Supplementary Materials

for

Phylodynamic modelling of bacterial outbreaks using nanopore sequencing

Eike Steinig, Sebastián Duchêne, Izzard Aglua, Andrew Greenhill, Rebecca Ford, Mition Yoannes, Jan Jaworski, Jimmy Drekore, Bohu Urakoko, Harry Poka, Clive Wurr, Eri Ebos, David Nangen, Moses Laman, Laurens Manning, Cadhla Firth, Simon Smith, William Pomat, Lachlan Coin, Steven Y.C. Tong, Emma McBryde, Paul Horwood

Table of Contents

Analysis S1: Cost estimate of multiplex sequencing

Table S1 for protocol component cost estimates (in Australian Dollars, May 2021)

Scenario 1: standard protocol per isolate

$4\ x\ FLOMIN + 8\ x\ RBK + 96\ x\ LS + 96\ x$ DNA + 96 x QBIT + 4 x WSH = \$4,882

Per genome, success across all isolates: \$50.22 (n = 192) or \$53.29 (n = 181)

Scenario 2: standard protocol with resequencing of already extracted isolates (n = 48)

$6\ x\ FLOMIN + 12\ x\ RBK + 96\ x\ LS + 96\ x$ DNA + 96 x QBIT + 6 x WSH = \$6,699

Per genome, resequencing on two flow cells: \$ 69.78 (n = 192+48) or \$75.28 (n = 181+48)

Analysis S2: Candidate-guided SNP calls using *Megalodon*

We evaluated a candidate-guided approach to reconstruct the phylogenetic divergences of nanopore-sequenced outbreak in PNG using a set SNPs at sites present in all isolates (core SNPs) called from existing population-wide background data of the ST93 lineage with *Snippy v4.6.0* (Illumina, n = 444, SNPs). SNPs from the known population were used as input to the candidate variant calling workflow in *Megalodon v2.2.10* (methylation-ware high-accuracy model, *Guppy v4.2.3*) and merged with the alignment of the background population (n = 495, SNPs). We used only isolates that passed genome assembly for the variant calling and phylogenetics (Fig. 1). Although slight variations in tree topology were observed in the divergence of the smaller monophyletic introduction into Papua New Guinea (PNG-2), the outbreaks diverged from their respective source populations (East Coast - PNG, Northeastern - FNQ) and deduction of their regional origin was not affected (Fig. S1). Importantly, putative transmissions from Papua New Guinea remained recognizable in the candidate-guided approach, thus allowing for the correct inference of sporadic regional transmission events (grey inside blue clade, Fig. S1). We estimated the date of divergence from the source population on the candidate phylogeny using *Treetime* *v0.8.1* (PNG-1: 2004.45, 90% maximum posterior region (MPR): 2003.02 – 2005.71, PNG-2: 2000.74, 90% MPR: 2000.11 – 2001.47) and found that it reasonably approximated the estimate from the Illumina reference phylogeny (PNG-1: 2002.09, 90% MPR: 2000.97 – 2003.81, PNG-2: 2000.36, 9% MPR: 1999.73 – 2001.34). Estimates of lineage-wide substitution rates estimates from the SNP alignments were moderately consistent between the candidate approach (2.884e-04 +- 1.30e-05 SD) and the Illumina reference (3.174e-04 +- 1.19e-05 SD) and fall within the expected range of other *S. aureus* lineages and previous estimates for ST93. However, it was not possible to recreate within-outbreak relationships, because novel variation in the outbreaks was not captured in the core-genome variants of the background population (blue). Since the outbreak in the highlands has been ongoing since at least the 2000s, sufficient novel variation has accumulated in the PNG clade. Thus, given the absence of informative branch lengths within outbreaks, we were unable to conduct additional outbreak-specific phylodynamic analysis of these data.

Table S1: Cost estimates of individual protocol reagents per genome

Protocol item	**Reactions**	**AUD**	**AUD per reaction**
SQK-RBK004	6	835	139.16
FLOMIN-106D	48	30,800	641.66
EXP-WSH003	6	110	18.30
Lysostaphin	315	835	3.15
DNeasy	250	1,728	6.91
Qubit	500	542	1.08

Table S2: Birth death skyline posterior estimates from nanopore sequencing data

Alignment	**MRCA**	**R_e**
PNG-Illumina	2000 (1997 - 2002)	2.54 (1.20 – 4.93)
PNG-sanderson	2000 (1998 – 2002)	2.59 (1.24 – 4.98)
PNG-saureus_mix	1995 (1992 - 2000)	2.55 (1.21 - 4.92)
PNG-saureus_fnq	*2000 (1997 – 2002)*	*2.58 (1.21 – 5.00)*
PNG-saureus_png	2000 (1998 – 2002)	2.64 (1.25 – 5.07)
FNQ-Illumina	2006 (2004 - 2008)	2.36 (1.17 - 4.62)
FNQ -sanderson	1857 (1845 – 1867)	1.02 (0.90 - 1.21)
FNQ -saureus_mix	1982 (1977 - 1986)	1.71 (1.04 - 3.24)
FNQ -saureus_fnq	2000 (1997 – 2004)	2.42 (1.17 - 4.68)
FNQ -saureus_png	*2004 (2002 – 2007)*	*2.54 (1.19 - 4.92)*

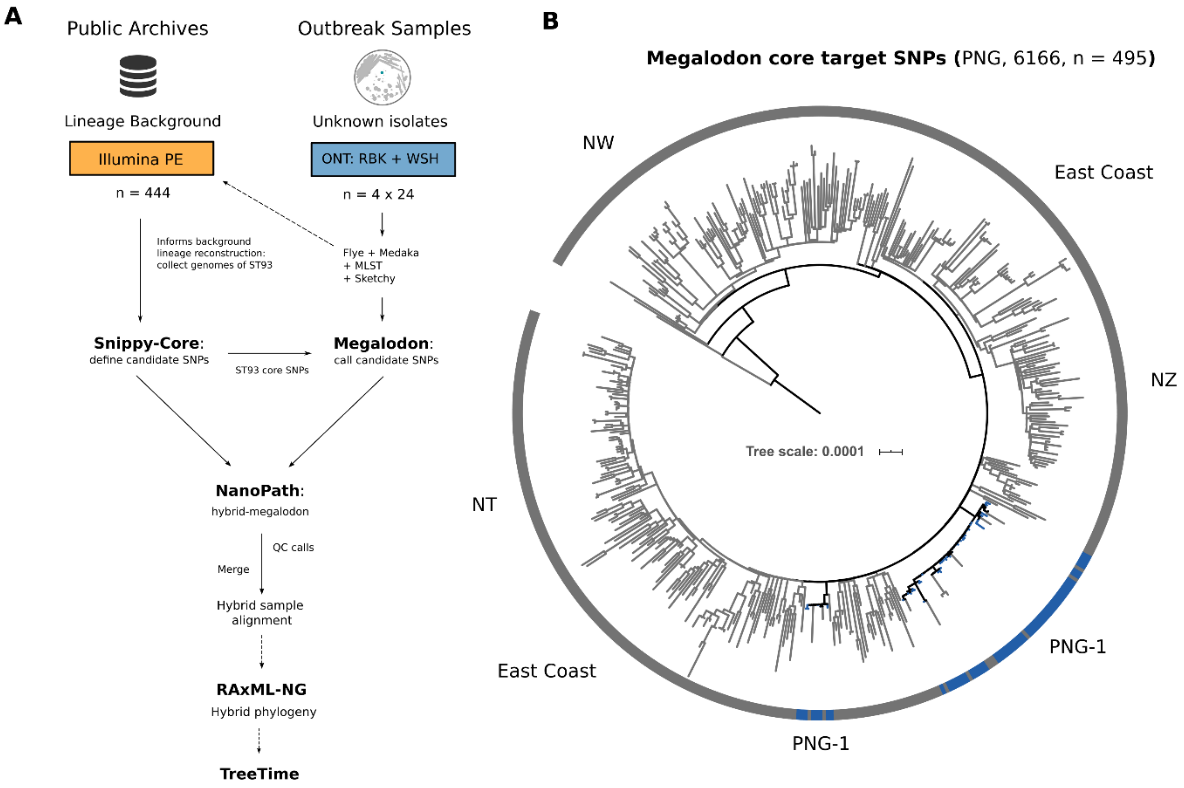

Fig. S1: Candidate-guided variant calling workflow **(A)** and phylogenetic reconstruction of the Papua New Guinea (PNG) clusters PNG-1 and PNG-2 in the ML phylogeny **(B)** using candidate-guided core SNP sites from the lineage background population.

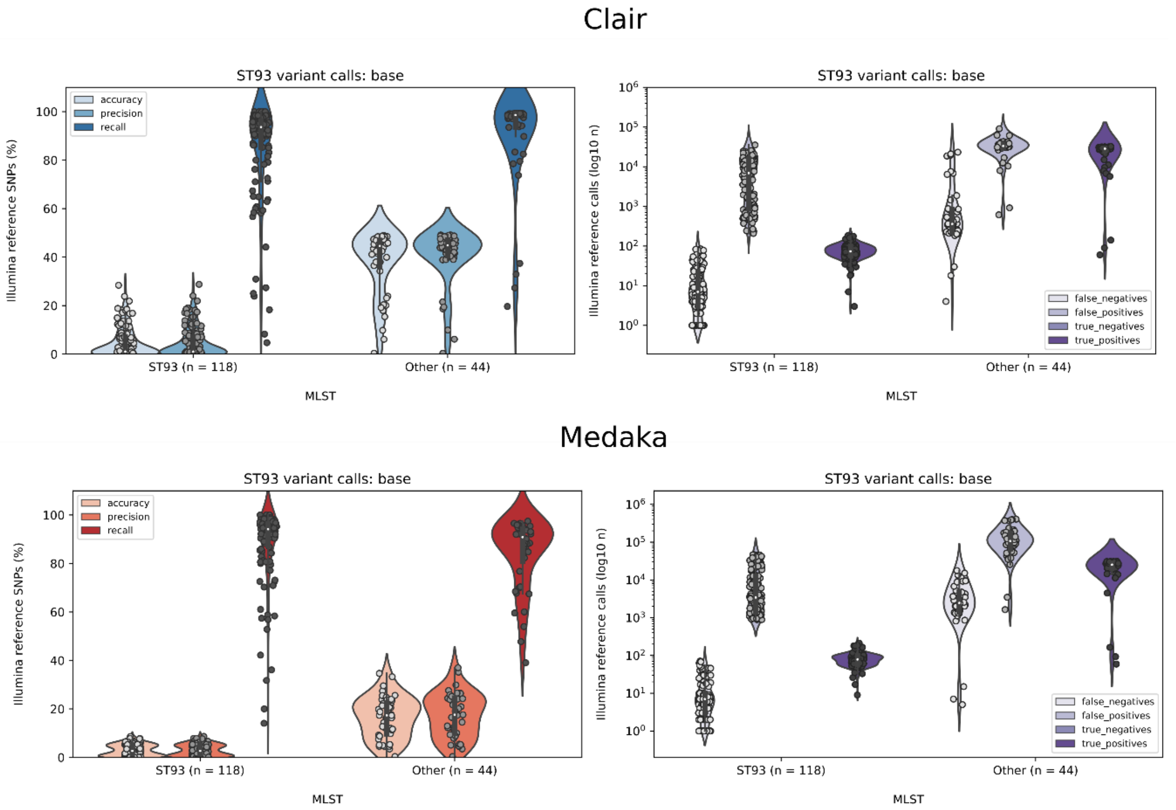

Fig. S2: Raw SNP call accuracy, precision and recall (left, all isolates split into ST93 and other sequence types) from *Clair* (blue) and *Medaka* (red). Right plots show absolute numbers of false negatives, false positives and true positives on a log scale.

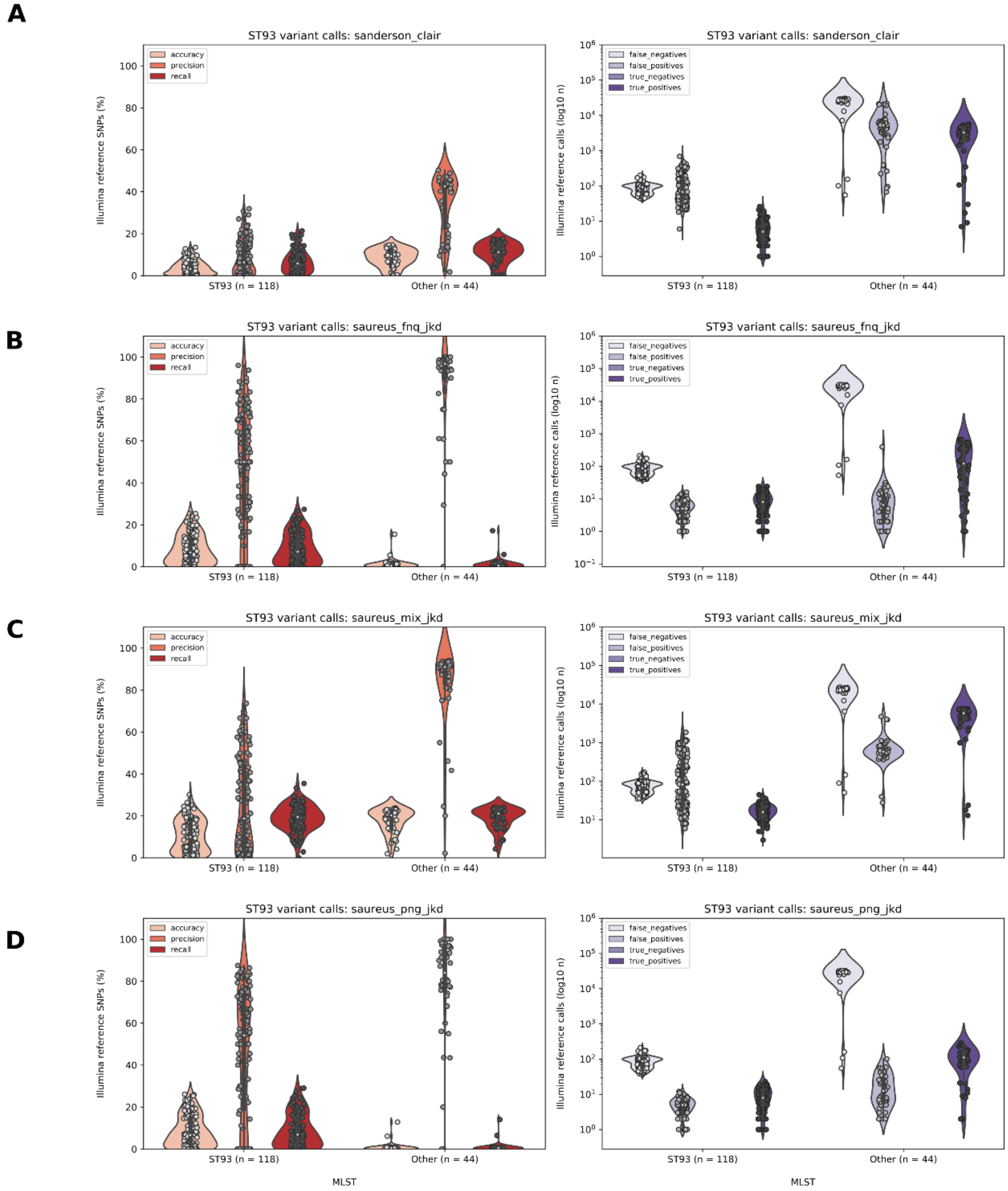

Fig S3: *Medaka* (all coverage) SNP calling accuracy, precision and recall compared to Snippy (Illumina reference) calls. Left plots show the metric distribution across ST93 (outbreak clades, n = 118) and other sequence types (split panels, n = 44). Right plots show absolute numbers of false negatives, false positives and true positives on a log scale.

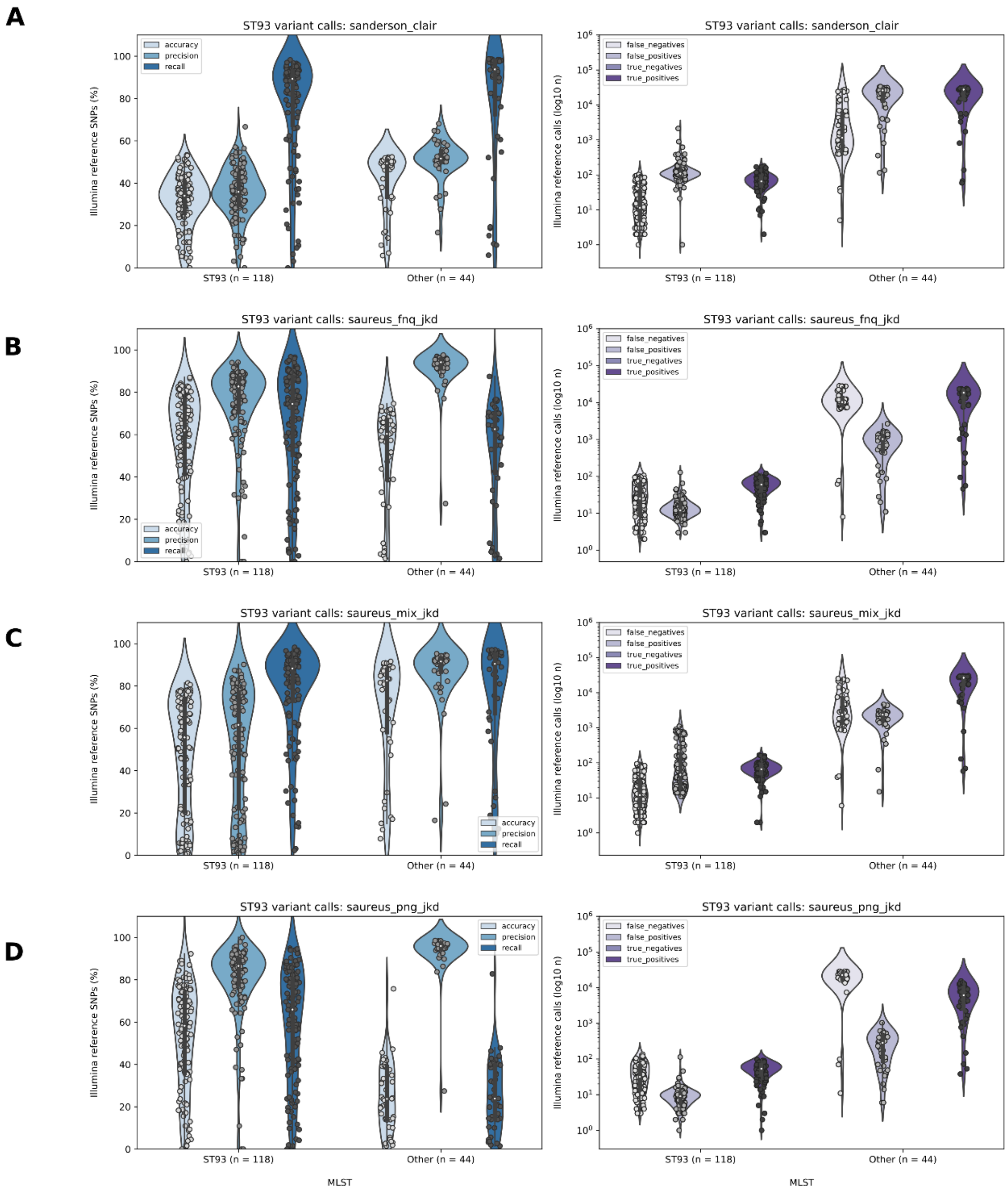

Fig. S4: *Clair* (all coverage, n = 159) SNP calling accuracy, precision and recall compared to *Snippy* (Illumina reference) calls. Left plots show the metric distributions across ST93 (outbreak clades) and other sequence types (split panels). Right plots show absolute numbers of false negatives, false positives and true positives on a log scale.

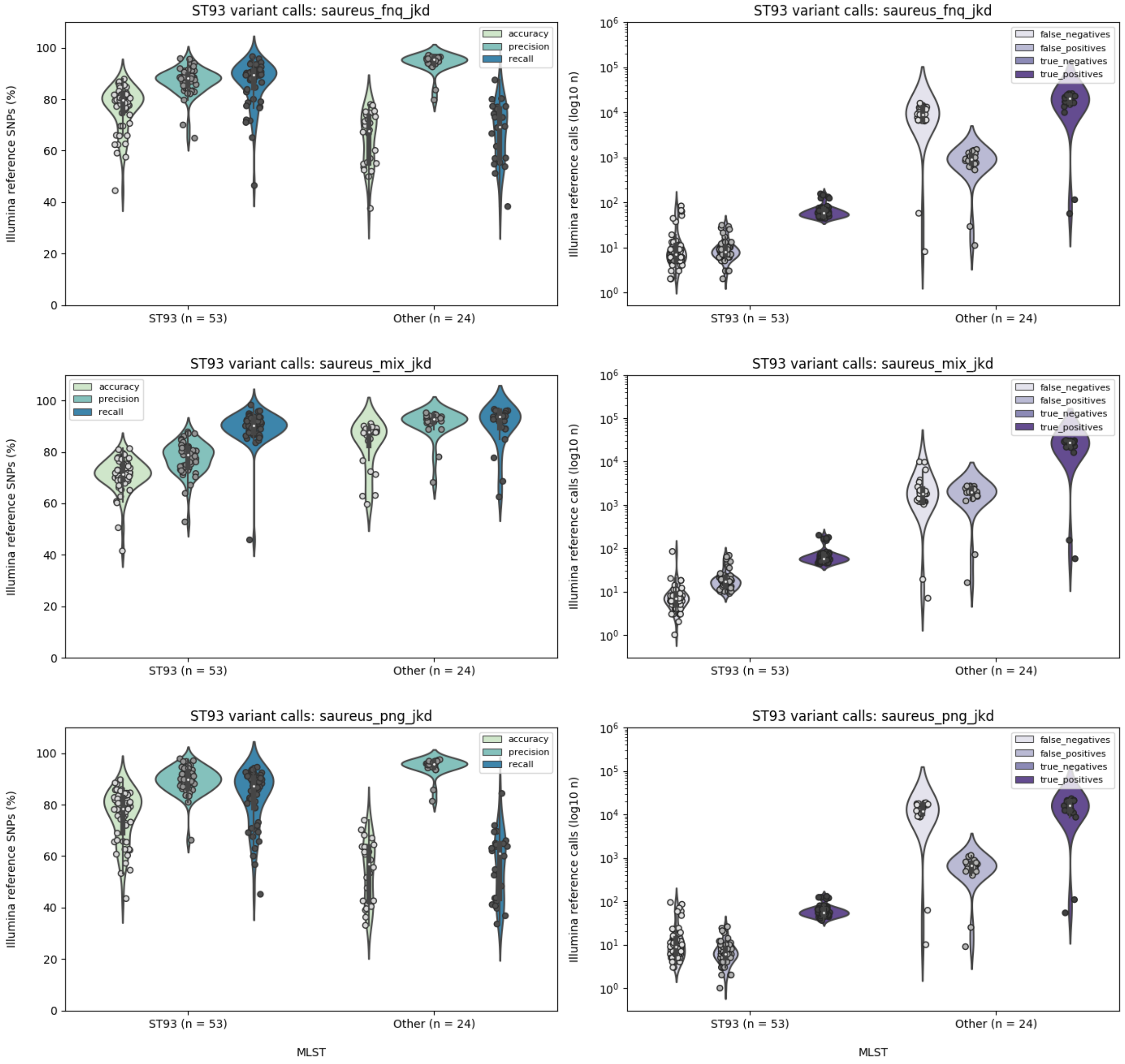

Fig S5: Isolates from the Papua New Guinean outbreak polished using Random Forest classifiers on *Guppy v.4.2.3* (high accuracy model) base called reads and SNP calls using *Clair* showing similar error profiles as *Bonito v0.3.6* base called reads and SNP calls.

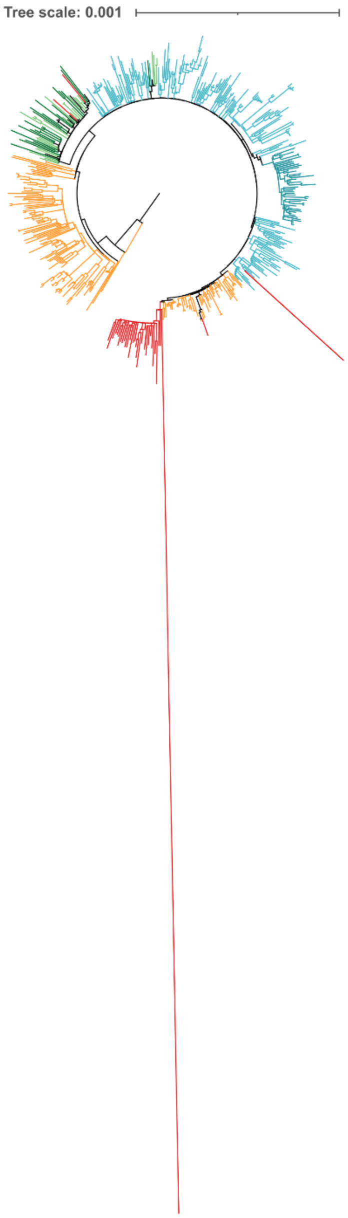

Fig. S6: Maximum likelihood phylogeny of *Neisseria gonorrhea* polished ONT SNPs of ST93-MRSA-IV outbreaks in Far North Queensland (FNQ, red) and Papua New Guinea (green). Complete branches are shown compared with Fig. 5, including extremely abnormal branch length of FNQ-36 and FNQ-62.

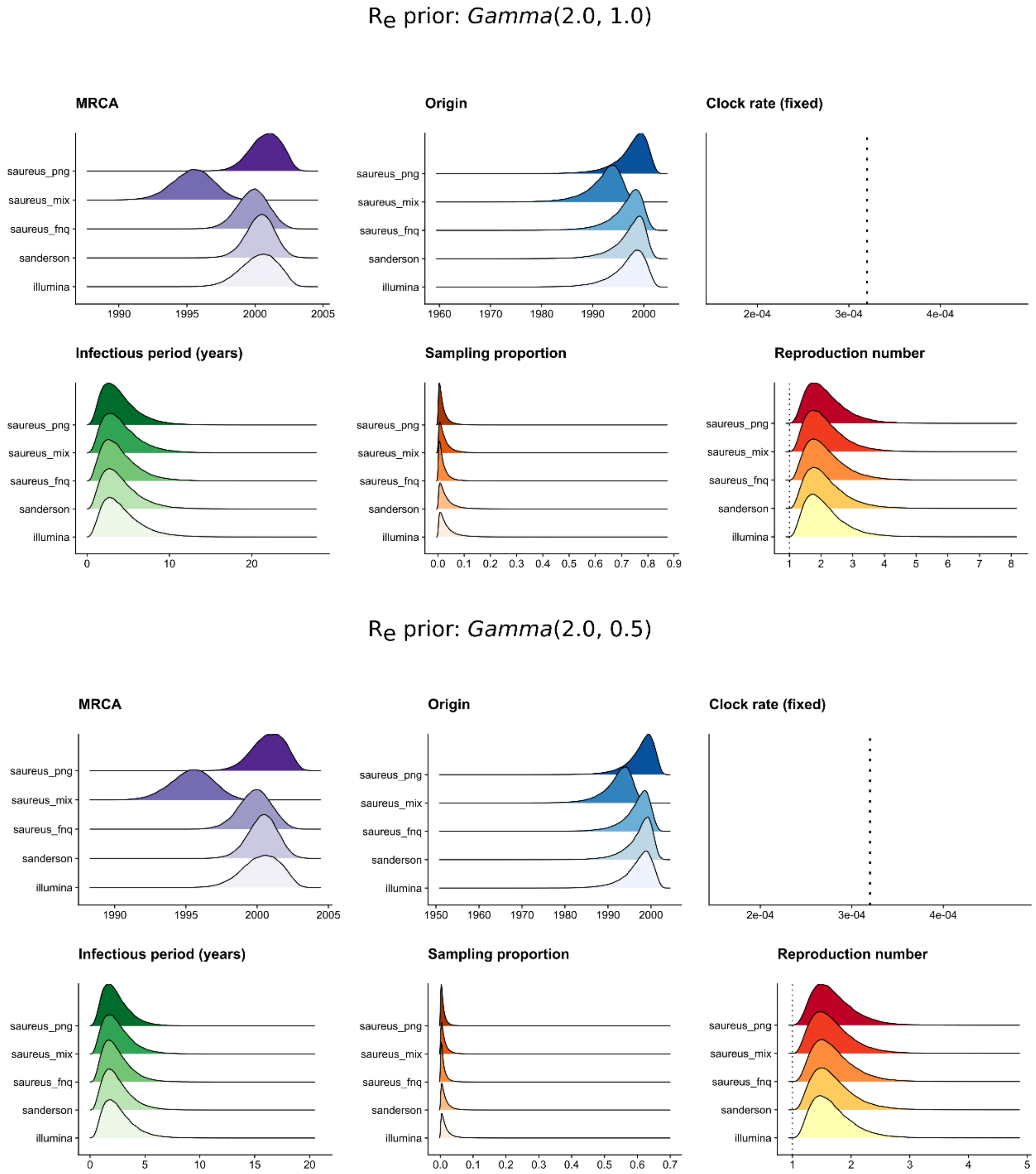

Fig. S7: Birth-death skyline posterior estimates of the serially sampled PNG outbreak of ST93-MRSA-IV with a different prior of the effective reproduction number (R_e) using a Gamma(2.0, 1.0) and Gamma(2.0, 0.5) prior distribution.

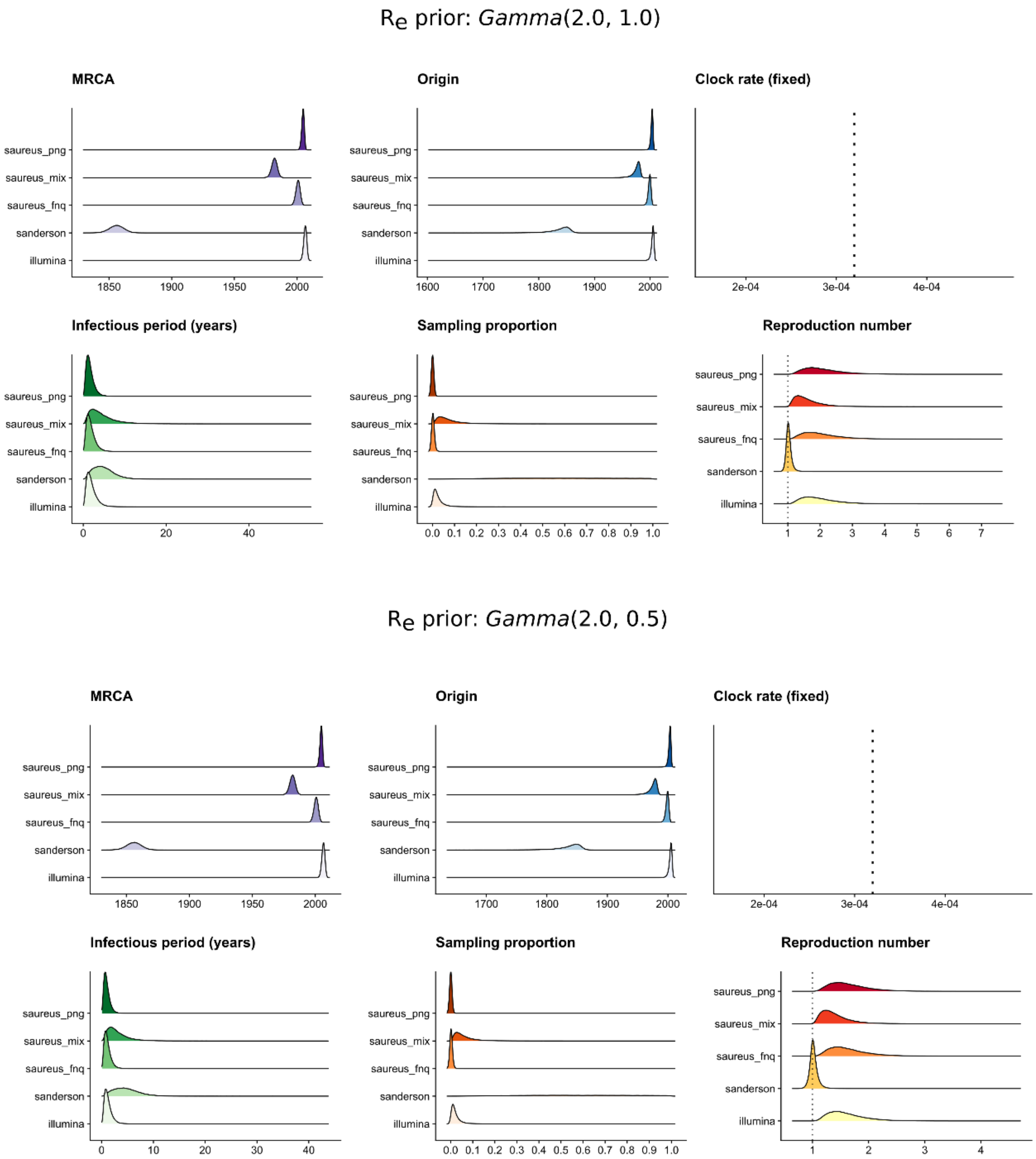

Fig. S8: Birth-death skyline posterior estimates of the contemporaneously sampled FNQ outbreak of ST93-MRSA-IV with a different prior of the effective reproduction number (R_e) using a Gamma(2.0, 1.0) and Gamma(2.0, 0.5) prior distribution.