/10.1016/S2666-5247(20)30149-X)

Figure legends

Figure 1. Maximum parsimony phylogeny of *Staphylococcus aureus* sequence type (ST)97. Single-nucleotide variants (SNVs) were derived from a core-genome alignment of ~2,718,000 bp and were called against the chromosome of sample 23MR1425. SNV density filtering in SPANDx (excluded regions with ≥ 3 SNVs in a 10 bp window). All phylogenetic trees were rooted according to the 21MR0293 and SRR8199312 outgroup. Bootstrap values >80% (1,000 replicates) are shown. **(a)** Illumina only data for the neonatal intensive care unit (NICU) isolates. The phylogeny was inferred from 406 core-genome SNVs from 18 genomes. The consistency index for the tree was 1.0. **(b)** Nanopore only data (Dorado v0.3.4) for the NICU isolates (purple taxa). The phylogeny was inferred from 502 core-genome SNVs from 18 genomes. The consistency index for the tree is 0.95. **(c)** Nanopore only data (Dorado v0.7.0, no Haplotype-aware ERRor cORrection (HERRO)) for the NICU isolates (purple taxa). The phylogeny was inferred from 424 core-genome SNVs from 18 genomes. The consistency index for the tree is 0.99. **(d)** Nanopore only data (Dorado v0.7.0 with HERRO correction) for the NICU isolates (purple taxa). The phylogeny was inferred from 407 core-genome SNVs from 18 genomes. The consistency index for the tree is 1.0.

Table 1. Nanopore quality control metrics for the nine *Staphylococcus aureus* genomes

Genome ID	Dorado v0.7.0					HERRO-corrected		
	Number of reads	Total bases	Median Chromosome depth (range)	Mean read length	Read length N50	Mean read quality	Number of reads	Median Chromosome depth (range)
<i>Included in analysis</i>								
23MR1425	104,568	556,031,878	185× (59 to 258)	5,317	10,536	18.2	40,467	153× (75 to 205)
23MR1429	100,616	603,621,103	198× (58 to 283)	5,999	11,439	18.4	43,544	173× (77 to 239)
23MR1431	83,153	602,562,445	194× (61 to 306)	7,246	14,014	18.4	40,617	176× (87 to 264)
23MR1432	21,700	129,463,338	40× (11 to 70)	5,966	12,374	17.7	8,841	35× (12 to 59)
23MR1433	59,252	374,049,295	120× (31 to 184)	6,313	12,711	18.5	25,509	105× (46 to 159)
<i>Excluded from analysis</i>								
23MR1426	204,179	292,225,105	90× (16 to 316)	1,431	2,104	16.8	6,291	11× (0 to 71)
23MR1427	125,786	191,055,409	58× (5 to 230)	1,519	2,330	16.7	5,176	10× (0 to 53)
23MR1428	200,733	370,361,286	116× (20 to 362)	1,845	2,884	17.1	14,839	33× (0 to 112)
23MR1434	50,675	44,938,964	13× (0 to 77)	887	1,117	15.9	106	0× (0 to 6)

Median chromosome depth (range) refers to the median depth at each nucleotide position of the chromosome. The lower bound of the range represents the position(s) with the lowest depth, while the upper bound represents the position(s) with the highest depth.

