University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Steinig, E;Duchêne, S;Aglua, I;Greenhill, A;Ford, R;Yoannes, M;Jaworski, J;Drekore, J;Urakoko, B;Poka, H;Wurr, C;Ebos, E;Nangen, D;Manning, L;Laman, M;Firth, C;Smith, S;Pomat, W;Tong, SYC;Coin, L;McBryde, E;Horwood, P

Title:
Phylodynamic Inference of Bacterial Outbreak Parameters Using Nanopore Sequencing

Date:
2022-03-01

Citation:
Steinig, E., Duchêne, S., Aglua, I., Greenhill, A., Ford, R., Yoannes, M., Jaworski, J., Drekore, J., Urakoko, B., Poka, H., Wurr, C., Ebos, E., Nangen, D., Manning, L., Laman, M., Firth, C., Smith, S., Pomat, W., Tong, S. Y. C. ,... Horwood, P. (2022). Phylodynamic Inference of Bacterial Outbreak Parameters Using Nanopore Sequencing. Molecular Biology and Evolution, 39 (3), https://doi.org/10.1093/molbev/msac040.

Persistent Link:
https://hdl.handle.net/11343/301605

License:
CC BY-NC

Phylodynamic Inference of Bacterial Outbreak Parameters Using Nanopore Sequencing

Eike Steinig,*[,1,2] Sebastián Duchêne,[1] Izzard Aglua,[3] Andrew Greenhill,[4] Rebecca Ford,[4] Mition Yoannes,[4] Jan Jaworski,[3] Jimmy Drekore,[5] Bohu Urakoko,[3] Harry Poka,[3] Clive Wurr,[6] Eri Ebos,[6] David Nangen,[6] Laurens Manning,[7,8] Moses Laman,[4] Cadhla Firth,[2] Simon Smith,[9] William Pomat,[4] Steven Y.C. Tong,[1,10] Lachlan Coin,[†,1] Emma McBryde,[†,2] and Paul Horwood[†,4,11]

[1]Department of Infectious Diseases, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia

[2]Australian Institute of Tropical Health and Medicine, James Cook University, Townsville and Cairns, Australia

[3]Joseph Nombri Memorial-Kundiawa General Hospital, Kundiawa, Papua New Guinea

[4]Papua New Guinea Institute of Medical Research, Goroka, Papua, Papua New Guinea

[5]Simbu Children's Foundation, Kundiawa, Papua New Guinea

[6]Surgical Department, Goroka General Hospital, Goroka, Papua New Guinea

[7]Department of Infectious Diseases, Fiona Stanley Hospital, Murdoch, Australia

[8]Medical School, University of Western Australia, Harry Perkins Research Institute, Fiona Stanley Hospital, Murdoch, Australia

[9]Cairns Hospital and Hinterland Health Service, Queensland Health, Cairns, Australia

[10]Victorian Infectious Diseases Service, The Royal Melbourne Hospital at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia

[11]College of Public Health, Medical & Veterinary Sciences, James Cook University, Townsville, Australia

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: eike.steinig@unimelb.edu.au.

Associate editor: Thomas Leitner

Abstract

Nanopore sequencing and phylodynamic modeling have been used to reconstruct the transmission dynamics of viral epidemics, but their application to bacterial pathogens has remained challenging. Cost-effective bacterial genome sequencing and variant calling on nanopore platforms would greatly enhance surveillance and outbreak response in communities without access to sequencing infrastructure. Here, we adapt random forest models for single nucleotide polymorphism (SNP) polishing developed by Sanderson and colleagues (2020. High precision *Neisseria gonorrhoeae* variant and antimicrobial resistance calling from metagenomic nanopore sequencing. *Genome Res*. 30(9):1354–1363) to estimate divergence and effective reproduction numbers (R_e) of two methicillin-resistant *Staphylococcus aureus* (MRSA) outbreaks from remote communities in Far North Queensland and Papua New Guinea (PNG; $n = 159$). Successive barcoded panels of *S. aureus* isolates (2×12 per MinION) sequenced at low coverage (>5× to 10×) provided sufficient data to accurately infer genotypes with high recall when compared with Illumina references. Random forest models achieved high resolution on ST93 outbreak sequence types (>90% accuracy and precision) and enabled phylodynamic inference of epidemiological parameters using birth–death skyline models. Our method reproduced phylogenetic topology, origin of the outbreaks, and indications of epidemic growth ($R_e > 1$). Nextflow pipelines implement SNP polisher training, evaluation, and outbreak alignments, enabling reconstruction of within-lineage transmission dynamics for infection control of bacterial disease outbreaks on portable nanopore platforms. Our study shows that nanopore technology can be used for bacterial outbreak reconstruction at competitive costs, providing opportunities for infection control in hospitals and communities without access to sequencing infrastructure, such as in remote northern Australia and PNG.

Key words: **nanopore, phylodynamics, bacteria, outbreaks, reproduction number, BEAST.**

Introduction

Sequence data from infectious disease outbreaks have provided critical information for infection control and inference of pathogen transmission dynamics, including during the West African Ebola virus epidemic (Quick et al. 2016) and the current severe acute respiratory syndrome coronavirus 2

© The Author(s) 2022. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

Downloaded from https://academic.oup.com/mbe/article/39/3/msac040/6529399 by University of Melbourne Library user on 14 April 2022

(SARS-CoV-2) pandemic (Bull et al. 2020). Maximum-likelihood (ML) and Bayesian phylodynamic methods are commonly used to date the emergence of lineages, variants, and outbreaks, and to estimate key epidemiological parameters, such as changes in the effective reproduction number over time (R_e; Hodcroft et al. 2021; Volz et al. 2021). Oxford Nanopore Technology (ONT) sequencing has emerged as viable technology for real-time genomic epidemiology and has been applied across national-scale SARS-CoV-2 surveillance efforts in the United Kingdom, Denmark, and Australia, amongst other countries (da Silva Filipe et al. 2021; du Plessis et al. 2021; Hammer et al. 2021; Nicholls et al. 2021). Moreover, nanopore sequencing devices can, in theory, be operated in low- and middle-income countries (LMICs) where local genomics infrastructure may be lacking or is difficult to access (Faria et al. 2017; Giovanetti et al. 2020), so that a timely outbreak response is not feasible (Gardy and Loman 2018). Accessible genomics infrastructure is particularly relevant for continuous surveillance at bacterial evolutionary time scales, where outbreak strains may circulate for years, and can persist in human and animal reservoirs or the environment outside their host. Furthermore, viral pathogen genomes, such as Ebola virus or SARS-CoV-2, are often sequenced directly from patient samples using targeted PCR-based enrichment approaches, achieving high-genome coverage and resolution capable of informing phylodynamic models (Quick et al. 2017; Bull et al. 2020). However, nanopore sequencing for bacterial pathogens, coupled to Bayesian phylodynamic models, has so far not been considered for routine epidemiological applications, mainly due the need for sufficiently accurate single nucleotide polymorphism (SNP) calling at bacterial whole-genome scales (Ingle et al. 2021). SNP calls from high-coverage ($>30\times$) Illumina data are the current standard for accurate SNP calls used in phylogenetic applications, but current generation nanopore SNP calling has suffered from low raw sequence read accuracy (R9.4.1, $<$ Guppy v5.0) and a focus on variant calling in human genomes, with much of the available callers developed specifically for human variants (Luo et al. 2020). This problem is further aggravated when attempting to sequence cost-effectively, for example, using low-coverage multiplexed runs ($<5–10\times$) and simple library preparation with ONT sequencing kits (at least R9.4.1 pore architecture, SQK-RBK004 libraries) that can be used in LMICs with large burdens of bacterial disease.

Phylodynamic inference on nanopore platforms is further complicated because (ideally) an outbreak reference genome is used, that is closely related to the outbreak sequence type, thus providing sufficiently high variant calling resolution for transmission inference, particularly in recent transmission chains or outbreaks (Gorrie et al. 2021). In addition, on bacterial time scales (years) little sequence variation will have occurred in newly emergent outbreaks, which places a disproportionate emphasis on correctly inferring the few available outbreak-specific SNPs. As a consequence, there is little room for systematic errors introduced by base and variant callers when using (low coverage) nanopore sequencing data to effectively survey bacterial outbreaks. Neural network-based, nanopore-native variant callers in particular can introduce excessive false-positive (FP) SNP calls, complicating transmission inference from ONT sequence data, where accuracy and precision are required (Sanderson et al. 2020). Within-lineage phylodynamic inference for bacterial outbreaks additionally depends on sufficient temporal signal to ascertain a phylodynamic threshold, at which sufficient molecular evolutionary change has accumulated in the sample to obtain robust phylodynamic estimates (Duchêne, Geoghegan et al. 2016; Duchene, Featherstone, et al. 2020; Duchene, Lemey, et al. 2020). Due to slower substitution rates in bacteria compared with viruses (Duchêne, Holt et al. 2016), longitudinal sample collections are optimal for genomic epidemiology and often require multiple years of data to infer transmission dynamics of the sampled population. Requirements for accurate whole-genome SNP calls across a large number of isolates, sequenced cost-effectively at low-genome coverage and over a sufficient interval of time, represent a significant barrier to the implementation of phylodynamic modeling for bacterial pathogens.

Illumina hybrid-corrected and ONT-native phylogenetic analyses methods have been demonstrated for a small number of distantly related bacterial nanopore genomes and genome assemblies from the same species for example, *Neisseria gonorrhoeae* (Golparian et al. 2018; Sanderson et al. 2020) or between species from environmental sources (Urban et al. 2021). Recently, a six-strain multiplex protocol for the MinION with genome assembly and determination of phylogenetic relationships to identify outbreaks has been tested for *Staphylococcus aureus* lineages sampled in Norway. However, it remains unclear whether full within-lineage phylodynamic modeling is possible at population-level scale, whether estimates from nanopore data match results obtained using SNP calling with Illumina reads and whether sequencing runs can be conducted cost-effectively (at least 24 isolates per run). In this study, we adapt a variant polishing approach first implemented by Sanderson et al. (2020) on metagenomic sequencing of *N. gonorrhoeae* using random forest classifiers to filter SNP calls from the nanopore-native variant callers *Medaka* v1.2.3 and *Clair* v2.1.1 (Luo et al. 2020). We use *Snippy* Illumina variant profiles as reference data and investigate caller performance across reference genomes and outbreak data sets. We show that random forest classifiers sufficiently remove incorrect calls from *Clair* in outbreak isolates with $>5\times$ coverage to allow for sequencing of 24 community-associated *S. aureus* isolates per MinION flow cell ($n = 181$) which successfully resolved phylodynamic parameters estimates of two outbreaks of ST93-MRSA-IV in remote Far North Queensland (FNQ) and Papua New Guinea (PNG).

New Approaches

In this study, we adapt the use of random forest classifiers for SNP polishing of nanopore sequence data to reduce excessive FP SNP calls, which have so far prevented accurate phylogenetic reconstruction of bacterial outbreaks. Our approach enables the inference of outbreak source, timing, and effective reproduction numbers using Bayesian phylodynamic models.

Downloaded from https://academic.oup.com/mbe/article/39/3/msac040/6529399 by University of Melbourne Library user on 14 April 2022

We validated the method on new sequence data from two outbreaks of community-associated *S. aureus* in remote communities of northern Australia and PNG.

Results

We sequenced a total of 181 unique isolates from a pediatric osteomyelitis outbreak (collected between 2012 and 2018) in the Papua New Guinean highland towns Kundiawa (Simbu Province, $n = 42$) and Goroka (Eastern Highlands Province, $n = 45$). We additionally sequenced haphazardly collected blood cultures from a hospital in Madang (Madang Province, $n = 8$) and strains from routine community surveillance across FNQ collected in 2019 (Cairns and Hinterlands, Cape York Peninsula, Torres Strait Islands, processed at Cairns Hospital, $n = 86$; fig. 1, supplementary tables, Supplementary Material online). ONT sequencing was conducted using a minimal, dual-panel barcoding scheme, multiplexing 2 × 12 isolates interspersed with a nuclease flush on a single MinION flow cell (R9.4.1, EXP-WSH-003) for a total of 96 barcodes per outbreak (including isolate re-runs that were merged, $n = 12$, and external isolates excluded here, $n = 3$). Rapid barcode sequencing libraries (RBK-004) were prepared omitting magnetic bead clean-ups after enzymatic digestion of cultured strains and simple spin column extraction. Panels produced between 0.506 and 6.47 Gb of sequence data per

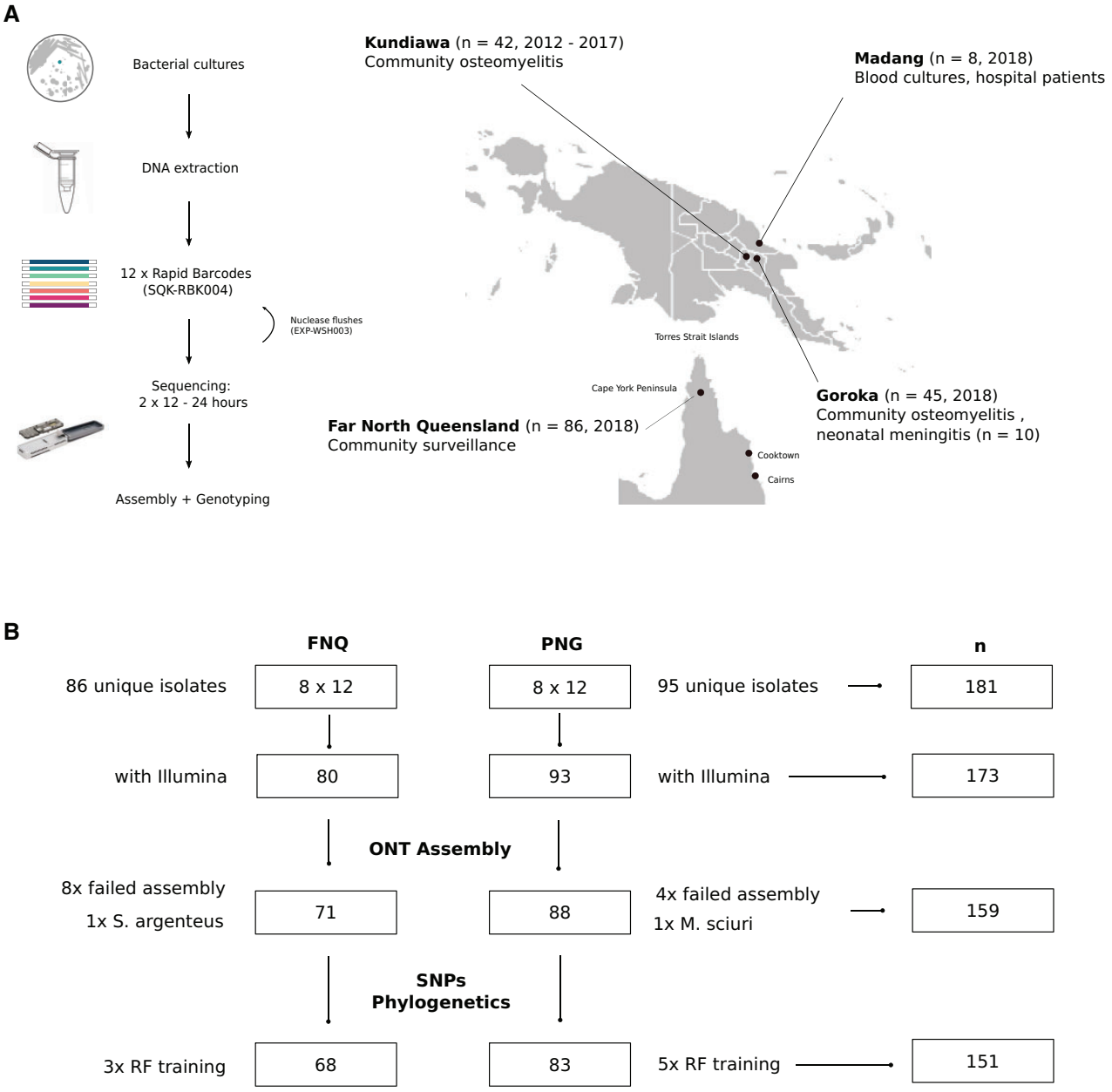


FIG. 1. Culture-based sequencing protocol and outbreak sampling locations in northern Australia and PNG. (*A*) Isolates were sequenced on 8 flow cells with 24 isolates per flow cell using a sequential nuclease flush protocol. (*B*) Sequenced data were subset to those matching Illumina sequencing of the isolates, assembled, and quality controlled. Several isolates were set aside for independent random forest classifier training used in the SNP polishing and phylogenetics pipeline.

Downloaded from https://academic.oup.com/mbe/article/39/3/msac040/6529399 by University of Melbourne Library user on 14 April 2022

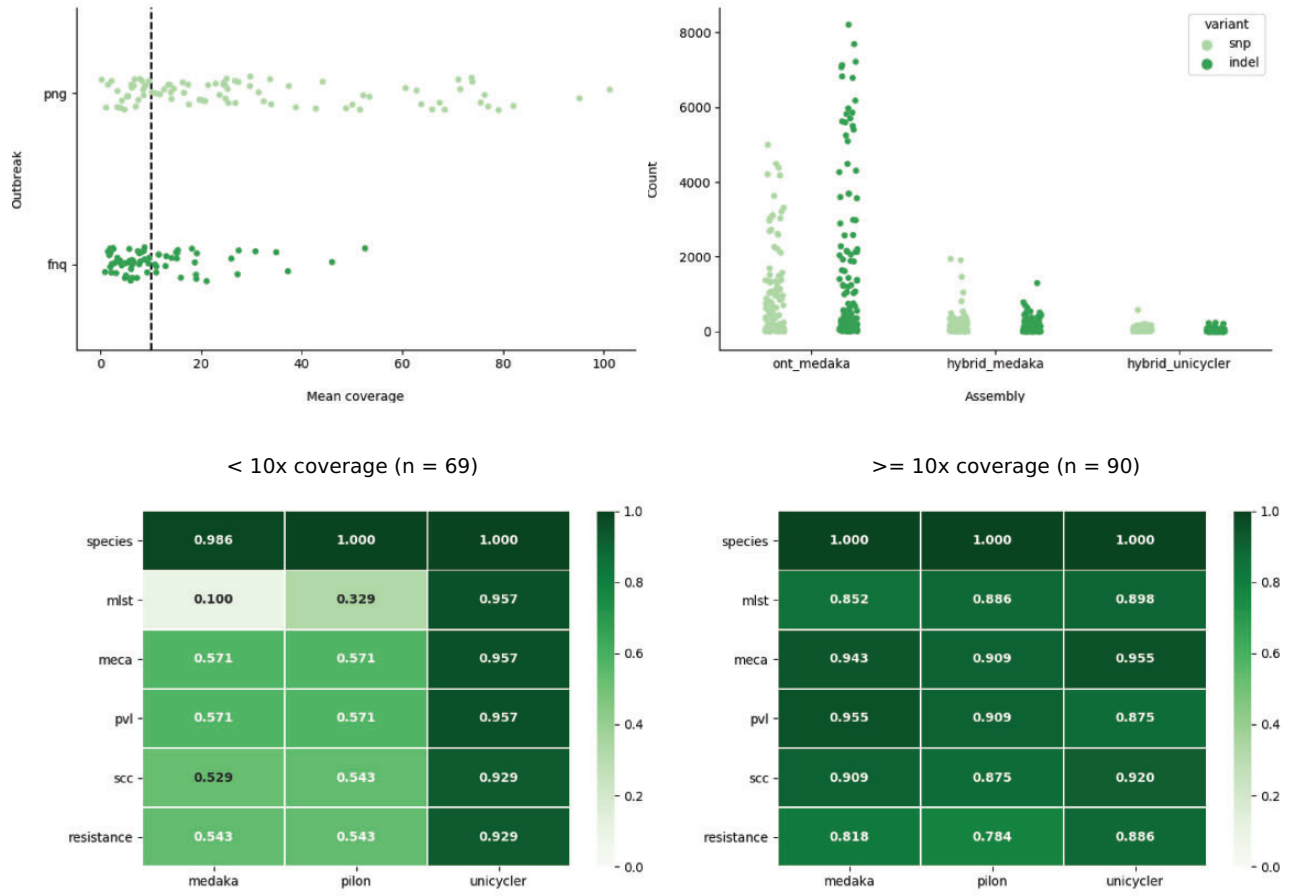


FIG. 2. (*A*) Average genome coverage (R9.4.1, RBK-004) of *Bonito* base-called nanopore reads against the JKD6159 (ST93) reference genome ($n = 159$) where the dashed lines indicate the coverage thresholds chosen to evaluate genotyping (10×) and phylodynamic models (5×) in the FNQ and PNG outbreaks. SNP and indel counts across three different assembly types: uncorrected nanopore reads polished with *Medaka* (ont_medaka), *Medaka* polished nanopore genomes Illumina corrected with *Pilon* (hybrid_medaka), and hybrid assembly in *Unicycler* (hybrid_-unicycler). (*B*) Assembly genotyping results are shown as proportion of assemblies matching the reference Illumina genotype across the three types of assemblies, and the 10× coverage threshold.

run (<24 h) resulting in low–medium coverage per isolate (ST93-JKD6159; fig. 2*A*). We excluded one infection with *S. argenteus* (FNQ) and one coinfection with *Mammaliococcus sciuri* (PNG). Isolates with matching Illumina data were retained to create a high-quality reference data set for further evaluation of genome assembly and variant calling ($n = 159$, fig. 1).

Genome Assembly and Genotype Validation

Short-read reference genomes (fragmented but highly accurate), long-read polished nanopore genomes (contiguous assemblies but less accurate), and long-read hybrid genomes corrected with *Pilon* (Walker et al. 2014) or *Unicycler* (Wick et al. 2017) (contiguous and highly accurate) were assembled using a standardized Nextflow (Di Tommaso et al. 2017) pipeline wrapping *Shovill*, *Flye* (Kolmogorov et al. 2019), *Medaka* and other components (Materials and Methods). Several isolates (12/159) failed long-read assembly due to excessive fragmentation of libraries and/or barcode attachment, but did not fail the short-read assemblies with *Skesa* (Souvorov et al. 2018) or the hybrid assemblies with *Unicycler* (supplementary tables, Supplementary Material online), which first assembles short reads and then scaffolds the assemblies with long reads to generate contiguous whole-genome assemblies.

Compared with Illumina reference assemblies, SNP and indels were frequently occurring in low-coverage uncorrected nanopore assemblies (fig. 2*A*, right). Errors were considerably reduced in high-coverage isolates leading to assembly identities ranging between 0.9993 and 0.9999 in the *dnadiff* metric (Marçais et al. 2018; supplementary tables, Supplementary Material online). Recovery of complete chromosomes and *S. aureus* specific genotypes from uncorrected long-read assemblies was sufficient for high-coverage isolates in our collection (fig. 2*B*, > 80–90%). Assembly genotyping for clinically relevant features such as the presence of *mecA* or the Panton-Valentine leukocidin (PVL), major subtypes of SCC*mec* elements, resistance genes, and other markers of interest showed high concordance with reference assemblies (fig. 2*B*). In contrast, low-coverage assemblies often failed to call genotypes—recovery was low for *mecA* and SCC*mec* types, as well as for PVL and other markers of interest (fig. 2*B*, <60%, supplementary tables, Supplementary Material online). Hybrid long-read correction with *Pilon* did

Downloaded from https://academic.oup.com/mbe/article/39/3/msac040/6529399 by University of Melbourne Library user on 14 April 2022

not markedly improve genotype recovery in low-coverage isolate; however, recovery improved in the *Unicycler* hybrid assemblies (fig. 2*A* and *B*). Lower SCC*mec* subtyping performance was likely due to remaining insertions or deletions from nanopore data impacting on the large cassette chromosomes (>20 kb). *Unicycler* produced more accurate hybrid assemblies than correction of long-read assemblies with *Pilon* alone and performed slightly better in hybrid assemblies of low-coverage nanopore data (fig. 2*B*). For genome assembly and genotyping, our dual-panel sequencing approach recovers nanopore genotypes in high-coverage isolates (>10×) although some errors remain, particularly in sequence type calling and SCC*mec* subtyping.

Training and Evaluation of Random Forest SNP Polishers

Next, we aimed to accurately reconstruct the PNG and FNQ outbreaks within the ML background phylogeny of ST93. Subsequent phylodynamic analysis is challenging because accurate reconstruction of branch lengths within the nanopore clades is required for reproduction of the Bayesian epidemiological parameters. We first tried a candidate-driven approach, using Illumina core SNP panels from the ST93 background population (*Snippy*, $n = 444$, 6,616 SNPs) and *Megalodon* which accurately reconstructed the divergence of the PNG clusters from the Australian East Coast (supplementary fig. S1, Supplementary Material online). However, within-outbreak branch lengths were not reconstructed, because novel variation had accumulated since the divergence from the Australian East Coast population in the 1990s. We therefore decided to use a de novo variant calling approach comparing two native nanopore variant callers based on neural network architectures, by default trained on *Homo sapiens* variant calls (*Clair* v2.1.1) or a mix of human and microbial data from *Escherichia coli*, *Saccharomyces cerevisiae*, and *H. sapiens* (v1.2.3). Although recall was high, raw basecaller performance was exceedingly low in both *Clair* and *Medaka* accuracy and precision, particularly in outbreak isolate calls against the outbreak reference genome (<20%, supplementary fig. S2, Supplementary Material online).

We next adopted SNP polishers using random forest classifiers originally developed by Sanderson and colleagues (2020) to correct nanopore variants in *N. gonorrhoeae* from metagenomic data (fig. 3, Materials and Methods). Each

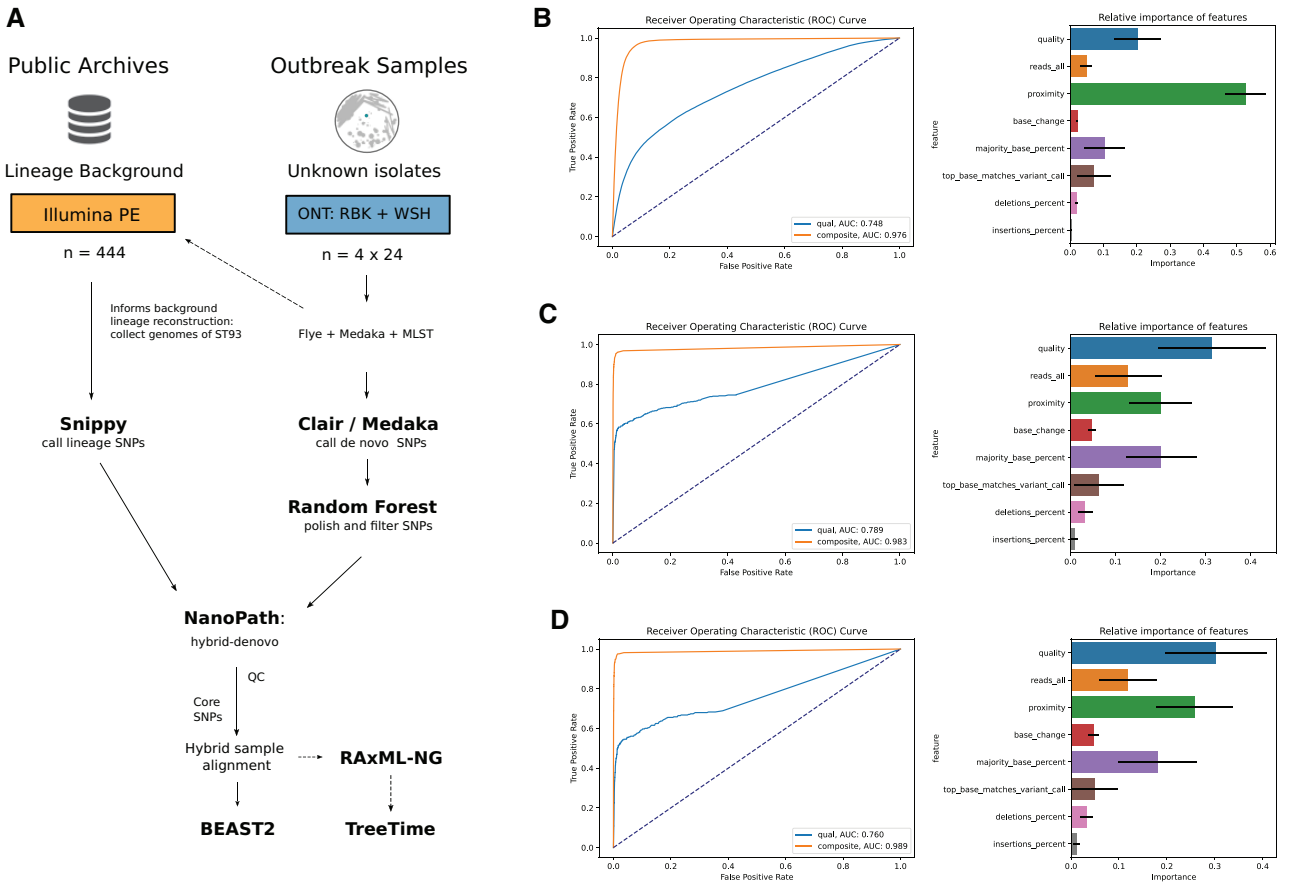


FIG. 3. (*A*) Workflow outlining computational analysis of community-associated *S. aureus* nanopore sequencing using successive barcode panels on ONT MinION flow cells (R9.4.1). MLST typing informs the background population genome collection from a previous study (Illumina). Outbreaks in PNG and FNQ were caused by the Australian clone (ST93-MRSA-IV). SNPs are called for the Illumina background with *Snippy* and ONT outbreak isolates with *Clair*. ONT SNP calls are polished using random forest SNP classifiers, trained on the outbreak reference genome (JKD6159 of ST93). (*B*–*D*) AUC scores of quality or composite features (left) used in training random forest classifiers for SNP polishing and relative feature importance of models (right) trained on (*B*) *S. aureus* mixed lineages (ST88, ST15, and ST93) (*C*) ST93 FNQ isolates and (*D*) ST93 from PNG with matching Illumina data and Snippy reference calls (all $n = 3$).

Downloaded from https://academic.oup.com/mbe/article/39/3/msac040/6529399 by University of Melbourne Library user on 14 April 2022

classifier was trained on three isolates with matching Illumina data and composite sequence features (fig. 3B–D); because there were no considerations of specific training sets used in the original *N. gonorrhoeae* classifier, we trained *S. aureus* classifiers on three combinations of isolates including a mixed set of three sequence types (ST93, ST88, ST15; *saureus_mixed*) and two sets of outbreak sequence type isolates (ST93) from either FNQ (*saureus_fnq*) or PNG (*saureus_png*). In combination with the original *N. gonorrhoeae* classifiers, the different training sets allowed us to evaluate whether SNP polishing was effective using models from a different species entirely (Sanderson), from the same species but without outbreak-related data (saureus_mixed) or from the same species, but with isolates from the same sequence type or outbreak (*saureus_fnq*, *saureus_png*). All models trained on composite sequence features (fig. 3, Materials and Methods) demonstrated high area under the curve (AUC) scores (0.976–0.989, orange) whereas models trained on quality features alone showed suboptimal AUC performance (0.748–0.760, blue) (fig. 3B–D) demonstrating that quality scores alone would be insufficient to discriminate false calls and remove them.

We next evaluated both the *N. gonorrhoeae* classifier, as well as the three *S. aureus* models against the remaining isolates from PNG and FNQ, excluding those used in training (fig. 1B). Evaluations indicated that all trained SNP polishers increased accuracy and precision with slight reductions in recall (fig. 4). However, suboptimal performance was observed in all metrics for the *N. gonorrhoeae* classifier across outbreak sequence types (<40%) as well as other sequence types (<50%). Performance improved considerably in the mixed *S. aureus* polisher (*saureus_mixed*) both among outbreak isolates (69.52% $\pm$ 12.48 accuracy, 75.94% $\pm$ 14.56 precision) and other sequence types (81.94% $\pm$ 14.56 accuracy, 90.11% $\pm$ 6.83 precision). However, despite significant baseline improvement, the interspecies and mixed-sequence type models the number of FP SNP calls remained in the range of 100s to 1,000s (right column, fig. 4A and B). Training the models with isolates from the same sequence type (ST93, FNQ) slightly improved performance (ST93: 71.69% $\pm$ 13.99 accuracy, 83.33% $\pm$ 10.42 precision) but reductions of accuracy and recall in other sequence types were observed (fig. 4C). PNG outbreak-derived model (saureus_png) performed best for polishing isolates from the same outbreak across all metrics in the high-coverage isolates (ST93: 69.28% $\pm$ 16.78 accuracy, 87.57% $\pm$ 9.83 precision) but incurred a steeper cost to accuracy and recall in nonoutbreak isolates (fig. 4D). Reductions indicate that the model trained on features specific to the outbreak genotype and became significantly less generalizable to other sequence type applications. We note that the levels of precision and accuracy of the ST93 polishers in absolute numbers translate to 1–10 s of false SNP calls compared with the *N. gonorrhoeae* and mixed-sequence type model (fig. 4).

Phylogenetic Trees and Transmission Dynamics from Polished SNPs

We next implemented Snippy's core alignment functionality, calling sites present in all isolates of the sampled population, with a minimum SNP site coverage of 1× (JKD6159). Hybrid alignments integrated Illumina background SNPs from the ST93 (outbreak) lineage ($n = 444$) in combination with polished ONT nanopore calls from Clair (fig. 1). The lineage background alignment, as one would use for short-read reference data, therefore served as a backbone for ONT data in the core-site alignment (fig. 4B). We retained isolates with at least 5× coverage ($n = 531/562$) due to low accuracy and precision of these isolates in the SNP polishing step (fig. 4, supplementary fig. S4, Supplementary Material online). We then used the between-species, within-species, within-lineage (FNQ and PNG) models to apply for variant polishing in our de novo core alignment and phylodynamics pipeline (fig. 3A).

NanoPath's core alignment construction reproduced *Snippy*'s core alignment from Illumina data closely (6,319 SNPs with *NanoPath* vs. 6,580 SNPs with *Snippy-core*, fig. 5A and B). When we called *Clair* SNPs on isolates with >5× coverage from PNG ($n = 56$) and FNQ ($n = 32$), we observed a vast excess of SNP calls, particularly in the raw *Clair* calls, where the hybrid core alignment contained 491,210 SNP sites and was considered unusable (supplementary table 7, Supplementary Material online). All polished SNPs produced reasonable alignments, where FNQ and PNG polishers produced alignments closest to the Illumina reference (fig. 5, supplementary table 2, Supplementary Material online). We reconstructed the ML phylogenies from these alignments in *RAxML-NG* using the *GTR + G* model with Lewis' ascertainment bias correction and rooted the trees on SRR115752 for comparison of topological consistency (van Hal et al. 2018). We also wanted to investigate whether the main introductions into FNQ and PNG could be reconstructed with accurate interpretations of their source divergence on the Australian East Coast. For reference, we used Illumina alignments constructed with *NanoPath* (Materials and Methods) and *Snippy-core* with matching isolates ($n = 531$, fig. 5).

All major clades and subpopulations of the background population (North West, East Coast, Northern Territory, and New Zealand) including the outbreaks in FNQ and PNG were accurately reconstructed as referenced by the Illumina trees (fig. 5). Minor topological variations were observed in the position of the PNG-1 and PNG-2 introductions (greens), and the southern East Coast and New Zealand subclade (seagreen) of the East Coast population (turquoise, fig. 5). However, there were no major topological inconsistencies that would affect interpretation of the source population. In all topologies, the outbreaks from PNG derived from the East Coast ST93-MRSA-IV clade, and the FNQ outbreak derived from the Northern Territory reintroduction (fig. 5). Regional transmissions into the United Kingdom and Australia within the outbreak clusters remained identifiable (black and red branches in PNG-1 and PNG-2). Introductions into FNQ from other parts of the population are evident from both the reference and the polished alignments (red branches in East Coast, PNG and Northern Territory clades). Branch lengths of the nanopore-sequenced clades were similar to the reference ML tree, but were excessive in the between-species *N. gonorrhoeae* polished alignments as well as in the mixed

Downloaded from https://academic.oup.com/mbe/article/39/3/msac040/6529399 by University of Melbourne Library user on 14 April 2022

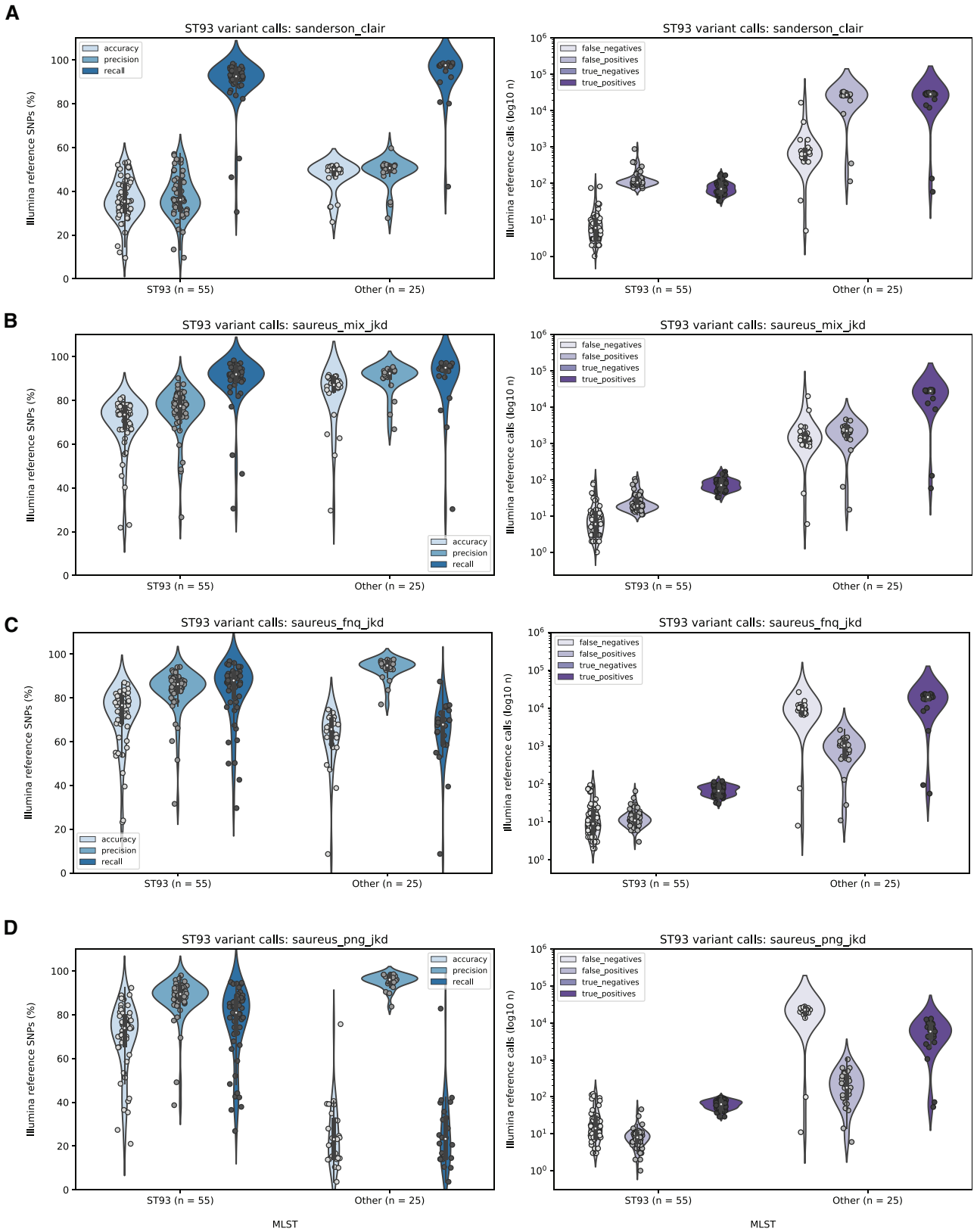


Fig. 4. Trained random forest SNP polisher evaluation showing left: accuracy, precision, and recall of *Clair* nanopore SNP calls against matching Illumina reference SNPs called with Snippy. Plots are split into ST93 outbreak isolates (inside left) and other sequence types (inside right) from PNG and FNQ combined. In the right-hand plots, the number of FNs, FPs, and TP SNP calls for the groups is shown on a log-scale. Models were trained on three Illumina matched isolates from between-species (*A*) *N. gonorrhea* from Sanderson et al. within species (*B*) *S. aureus* ST88, ST93, ST15 from PNG, (*C*) within-lineage (ST93) using samples from FNQ and separately from PNG (*D*) (ST93). Polishing models were evaluated on all PNG and FNQ isolates excluding those used in training (ST93: $n = 55$, other sequence types: $n = 25$, $>10\times$ coverage). Outliers in the tails of the distributions are novel multilocus sequence type variants of ST93.

Downloaded from https://academic.oup.com/mbe/article/39/3/msac040/6529399 by University of Melbourne Library user on 14 April 2022

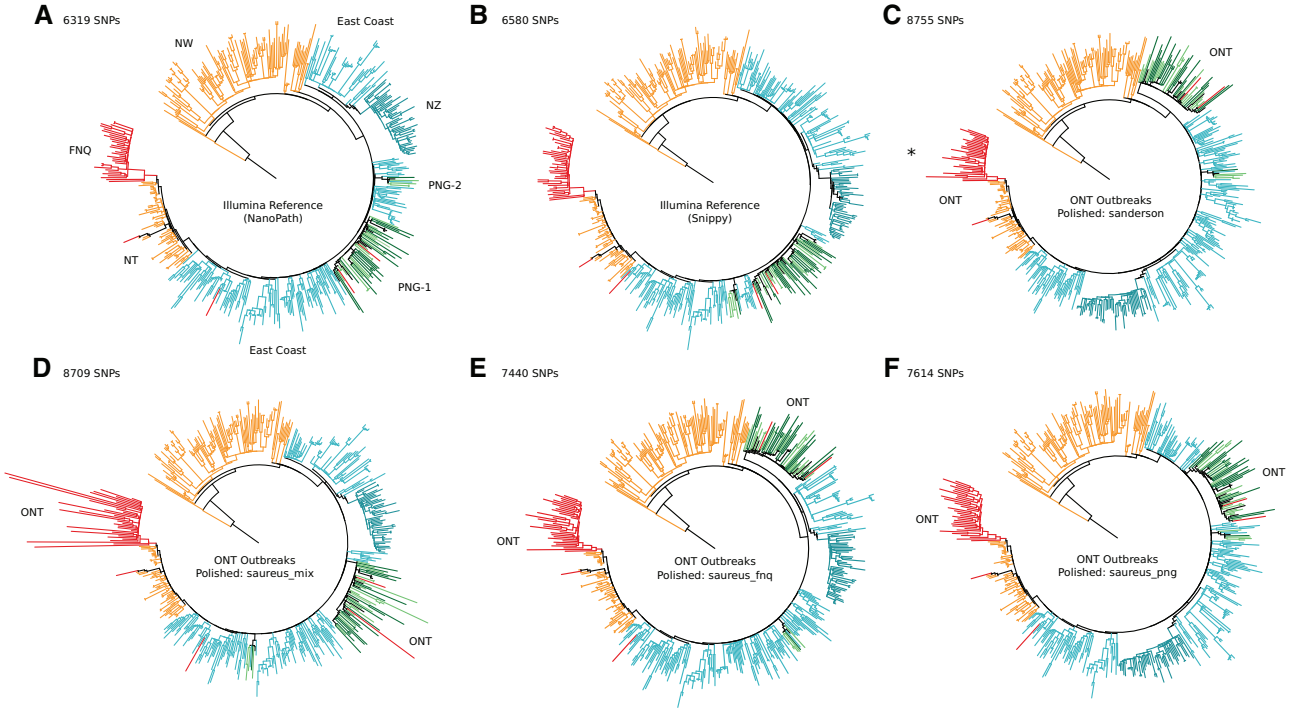


FIG. 5. Comparison of ML phylogenetic topologies of ST93. Illumina reference trees were constructed with *NanoPath* (*A*) and *Snippy* (*B*). All other trees are hybrid phylogenies including the nanopore data of the outbreaks in FNQ and PNG (>5× coverage) within the ST93 background population (Illumina, $n = 531$) (*B*) after polishing *Clair* SNPs using the trained random forest classifiers (*C*: *N. gonorrhoeae*; *D and E*: *S. aureus* mixed and lineage-specific). Asterisk (*) denotes two isolates with excessive branch lengths that were removed for visual clarity (supplementary fig. S6, Supplementary Material online).

sequence type alignments (fig. 5, in particular due to two isolates: PNG-36 and PNG-62, supplementary fig. S6, Supplementary Material online). The alignment based on SNPs polished using outbreak sequence type (ST93) isolates was most consistent with the Illumina reference phylogeny of ST93. We note that within-lineage polishing did not require within-outbreak polishers, for example, FNQ-trained polishers reproduced PNG outbreak divergence and vice versa.

We next investigated the performance of Bayesian phylodynamic methods to estimate the divergence date and effective reproduction number using birth–death skyline models with serial (PNG) or contemporaneous (FNQ) sampling and lineage-specific prior configurations. We ran *BEAST2* Markov chain Monte Carlo (MCMC) chains on the outbreak subsets of the full SNP alignment with sufficient isolates ($n_{\text{PNG-1}} = 53$; $n_{\text{FNQ}} = 32$) using a fixed substitution rate of the whole-lineage median posterior estimate (3.199×10^{-4}). This was necessary as nonrandom sampling (subsetting the alignment to the outbreak clade) removes the temporal signal in the comparatively recent outbreaks, and thus leads to an overestimation of the outbreak MRCA. We note that the models were efficiently run on a standard NVIDIA GTX1080-Ti graphical processing unit (GPU) using *BEAST2* with the *BEAGLE* library at speeds of <3–4 min/million steps in the MCMC (5–7 h per run and GPU) making timely parameter estimation for outbreak responses feasible on low-cost hardware. On an NVIDIA P100 GPU, walltime decreased to <50 s to 1 min/million steps in the MCMC, around 1–2 h walltime per run and GPU on a distributed system.

MCMC chains converged onto similar posterior distributions across all polished alignments in the PNG clade (fig. 6). Polished models in the PNG clade were highly stable across posterior estimates, including those polished with between-species polisher from *N. gonorrhoeae*, and showing only slightly aberrant estimates of the most recent common ancestor in the mixed polishing model (fig. 6*B*, supplementary table S2, Supplementary Material online). More variable posterior estimates were observed in the FNQ clade (fig. 6), consistent with higher variability in branch lengths as a result of excessive FP SNP calls retained in low-coverage FNQ isolates (fig. 5). Nevertheless, when compared with the *NanoPath* Illumina reference estimates, ST93-polished estimates (saureus_png, saureus_fnq) closely resembled those of the reference, with only minor deviations (fig. 6, supplementary table S2, Supplementary Material online). Estimates were consistent with full lineage-wide analysis ($R_e > 1.5$–2.0) and we observed robust estimates in an exploration of the R_e prior (supplementary table S2 and figs. S6 and S7, Supplementary Material online). We therefore demonstrate that SNP polishing enables the use of birth–death skyline models for outbreak parameter estimation, even with low-coverage nanopore sequencing data (5× to 10×). Finally, we implemented training, evaluation, and deployment of SNP polishers for within-lineage transmission modeling in Nextflow.

Discussion

In this study, we adopted a method for variant polishing for phylodynamic modeling of bacterial whole-genome data

Downloaded from https://academic.oup.com/mbe/article/39/3/msac040/6529399 by University of Melbourne Library user on 14 April 2022

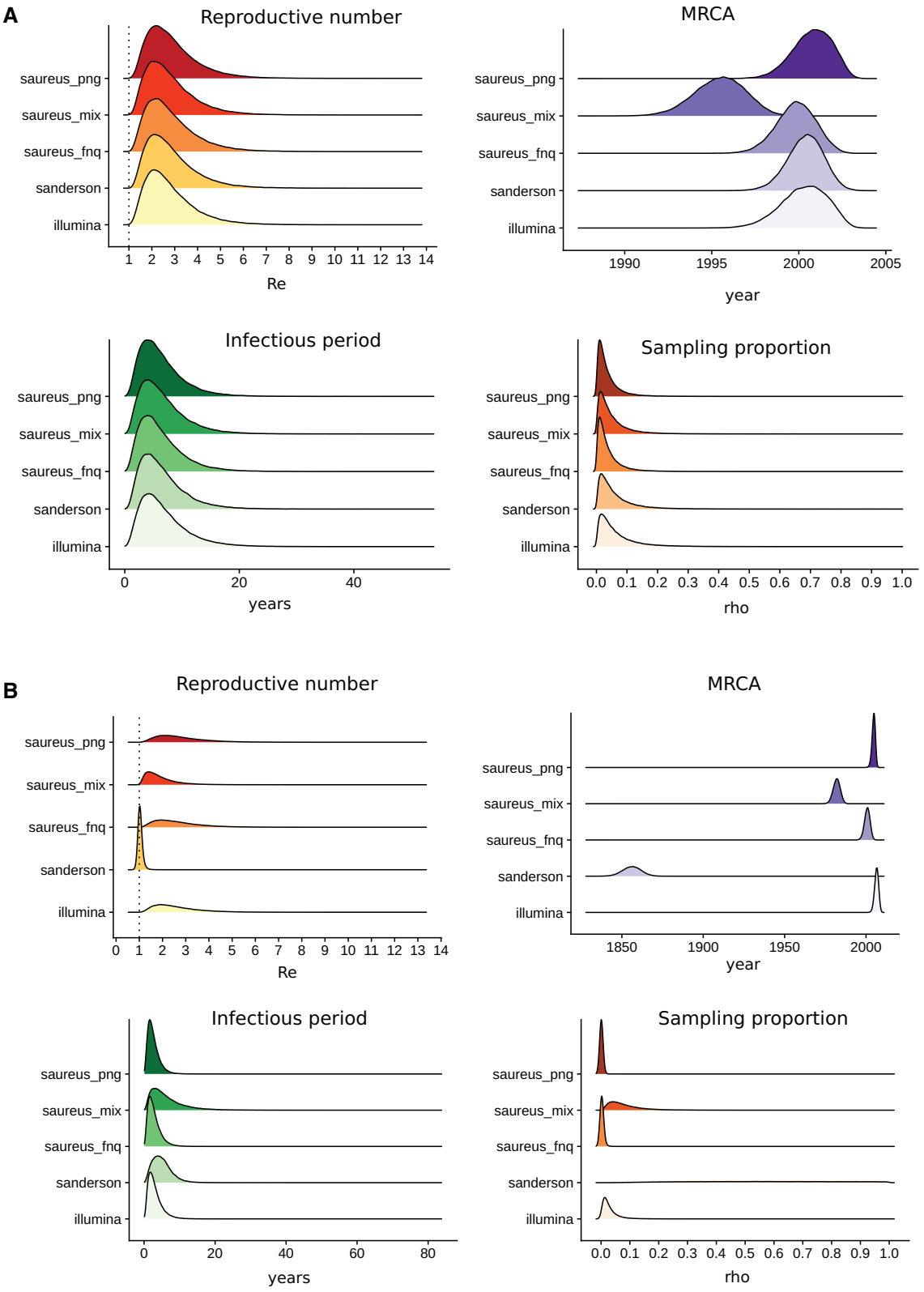


FIG. 6. Posterior distributions of the effective reproduction number (R_e), most recent common ancestor of the outbreak (MRCA), infectious period ($1 \div \delta$) and sampling proportion (ρ) for the nanopore-sequenced outbreak clades in PNG (*A*, $n = 56$) and FNQ (*B*, $n = 32$). Birth–death skyline models were run on the clade subsets of the polished hybrid alignments with $>5\times$ coverage (ridge labels) including the Illumina reference alignment (illumina, bottom ridge), the between-species *N. gonorrhoeae* polished alignment by Sanderson and colleagues (Sanderson), as well as the *S. aureus* mixed lineages (saureus_mix), ST93 FNQ (saureus_fnq), and ST93 PNG (saureus_png).

Downloaded from https://academic.oup.com/mbe/article/39/3/msac040/6529399 by University of Melbourne Library user on 14 April 2022

using low-coverage nanopore sequencing, and applied it to outbreaks of community-associated *S. aureus* in remote FNQ and PNG. Previous studies using (high coverage) nanopore data have evaluated phylogenetic reconstructions on few and distantly related isolates of *N. gonorrhoeae* as well as other bacterial genomes from assembly (Golparian et al. 2018; Sanderson et al. 2020; Urban et al. 2021). A recent pipeline for cluster identification using six strains per MinION flow cell (42 on 7 flow cells) successfully identified clusters in four distinct lineages, using a whole-genome assembly-based phylogeny (Ferreira et al. 2021). However, full outbreak reconstruction within the outbreak lineage—allowing for Bayesian model applications to estimate epidemiological parameters within the phylogeny—has so far not been conducted. Here, we show that the application of random forest SNP polishers developed by Sanderson and colleagues (2020) can sufficiently reduce the number of FP SNP calls from neural network variant caller *Clair* v.2.1.1 (Luo et al. 2020). Hybrid lineage alignments of ONT sequence and Illumina background data of the outbreak lineage (ST93) can be constructed, and effective reproduction numbers accurately modeled using birth–death skyline models in *BEAST2*.

We evaluated genotype reconstruction against previously sequenced Illumina data (Steinig et al. 2021) demonstrating the superior quality of hybrid assembly with *Unicycler*. Our genotyping analysis showed that for high-coverage isolates ($>10\times$) genotyping directly from polished nanopore assembly was comparable to hybrid Illumina-ONT approaches (fig. 2). We used the most recent models at the time of writing for base calling (*Bonito* v0.3.6) followed by polished long-read assembly or hybrid assembly. With the imminent release of R10.3 pores and expected increases in raw-read accuracy the remaining misclassifications in genotypes from assemblies (mostly in MLST and SCC*mec* subtyping) will be eliminated and produce nanopore assemblies comparable to reference assemblies at $>10\times$ coverage. We chose here to implement a rapid and minimal protocol to evaluate its application in reference laboratories without local access to sequencing infrastructure, such as at the Australian Institute of Tropical Health and Medicine or the Papua New Guinea Institute of Medical Research. Our method requires some context from genomic surveillance at the level of full lineages (e.g., ST93 or ST772) in order to situate nanopore-sequenced outbreaks within the wider lineage context and fix the clade birth–death model substitution rate. Given that substitution rates vary between *S. aureus* lineages (Steinig et al. 2021), an estimate from the background data is required to fix substitution rates within the outbreak clusters. For optimal polishing results, it appears to be effective to train the random forest polishers on lineage-specific data, noting that effective polishing was still achieved when training isolates derived from a different part of the tree within the lineage (e.g., FNQ-trained polishers were effective on PNG isolates). In higher-coverage isolates effective polishing was also achieved with the mixed *S. aureus* and *N. gonorrhoeae* models; we note that only three isolates with matching Illumina and ONT data are required for training the polishers.

Interestingly, the random forest classifiers failed to polish *Medaka* v1.2.3 reference-specific SNP calls (supplementary fig. S3, Supplementary Material online) even though the *Medaka-Bonito* model is trained explicitly on microbial signal data from *E. coli* and an experimental version (v0.1.0) was successfully used for polishing by Sanderson and colleagues (2020). Polishing success of *Clair* calls suggests that the features selected here—in particular, the proximity and quality features (fig. 3*B–D*)—were effective at removing systematic FP SNP calls. SNP calling did not improve considerably using *Bonito v0.3.6* R9.4.1 DNA models compared with Guppy high accuracy (supplementary fig. S5, Supplementary Material online) and methylation-aware models (data not shown). It remains to be seen whether retraining *Clair* or *Medaka* neural networks on *S. aureus* specific signal and sequence data would improve species-specific SNP calls without polishing.

We demonstrate the utility of our method by sequencing novel isolates of community-associated MRSA from a pediatric osteomyelitis outbreak in the highland towns of Kundiawa and Goroka (PNG) and routine surveillance in remote northern Australia (FNQ; fig. 1, $n = 181$). A protocol that minimized cost (without optimization) allowed us to sequence 2 consecutive panels of 12 isolates with rapid barcoded libraries on a MinION flow cell (SQK-RBK004), by using an interspersing nuclease flush (ONT, EXP-WSH-003). We note that spin column extractions resulted in several fragmented barcodes that failed assembly (12/96). Overall, phylodynamic models were mostly affected by very low-coverage isolates ($<5\times$) whereas even low-medium coverage isolate ($\geq 5\times$) produced consistent estimates of the effective reproduction number for the PNG and FNQ clades, when compared with the Illumina reference (fig. 6). Accurate modeling was possible even with interspecies polishers trained on *N. gonorrhoeae* in higher-coverage isolates in PNG. Estimates were more variable in the low-coverage FNQ outbreak clade and for optimal performance, some protocol optimization will be required, and may include extraction protocols for long reads, inclusion of a magnetic bead cleanup step (obligatory in the latest iteration of the ONT rapid kit protocols, 2021) or short-read elimination kits. Although we were ultimately unable to use a total of 32 isolates ($<5\times$ coverage) in the phylodynamic estimation, the cost per *S. aureus* genome using the $24\times$ multiplex protocol ranges between USD $40 (no failures over 181 unique samples) and around USD $50 per genome with two repeat flow cells from already extracted cultures (supplementary material 2, Supplementary Material online). Further optimization would incur small additional cost and can be conducted for bacterial pathogens of interest in sufficiently resourced laboratories. Further improvements in cost for genome selection by first scanning genomes with approximate genomic neighbor typing approaches may also be feasible (Steinig et al. 2022). Although we chose *S. aureus* as a model organism for this work mainly due to our interest in sequencing the outbreaks in PNG and FNQ, as well as because of its relatively small genome (2.8 Mbp), core principles and methods used in this study are immediately applicable to other bacterial pathogens

Downloaded from https://academic.oup.com/mbe/article/39/3/msac040/6529399 by University of Melbourne Library user on 14 April 2022

and all steps of the pipelines are implemented in replicable Nextflow workflows (Materials and Methods).

We did not expect significant rate variation in the outbreak clades, which made computation of clade parameters with a lineage-wide fixed substitution tractable. We note that within-outbreak patterns of divergence vary between phylogenies (fig. 5), and considering the number of remaining FP and false-negative (FN) SNPs after polishing (fig. 4), we did not expect within-outbreak transmission chains to be reproducible. Optimization of SNP polishing or variant calling, for example, with species-specific neural networks, remains to be investigated. For this study, we accelerated computation using the *BEAGLE* library (Ayres et al. 2019) in combination with *BEAST2*. Moderate acceleration on standard hardware (<5–7 h) and increased acceleration on NVIDIA P100 GPUs (<2 h) were achieved. Nanopore-driven outbreak sequencing and GPU acceleration in *BEAST2* thus enable the rapid deployment of phylodynamic models and responsive surveillance of bacterial diseases.

Ultimately, our cost-effective protocol for multiplex nanopore sequencing and phylodynamic inference of outbreak parameters lowers the barrier of conducting these analyses in scenarios where access to sequencing infrastructure is difficult or infeasible, including in low- or middle-income countries where the burden of bacterial disease outbreaks remains high. In particular, the effective cost of monitoring disease transmission is considerably lower on nanopore sequencing platforms than with gold-standard Illumina sequencing, which may facilitate the sustainable integration of genomic surveillance in reference laboratories located in these regions, including in remote northern Australia and the highlands of PNG. Improvements to the sequencing protocol, for example, by further reducing cost of nucleic acid extraction, increasing the number of isolates per flow cell or balancing throughput per barcode using in silico adaptive sequencing (Payne et al. 2021) will further enable the adoption of phylodynamic surveillance for bacterial outbreaks on nanopore devices.

Materials and Methods

Outbreak Sampling in FNQ and PNG

We collected isolates from outbreaks in two remote populations in northern Australia and PNG (fig. 1). Isolates associated with pediatric osteomyelitis cases (mean age of 8 years) were collected from 2012 to 2017 ($n = 42$) from Kundiawa, Simbu Province (27), and from 2012 to 2018 ($n = 35$) from patients in the neighboring Eastern Highlands province town of Goroka. We supplemented the data with methicillin sensitive *S. aureus* isolates associated with severe hospital-associated infections and blood cultures in Madang (Madang Province; $n = 8$) and Goroka ($n = 12$). Isolates from communities in FNQ, including urban Cairns, the Cape York Peninsula and the Torres Strait Islands ($n = 91$), were a contemporary sample from routine surveillance at Cairns Hospital in 2019. Isolates were recovered on LB agar from clinical specimens using routine microbiological techniques at Queensland Health and the Papua New Guinea Institute of Medical Research (PNGIMR). Isolates were transported on swabs from monocultures to the Australian Institute of Tropical Health and Medicine (AITHM Townsville) where they were cultured in 10 ml Luria-Bertani (LB) broth at 37 °C overnight and stored at −80 °C in glycosol and LB. Illumina short-read data from the ST93 lineage (van Hal et al. 2018) included in this study were collected from the European Nucleotide Archive (supplementary tables, Supplementary Material online).

Nanopore Sequencing and Basecalling

Two milliliters of LB broth was spun down at 5,000 × g for 10 min and after removing the supernatant, 50 µl of 0.5 mg/ml lysostaphin were added to the tube and vortexed. Cell lysis was conducted at 37 °C for 2 h with gentle shaking followed by a *proteinase K* digestion for 30 min at 56 °C. DNA was extracted using a simple column protocol from the DNeasy Blood & Tissue kit (QIAGEN) following the manufacturer's instructions. DNA was eluted in 70 µl of nuclease-free water, quantified on Qubit, and DNA was stored at 4 °C until library preparation. Library preparation was done using approximately 420 ng of DNA and the rapid barcoding kit with 12 barcodes (ONT, SQK-RBK004) as per manufacturer's instructions. Basecalling was done using the R9.4.1 high accuracy (HAC, supplementary fig. S5, Supplementary Material online), the HAC methylation model (not shown), and the all context methylation *Rerio* model (not shown) in *Guppy v4.2.3*, as well as the final *Bonito v0.3.6 R9.4.1* DNA model (used for all analyses), run on a local NVIDIA GTX1080-Ti or a remote cluster of NVIDIA P100 GPUs. Sequence runs were conducted with 2 × 12 barcoded (SQK-RBK004) isolates per flow cell in two consecutive 18–24 h runs. Libraries were nuclease flushed using the wash kit between consecutive runs (EXP-WSH-003). This is sufficiently effective to remove read carry-over, as demonstrated previously with hybrid assemblies of sequentially sequenced Enterobacteriaceae (Lipworth et al. 2020) and our analysis of a single library panel (FNQ-2) sequenced on a previously used flow cell with a human library. After washing with EXP-WSH-003, a total of 2,910/294,461 reads were classified as human in the *S. aureus* library, about twice as much as human contamination in other runs. Sequencing runs were managed on two MinIONs and monitored in *MinKNOW* > v20.3.1.

Nanopore Genome Assembly and Quality Control

Genome assemblies for genotyping were constructed using our Nextflow assembly pipeline (https://github.com/np-core/np-assembly), which first randomly subsamples reads to a maximum of 200× coverage with *rasusa* v0.3.0 (Hall 2022) and filtered Q > 7 with minimum read length of 100 bp using *nanoq v0.8.0* (Steinig and Coin 2022). Fastp v0.20.1 (Chen et al. 2018) was used to trim adapter and low-quality Illumina sequences. We then constructed three types of assemblies: a polished long-read assembly using ONT data only (flye), one with short-read correction of the ONT long-read assembly (pilon) and one that first assembles short reads and scaffolds the assembly with long reads. For the polished long-read assembly, *Flye* v2.8.3 (Kolmogorov et al. 2019) was used

Downloaded from https://academic.oup.com/mbe/article/39/3/msac040/6529399 by University of Melbourne Library user on 14 April 2022

in conjunction with four iterations of *minimap2* v2.17-r941 (Li 2018) + *Racon* 1.4.20 (Vaser et al. 2017) and subsequent polishing with *Medaka* 1.2.3. For the long-read hybrid assembly, corrections were conducted with Illumina paired-end reads for each genome using two iterations of *Pilon* v1.2.3. For the short-read hybrid assembly, we used *Unicycler* v0.4.8. Reference Illumina assemblies were generated with the pipeline *Shovill* v1.1.0 (https://github.com/tseemann/shovill) using *Skesa* v2.4.0 and genotyped with *Mykrobe* v0.9.0 (Hunt et al. 2019) (from reads) and *SCCion* v0.4.0 (https://github.com/esteinig/sccion), a wrapper around common assembly-based genotyping tools and databases (Zankari et al. 2012; Chen et al. 2016; Kaya et al. 2018) for *S. aureus*. We called species, resistance genes, virulence factors, PVL, multilocus sequence type, *mecA*, and major SCC*mec* cassette subtypes. We assessed differences between the Illumina references and hybrid- or nanopore assemblies using the *dnadiff* v1.3 to determine assembly-based differences in SNPs and Indels, as well as assess overall identity between genomes (fig. 2). Coverage against the reference genome (ST93: JKD6159; Chua et al. 2010) was assessed using *CoverM* v0.6.0 (https://github.com/wwood/CoverM).

De Novo Variant Calling and Random Forest SNP Polishers

We called SNPs de novo using the neural network callers *Medaka* v1.2.3 (https://github.com/nanoporetech/medaka) and *Clair* v2.1.1 (shown in example pipeline executions). *Snippy* v4.6.0 (https://github.com/tseemann/snippy) was used to generate a core-site alignment of the ST93 background population ($n = 444$, 6,161 SNPs) and reference Illumina core alignments including the outbreaks in FNQ and PNG isolates ($>5\times$, $n = 531$, 6,580 SNPs). *Snippy* variant calls (SNP type) were used as reference truth for matching ONT and Illumina sequenced isolates. We implemented the feature extraction and random forest design from Sanderson and colleagues (2020) who use the RandomForest classifier from *scikit-learn* (Pedregosa et al. 2011) with default hyperparameter settings and feature extraction with *pysamstats*. Like the original implementation, we subsampled isolates to 2, 5, 10, 20, 50, and 100× coverage with *rasusa* to account for read coverage in training and evaluating the classifiers. For training, we created three sets of matching Illumina and ONT sequence data, each with three isolates for training: three mixed sequence types (ST88, ST15, and ST93; *saureus_mixed*), one of FNQ within-lineage isolates (ST93; *saureus_fnq*), and one of Papua New Guinean within-lineage isolates (ST93; *saureus_png*). Training and validation sets for the classifiers were split into 60% training and 40% validation data.

Next, we evaluated the classifiers, including the *N. gonorrhoeae* classifier trained by Sanderson and colleagues, using the remaining isolates from FNQ and PNG as an independent test data set (fig. 1). We defined true positive (TP) SNPs as those that were called by both Illumina *Snippy* and ONT *Clair*, FP as ONT SNPs that were not called with *Snippy*, and FN *Snippy* calls that were missed by ONT calls or later excluded in the random forest filtering step. Since we used the de novo *Snippy* calls as reference, true negative (TN) calls (sites called as wild type by ONT and *Snippy*) were not able to be considered. We combined data from both outbreaks ($n_{ST93} = 118$, $n_{other} = 44$) and computed accuracy, precision, recall, and F1 scores for each evaluation against Illumina reference data (supplementary tables, Supplementary Material online, fig. 4).

Hybrid Core-Site Outbreak Alignments

To contextualize polished ONT isolates called with *Clair* within the wider background of the ST93 lineage, we adopted the core functionality from *Snippy*'s core alignment caller (*Snippy-core*) into an ONT and Illumina core SNP alignment caller in the *NanoPath* package (https://github.com/np-core/nanopath). Core SNP sites were defined by polymorphic SNP sites present in genomes of all isolates included in the alignment, excluding any site that in any one isolate falls into a gap, or any site with less than a predefined minimum coverage (default: 1×). We first polished ONT SNPs from *Clair* with the trained random forest models, including the *N. gonorrhoeae* data set from Sanderson et al. (2020). We then created reference alignments of the Illumina data (ST93 background and outbreaks, $n = 531$, $>5\times$) with *Snippy-core*, as well as a reference Illumina and polished hybrid alignments with ONT outbreak SNPs in *NanoPath* (fig. 5).

ML Phylogenetics and Bayesian Model Configurations

ML phylogeny of the ST93 lineage was reconstructed from the Illumina and ONT polished alignments, including the outbreaks. We used *RAXML-NG* with the $GTR + G$ and Lewi's ascertainment bias correction for SNP alignments. Trees were rooted on SRR115236 (early isolate from 1992), near the root of the phylogeny (van Hal et al. 2018) and decorated with metadata of sample origin at state level in ITOL (Letunic and Bork 2019). Sampling dates in years were provided for each isolate. We next subset the full lineage alignments to the isolates in the large clades of the FNQ ($n = 36$) and PNG ($n = 62$) outbreaks. We then configured birth–death skyline models in *BEAST2* using a custom Python interface (*NanoPath Beastling*) that stores model configurations of the serially (PNG) and contemporaneously sampled models (FNQ) in YAML files. Birth–death models consider dynamics of a population forward in time using the (transmission) rate λ, the death (become uninfectious) rate δ, the sampling probability ρ, and the time of the start of the population (outbreak; also called origin time) T. The effective reproduction number (R_e), can be directly extracted from these parameters by dividing the birth rate by the death rate ($\lambda \div \delta$). We configured the model priors as outlined in table 1. Importantly, we set a lineage-wide fixed substitution rate prior at 3.199×10^{-4} (Steinig et al. 2021) to account for the loss of temporal signal in the outbreak subset alignments. *NanoPath* constructs the *BEAST2* XML model files which can be run with the *BEAGLE* library on GPU. Results were summarized using the *bdskytools* package in R, where median higher posterior density intervals were computed in custom plotting scripts that can be found along with all other results from the pipelines and model runs at the data repository.

Downloaded from https://academic.oup.com/mbe/article/39/3/msac040/6529399 by University of Melbourne Library user on 14 April 2022

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by a joint Policy Relevant Infectious Disease Simulation and Mathematical Modelling & Improving Health Outcomes in the Tropical North pilot grant (Williams et al. 2021) by the Australian National Health and Medical Research Council (1131932 to E.S. and E.M.), an Improving Health Outcomes in the Tropical North fellowship by the Australian National Health and Medical Research Council (1131932 to C.F.), an Australian National Health and Medical Research Council fellowship (1145033 to S.Y.C.T.), a joint Australian National Health and Medical Research Council and European Union collaborative research grant (GNT1195743 to L.C.), a Queensland Genomics project grant (Vidgen et al. 2021) and a National Health and Medical Research Council Ideas grant (2012286 to P.H., I.A., A.G., C.F., R.F., S.S., E.S., L.C., S.Y.C.T., E.M., and W.P.). Models were run on graphical processing units supported by the Linkage Infrastructure, Equipment and Facilities (LIEF) at the high-performance computing facility hosted at the University of Melbourne (LE170100200; Lafayette et al. 2016).

Author Contributions

E.S., P.H., E.M., L.C., and S.T. planned and conceived of the study. E.S. conducted sequencing, wrote the code, and conducted bioinformatic analysis; E.S. and S.D. conducted phylodynamic analyses; I.A., A.G., R.F., M.Y., J.J., J.D., B.U., H.P., C.W., E.E., D.N., M.L., L.M., C.F., S.S., W.P., and P.H. collected, maintained, and provided the outbreak strains for sequencing, and managed all work in Papua New Guinea and Far North Queensland; E.S. wrote the initial manuscript draft, all authors contributed to the final version.

Data Availability

Sequence data (Illumina, ONT) generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA657380. Additional model results and configuration files may be found in our repository (https://github.com/esteinig/ca-mrsa).

References

Ayres DL, Cummings MP, Baele G, Darling AE, Lewis PO, Swofford DL, Huelsenbeck JP, Lemey P, Rambaut A, Suchard MA. 2019. BEAGLE 3: improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Syst Biol*. 68(6):1052–1061.

Bull RA, Adikari TN, Ferguson JM, Hammond JM, Stevanovski I, Beukers AG, Naing Z, Yeang M, Verich A, Gamaarachchi H, et al. 2020. Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat Commun*. 11(1):6272.

Chen L, Zheng D, Liu B, Yang J, Jin Q. 2016. VFDB 2016: hierarchical and refined dataset for big data analysis–10 years on. *Nucleic Acids Res*. 44(D1):D694–D697.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34(17):i884–i890.

Chua K, Seemann T, Harrison PF, Davies JK, Coutts SJ, Chen H, Haring V, Moore R, Howden BP, Stinear TP. 2010. Complete genome sequence of *Staphylococcus aureus* strain JKD6159, a unique Australian clone of ST93-IV community methicillin-resistant *Staphylococcus aureus*. *J Bacteriol*. 192(20):5556–5557.

da Silva Filipe A, Shepherd JG, Williams T, Hughes J, Aranday-Cortes E, Asamaphan P, Ashraf S, Balcazar C, Brunker K, Campbell A, et al.; COVID-19 Genomics UK (COG-UK) Consortium. 2021. Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland. *Nat Microbiol*. 6(1):112–122.

Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017. Nextflow enables reproducible computational workflows. *Nat Biotechnol*. 35(4):316–319.

du Plessis L, McCrone JT, Zarebski AE, Hill V, Ruis C, Gutierrez B, Raghwani J, Ashworth J, Colquhoun R, Connor TR, et al.; COVID-19 Genomics UK (COG-UK) Consortium. 2021. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* 371(6530):708–712.

Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G. 2020. Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol*. 6(2):veaa061.

Duchêne S, Geoghegan JL, Holmes EC, Ho SYW. 2016. Estimating evolutionary rates using time-structured data: a general comparison of phylogenetic methods. *Bioinformatics* 32(22):3375–3379.

Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, Fourment M, Holmes EC. 2016. Genome-scale rates of evolutionary change in bacteria. *Microb Genom*. 2(11):e000094.

Duchene S, Lemey P, Stadler T, Ho SYW, Duchene DA, Dhanasekaran V, Baele G. 2020. Bayesian evaluation of temporal signal in measurably evolving populations. *Mol Biol Evol*. 37(11):3363–3379.

Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, Kraemer MUG, Hill SC, Black A, da Costa AC, et al. 2017. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* 546(7658):406–410.

Ferreira FA, Helmersen K, Visnovska T, Jørgensen SB, Aamot HV. 2021. Rapid nanopore-based DNA sequencing protocol of antibiotic-resistant bacteria for use in surveillance and outbreak investigation. *Microb Genom*. 7:000557.

Gardy JL, Loman NJ. 2018. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat Rev Genet*. 19(1):9–20.

Giovanetti M, Faria NR, Lourenço J, Goes de Jesus J, Xavier J, Claro IM, Kraemer MUG, Fonseca V, Dellicour S, Thézé J, et al. 2020. Genomic and epidemiological surveillance of zika virus in the amazon region. *Cell Rep*. 30(7):2275–2283.e7.

Golparian D, Donà V, Sánchez-Busó L, Foerster S, Harris S, Endimiani A, Low N, Unemo M. 2018. Antimicrobial resistance prediction and phylogenetic analysis of *Neisseria gonorrhoeae* isolates using the Oxford Nanopore MinION sequencer. *Sci Rep*. 8(1):17596.

Gorrie CL, Da Silva AG, Ingle DJ, Higgs C, Seemann T, Stinear TP, Williamson DA, Kwong JC, Grayson ML, Sherry NL, et al. 2021. Key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria: a systematic analysis. *Lancet Microb*. 2(11):e575–e583.

Hall M. 2022. Rasusa: randomly subsample sequencing reads to a specified coverage. *J Open Source Softw*. 7(69):3941.

Hammer AS, Quaade ML, Rasmussen TB, Fonager J, Rasmussen M, Mundbjerg K, Lohse L, Strandbygaard B, Jørgensen CS, Alfaro-Núñez A, et al. 2021. SARS-CoV-2 transmission between mink (*Neovison vison*) and humans, Denmark. *Emerg Infect Dis*. 27(2):547–551.

Hodcroft EB, Zuber M, Nadeau S, Vaughan TG, Crawford KHD, Althaus CL, Reichmuth ML, Bowen JE, Walls AC, Corti D, et al.; SeqCOVID-SPAIN consortium. 2021. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* 595(7869):707–712.

Hunt M, Bradley P, Lapierre SG, Heys S, Thomsit M, Hall MB, Malone KM, Wintringer P, Walker TM, Cirillo DM, et al. 2019. Antibiotic resistance prediction for *Mycobacterium tuberculosis* from genome sequence data with Mykrobe. *Wellcome Open Res*. 4:191.

Downloaded from https://academic.oup.com/mbe/article/39/3/msac040/6529399 by University of Melbourne Library user on 14 April 2022

Ingle DJ, Howden BP, Duchene S. 2021. Development of phylodynamic methods for bacterial pathogens. *Trends Microbiol*. 29(9):788–797.
Kaya H, Hasman H, Larsen J, Stegger M, Johannesen TB, Allesøe RL, Lemvigh CK, Aarestrup FM, Lund O, Larsen AR. 2018. SCCmecFinder, a web-based tool for typing of Staphylococcal Cassette chromosome *mec* in *Staphylococcus aureus* using whole-genome sequence data. *mSphere* 3(1):e00612-17.
Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*. 37(5):540–546.
Lafayette L, Sauter G, Vu L, Meade B. 2016. Spartan performance and flexibility: an HPC-cloud chimera. *OpenStack Summit*. 27:1.
Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. 47(W1):W256–W259.
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34(18):3094–3100.
Lipworth S, Pickford H, Sanderson N, Chau KK, Kavanagh J, Barker L, Vaughan A, Swann J, Andersson M, Jeffery K, et al. 2020. Optimized use of Oxford Nanopore flowcells for hybrid assemblies. *Microb Genom*. 6:11.
Luo R, Wong C-L, Wong Y-S, Tang C-I, Liu C-M, Leung C-M, Lam T-W. 2020. Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nat Mach Intell*. 2(4):220–227.
Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol*. 14(1):e1005944.
Nicholls SM, Poplawski R, Bull MJ, Underwood A, Chapman M, Abu-Dahab K, Taylor B, Colquhoun RM, Rowe WPM, Jackson B, et al.; COVID-19 Genomics UK (COG-UK) Consortium. 2021. CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. *Genome Biol*. 22(1):196.
Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. 2021. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat Biotechnol*. 39(4):442–450.
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 12:2825–2830.
Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, Oliveira G, Robles-Sikisaka R, Rogers TF, Beutler NA, et al. 2017. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc*. 12(6):1261–1276.
Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, et al. 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530(7589):228–232.
Sanderson ND, Swann J, Barker L, Kavanagh J, Hoosdally S, Crook D, Street TL, Eyre DW; The GonFast Investigators Group. 2020. High precision *Neisseria gonorrhoeae* variant and antimicrobial resistance calling from metagenomic nanopore sequencing. *Genome Res*. 30(9):1354–1363.
Souvorov A, Agarwala R, Lipman DJ. 2018. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biol*. 19(1):153.
Steinig E, Aglua I, Duchêne S, Meehan MT, Yoannes M, Firth C, Jaworski J, Drekore J, Urakoko B, Poka H, et al. 2021. Phylodynamic signatures in the emergence of community-associated MRSA. *bioRxiv* [Internet; cited 2021 May 1]. Available from: https://www.biorxiv.org/content/early/2021/04/30/2021.04.30.442212
Steinig E, Coin L. 2022. Nanoq: ultra-fast quality control for nanopore reads. *J Open Source Softw*. 7(69):2991.
Steinig E, Pitt M, Aglua I, Suttie A, Greenhill A, Heather C, Firth C, Smith S, Pomat W, Horwood P, et al. 2022. Genomic neighbor typing for bacterial outbreak surveillance. *bioRxiv* [Internet; cited 2022 Feb 6]. Available from: https://www.biorxiv.org/content/early/2022/02/06/2022.02.05.479210
Urban L, Holzer A, Baronas JJ, Hall MB, Braeuninger-Weimer P, Scherm MJ, Kunz DJ, Perera SN, Martin-Herranz DE, Tipper ET, et al. 2021. Freshwater monitoring by nanopore sequencing. *Elife* 10:e61504.
van Hal SJ, Steinig EJ, Andersson P, Holden MTG, Harris SR, Nimmo GR, Williamson DA, Heffernan H, Ritchie SR, Kearns AM, et al. 2018. Global scale dissemination of ST93: a divergent *Staphylococcus aureus* epidemic lineage that has recently emerged from remote Northern Australia. *Front Microbiol*. 9:1453.
Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*. 27(5):737–746.
Vidgen ME, Williamson D, Cutler K, McCafferty C, Ward RL, McNeil K, Waddell N, Bunker D. 2021. Queensland Genomics: an adaptive approach for integrating genomics into a public healthcare system. *NPJ Genom Med*. 6(1):71.
Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, Hinsley WR, Laydon DJ, Dabrera G, O'Toole Á, et al.; COVID-19 Genomics UK (COG-UK) consortium. 2021. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* 593(7858):266–269.
Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9(11):e112963.
Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol*. 13(6):e1005595.
Williams K, Rung S, D'Antoine H, Currie BJ. 2021. A cross-jurisdictional research collaboration aiming to improve health outcomes in the tropical north of Australia. *Lancet Reg Health West Pac*. 9:100124.
Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother*. 67(11):2640–2644.

Downloaded from https://academic.oup.com/mbe/article/39/3/msac040/6529399 by University of Melbourne Library user on 14 April 2022