Genomic epidemiology of community-associated *Staphylococcus aureus* in northern Australia and Papua New Guinea

Eike Joachim Steinig (BSc. Hons.)

in fulfillment of the degree of

Doctor of Philosophy (Medical and Biomedical Sciences)

at the

Australian Institute of Tropical Health and Medicine

College of Public Health, Medical and Veterinary Sciences
James Cook University, Townsville, Queensland, Australia

July 2021

Christian Drosten & Sandra Ciesek

(Deutschlandfunk, 2020 - 2022)

ACKNOWLEDGEMENTS

This project would not have been possible, if not for the extraordinary trust and support that my advisors Emma McBryde and Steven Tong have sustained for all this time. This includes my honorary advisors Lachlan Coin and Paul Horwood, who have been beyond generous with their time and patience. None of this work would have been possible without our amazing collaborators across the world: at the Papua New Guinea Institute of Medical Research, the Simbu Childrens' Foundation, Kundiawa and Goroka hospitals, in particular Izzard Aglua, William Pomat, Rebecca Ford, Andrew Greenhill and Mition Yoannes, who have gone above and beyond to investigate and control transmission in the community and have provided us with samples for sequencing; Mohammed Ali Syed and Bushra Jamil, who have generously provided us with additional samples from Pakistan; Simon Smith and his team at Cairns Hospital for their sample collection across the Cape York Peninsula; Christopher Heather at Townsville Hospital, PI of our SERTA grant and coordinator for COVID testing in Townsville; Bart Currie and Kevin Williams for their amazing support for sequencing the FNQ strains through the HOT North program; Dana Williamson and Patrick Harris for their kindness and support through the QLD Genomics program; my colleagues in the *Staphylococcus* community, who have been a great pleasure to work with, in particular (but not limited to) Marc Stegger, Ralf Ehricht, Stefan Monecke, Megan Earls and David Coleman; Tim Stinear and colleagues at the Doherty; Kyall Zenger for his support over the years; our lab- and admin support staff, without whom none of this would have happened, especially Chris Wright, Lachlan Pomfrett and Jasmine Bell; Tania Duarte and Miranda Pitt for their support with the chapters, my friends and colleagues and students in the sciences, especially Michael Meehan, Sebastian Duchene, Sandra Erdmann, Paul O'Brien and Dan Rawlinson; Roger Huerlimann who is the reason I am in this cursed science thing in the first place; Cadhla Firth who has kept my sanity the last few years and has been my source for much needed cynicism and snark; my friends in Townsville and Berlin, whom I have missed dearly in the last year; and Jessica Hay, who has been my rock (rock) throughout all this madness.

STATEMENT of CONTRIBUTION OF OTHERS

Assistance	Contribution	Co-Contributors
Intellectual	Grant applications	Prof. Emma McBryde, Dr. Christopher Heather, Dr. Paul Horwood, Dr. Cadhla Firth, A/Prof. Steven Tong
	Data Analysis	Dr. Michael Meehan, Dr. Sebastian Duchene, Dr.Cadhla Firth, Prof. Lachlan Coin
	Editorial assistance	Dr. Cadhla Firth, Dr. Michael Meehan, A/Prof Steven Tong
Financial	Fee offset/waiver	Prof. Emma McBryde
	Research costs	Prof. Emma McBryde, A/Prof. Steven Tong
	Stipend	Prof. Emma McBryde
Data	Microbiology and sequencing assistance	Dr. Peter Mulvey, Dr. Annika Suttie, Dr. Miranda Pitt, Dr. Paul Horwood, Dr. Andrew Greenhill, Dr. Elke Müller, Dr. Mition Yoannes, Dr. Rebecca Ford, Dr. Tania Duarte, Prof. Tim Stinear, A/Prof. Torsten Seemann
	Clinical sampling, field-work and laboratory maintenance of strains	Dr. Christopher Heather, Dr. Simon Smith, Dr. Stefan Monecke, Prof. Ralf Ehricht, A/Prof. Muhammed Ali Syed, Prof. Bushra Jamil, Dr. Izzard Aglua, Dr. Rebecca Ford, Dr. Mition Yoannes, Dr. Jan Jorowski, Dr. Jimmy Drekore, Dr. Bohu Urakoko, Dr. Harry Poka, Dr. Clive Wurr, Dr. Eri Ebos, Dr. David Nangen, Dr. Moses Laman, Prof. Laurens Manning, Prof. WIlliam Pomat
	Unpublished data and strains	Dr. Marc Stegger, Dr. Megan Earls, Prof. David Coleman, Dr. Tania Duarte, Prof. Lachlan Coin, Dr. Hege Vangstein Aamot, Dr. Anita Blomfeldt, Dr. Charlene Jackson, Dr. Erin Price

ABSTRACT

Community-associated, methicillin-resistant *Staphylococcus aureus* (MRSA) lineages have emerged in geographically distinct regions around the world during the past thirty years. Despite comprehensive collection efforts reaching back decades and recent analysis of whole genome datasets that chronicle the evolution of some of these lineages, the genomic and epidemiological drivers behind their emergence are poorly understood. Application of genomic technology has been particularly limited in rural and remote populations of the South-Western Pacific, such as in northern Australia and Papua New Guinea, where sequencing infrastructure has not been accessible. In this study, we aimed to investigate the genomic epidemiology and transmission dynamics of community-associated *S. aureus* from community outbreaks in Far North Queensland (FNQ) and the remote highland regions of Papua New Guinea (PNG). Simultaneously, we evaluated the application of portable nanopore sequencing technology for bacterial outbreak transmission and genotype inference.

In the first chapter, we used traditional Illumina sequencing to investigate provenance and transmission dynamics of community-associated *S. aureus*. Samples were collected from routine presentation of patients from FNQ communities, as well as from a pediatric osteomyelitis outbreak in the highlands of PNG, including the first *S. aureus* genomes ever sequenced from the country. We found that the outbreak in PNG was caused by the dominant Australian community lineage (ST93-MRSA-IV). Surges in the effective reproduction number indicating epidemic growth ($R > 1$) were associated with drug-resistance acquisition and spread into vulnerable host populations. We used comparative phylodynamic modelling to investigate the transmission dynamics of global community-associated lineages, detecting signatures of epidemic growth at the divergence of drug-resistant strains, that indicated sustained, epidemic

growth in nearby population centers, such as on the Australian East Coast or the Indian subcontinent. Our data suggests that community-associated strains emerge by secondary spillover into vulnerable host populations, facilitated by the acquisition of drug resistance, and leaving distinct signatures of epidemic growth in lineage- and outbreak-resolved phylogenetic trees.

In the second and third chapter, we developed nanopore-driven methods to enable the application of phylodynamic modelling and rapid inference of clinical genotypes for bacterial pathogens. Outbreaks from FNQ and PNG were re-sequenced using sequential, low-coverage multiplex panels at a cost of ~$50 and ~5-15x coverage per *S. aureus* genome (MinION). We then adopted Random Forest machine learning classifiers to reduce the false positive rate of variant calls from nanopore data, demonstrating that polished alignments were as accurate as Illumina reference data for inferring geographical source, date of divergence and effective reproduction numbers of the outbreaks. In the third chapter, we adopted the heuristic principle of genomic neighbor typing for rapid genotype inference on nanopore devices using MinHash techniques and expanded its application to species-wide genome collections, comprising tens of thousands of whole genome genotypes from public archives. Using the re-sequenced outbreak data as independent test set, we demonstrated high recall and precision across clinically important features for outbreak genotype surveillance, including on Flongle adaptors, multiplexing 48 strains on a single flow cell, and conducting genotyping *in situ* under challenging conditions at the Papua New Guinea Institute of Medical Research.

TABLE of CONTENTS

1. Introduction

Antibiotic resistance is considered amongst the great challenges of our time. For a considerable part of human history, bacterial diseases such as diphtheria, the bubonic plague or tuberculosis have posed an often insurmountable challenge to their hosts (*1*, *2*). The discovery of drugs that are capable of combating human pathogens, as well as their subsequent industrial scale production and proliferation in hospitals and healthcare systems, is nothing short of remarkable and a testament to the ingenuity and labor of generations of scientists (*3*–*5*). Since the early 20th century modern medical practice has been dependent on the use of antibiotics for effective prevention and treatment of infectious diseases in everyday clinical practice, from urinary tract infections to surgical prophylaxis and critical care (*6*). Given the immense evolutionary pressures associated with its genetic maintenance and proliferation in favorable environments (*7*), the ease of vertical and horizontal transmission of genetic determinants (*8*), and wide-spread use in the late 20th century in communities and agricultural sectors amongst industrialized and industrializing nations (*9*, *10*), it is not surprising that resistant bacterial lineages have emerged across the world. While we have tracked and conducted extensive microbiological and epidemiological surveillance of the transmission and evolutions of these lineages, only in the last two decades the widespread adoption of high throughput sequencing technology has enabled us to better probe the genetic changes that interact with and drive the epidemiological phenotype of communicable diseases (*11*). Whole genome sequencing of viral, bacterial and eukaryotic pathogens has accelerated our understanding of the way microbial pathogens evolve, emerge and are transmitted on a national and international scale. Writing this with some (limited) hindsight in 2021, we are extremely fortunate to have built on two decades of genomic sequencing, where methods developed during the rise and adoption of genomic technology in the healthcare system, including for infectious disease transmission, have become front-line surveillance and outbreak response tools in the current SARS-CoV-2

pandemic (*12*–*15*) and in communities across the world (*16*). In cases such as Sam Nicholls' MAJORA (*12*) and many others, these computational tools have been developed by individuals and students, who have built on years of experience on genome-informed disease surveillance research.

Nanopore-based sequencing technology is a good example of thirty years of foundational research in biochemistry and genomics that have enabled commercial production and dissemination as a viable platform (*17*). Portable Oxford Nanopore Technologies (ONT) sequencing devices have now been widely deployed to generate real-time genomic data on transmission and surveillance of viral diseases (*18*–*21*) generating around one third of the global sequencing data on the current SARS-CoV-2 pandemic (approximation based on GISAID data presented at London Calling, May 2021). Having prototyped the technology during the Ebola and Zika outbreaks, with collaborators across the globe, research groups in the United Kingdom were able to build a nationally integrated genomic surveillance system on the top of a national computational laboratory and information management system (LIMS) for genomic data (*12*, *22*) (with over a million genomes contributed to date). Denmark, which currently sequences around 90% of national SARS-CoV samples initially started with a dozen MinION platforms operated by Mads Albertsen's research group (May 2021, *pers. comm.* Marc Stegger, Statens Serum Institute). At the height of the second wave in Denmark this setup (now adopted and expanded to other sequencing technologies by the Statens Serum Institute) was providing unprecedented resolution on transmission pathways and lineage evolution, chronicling the rise of the B1.1.7 (Alpha) variant in the following months (*21*). Nanopore technology has also been critical in outbreaks occurring in rural locations with no existing (and expensive) sequencing infrastructure, such as during the initial surveillance deployments in the Sierra Leone Ebola outbreak (*23*), the Zika epidemic (*24*) and yellow fever transmission (*25*) in Brazil.

While nanopore sequencing has also been extensively used for genome assembly of bacterial pathogens, including tracing outbreaks of important pathogens such as hospital-associated *Salmonella* (*26*) or *Klebsiella pneumoniae (27)*, advanced methods including dated phylogenetic tree reconstruction and estimation of epidemiological population parameters using Bayesian phylodynamic models rely on single nucleotide polymorphism (SNP) resolution of whole genome sequence data (*28–31*). However, nanopore raw read accuracy is at the time of writing insufficient to allow for accurate variant calling from mapped reads. Relatively larger bacterial genomes also require higher throughput for sufficient coverage and operational haploid variant callers, which have not been available or tested specifically for diverse bacterial nanopore data, despite rapid improvements in variant calling of human sequences, when reference genome diversity is limited (e.g. with Google's Deep Variants and phasing pipelines) (*32*, *33*). Applying the same successful phylodynamic methods we have used in real-time surveillance of SARS-CoV-2 to bacterial genome surveillance to estimate critical response parameters - such as the effective reproduction number (R_e) - or conduct surveillance on emerging variants and outbreaks (*18*, *30*, *34–36*), will require improvements in our ability to use nanopore sequencing data for bacterial genomics where it is otherwise too difficult to conduct sequencing operations, including in remote places like the Papua New Guinean highlands and regional or rural hospitals in northern Australia that lack sequencing infrastructure. Sampling from community outbreaks and contextualisation of outbreak data within the wider, known genomic lineage background of the circulating bacterial pathogen is critical to carry methods that have been successfully implemented in healthcare systems into the community. Finally, a challenge that requires careful and long-term operational planning is the availability of lineage-specific samples spanning at least a decade. As bacterial evolution is around two orders of magnitude slower than viral evolution, sufficient temporal and geographical sample coverage has to exist to trace the evolutionary origins and processes in the emergence of bacterial pathogens, especially when compared to the fast paced emergence dynamics operating for

viruses (*37*). Just like viral transmission dynamics methodology is adoptable to bacterial settings (for example by parameterising the Bayesian models with appropriate prior configurations and using sequence data spanning decades of lineage evolution) similar evolutionary and epidemiological processes operate in bacterial disease transmission, including spillovers from animal reservoirs or lineage (variant) evolution driven by genomic changes that interact with host population dynamics, as the lineage emerges and disseminates (*11*). However, in contrast to the more recent viral epidemics, bacterial transmission can proceed silently, in part due to our inability to routinely track important epidemiological parameters rooted in biological evidence (e.g. genome sequences), the long time spans over which evolutionary dynamics operate in bacteria (and the required planning required to obtain samples), modes of transmission generally not observed in viral pathogens (including decade long stable host colonization) as well as absence of monitoring of critical transmission settings, including in disadvantaged and poverty-stricken communities and in rural settings, where lack of access to healthcare and contact with animal populations is more severe than in industrialized settings.

In this work, we address the sparsity of genomic surveillance data available for community-associated *Staphylococcus aureus* transmission in remote Papua New Guinea, where collaborators have tracked an outbreak of severe osteomyelitis in children from the highland towns of Kundiawa and Goroka (*38*). At the same time, we had become aware of similar presentations of community-associated *S. aureus* infections in Far North Queensland, a remote region of northern Australia, encompassing the Cape York Peninsula and the Torres Strait Island, which borders Papua New Guinea (*39*). In the pursuit of data on the transmission dynamics of community-associated *S. aureus* in this region, we find a pattern in the emergence of global community-associated lineages of *S. aureus,* building on previous work of the main authors on the global dissemination of the Indian clone (ST772) (*40*) and the genomic epidemiology and spread of the Australian clone (ST93) (*41*, *42*). Using these data as

reference, our work explores the application bacterial phylodynamic models to estimate outbreak parameters and the use of machine learning models to improve SNP resolution of cost-effective (multiplex) nanopore data of bacterial genomes (*43*), allowing for the application of phylodynamic models to the outbreaks in Far North Queensland and Papua New Guinea. In addition, this work also explores the development of a genomic neighbor typing algorithm based on MinHash techniques (*44–46*), which allows us to rapidly scan outbreak genotypes and infer lineage provenance and antibiotic susceptibility profiles.

In the next paragraphs, an overview of the state of genomic epidemiology of *S. aureus* is provided, followed by a brief summary of advances in nanopore sequencing and phylodynamic inference. In the main body of this work, three comprehensive data chapters describe the research conducted by the main author, and are followed by a summary discussion that draws on the insights gained in the data chapters. **Please note that all Tables and Supplementary Tables or Figures associated with the main chapters are found in the print-edited versions of the preprints in Appendix 1.** Additional published research conducted by the main author and collaborators, some of which relates to this work, but is unrelated to the fulfillment of the degree, can be found in Appendix 2 (*47–52*).

1.1. Genomic epidemiology of *Staphylococcus aureus*

1.2.1. Evolution of natural populations of *Staphylococcus aureus*

Staphylococcus aureus is a prominent bacterial pathogen of humans and animals. Since its discovery from contaminated surgical wounds (Ogston, 1880) it has become one of the most intensely studied bacterial organisms, largely due to its extraordinary impact on public health

and the rise of antibiotic resistant strains (*53*). *S. aureus* is an opportunistic pathogen that causes a variety of clinical manifestations, from superficial skin and soft tissue infections to life-threatening systemic diseases, including osteomyelitis, bloodstream infections, and necrotizing pneumonia (*54*). Patients with compromised immune systems or implanted medical devices are particularly at risk. Despite its eminent role as human pathogen, *S. aureus* is a commensal bacterium, colonizing the anterior nares, nasopharynx and intestines of around 30% of the population, with considerable heterogeneity in carriage duration, ranging from transient (weeks) to permanent colonization (decades) (*55*). It belongs to a group of around 80 other *Staphylococcal* species, most of which are not pathogenic, but interact with *S. aureus* in complex natural communities (*53*). For example, it was recently shown that the emergence of a main hospital associated (pathogenic) sequence type of *S. epidermidis*, a commensal species and often co-occurring with *S. aureus* in the skin microbiome, was associated with the uptake of an accessory genetic element *tarIJLM* from *S. aureus* which led to the expression of a surface-exposed wall teichoic acid that augmented epithelial attachment and host mortality in a mouse sepsis model (*56*). Additionally, the element allowed *S. epidermidis* to transfer bacteriophages from *S. aureus* opening a corridor for genetic exchange of virulence, factors, antibiotic resistance genes and other mobile elements such as the staphylococcal cassette chromosome (SCC*mec*) which usually carries the beta-lactam resistance encoding penicillin-binding protein *mecA* and other mobile genes inserted into these mobile elements (including gene shuttling transposons and epigenetic restriction-modification systems) (*57*).

In the last twenty years, the development of high-throughput whole genome sequencing technology has significantly enhanced our understanding of the evolutionary processes driving the virulence potential and emergence of antibiotic resistant strains in various transmission settings (*53*). Since the publication of the first *S. aureus* genomes by Kuroda and colleagues (*58*, *59*) the field has undergone a remarkable expansion. Public databases now hold tens of

thousands of *S. aureus* genomes, facilitating the analysis of micro-evolutionary processes and integration of epidemiological data to further characterise and understand lineage evolution and emergence. During these initial explorations of the genome architecture of *S. aureus* the research and clinical community gained important insights into the evolutionary drivers that contribute to important clinical characteristics, such as virulence and antibiotic resistance factors. It quickly became clear that *S. aureus* populations evolved predominantly clonally (rate of mutations exceeded rate of recombination) (*60*) and that a significant portion of the genome sequence of these initial strains carried phage genes and mobile elements, including signature elements of *S .aureus* (and other staphylococcal species) such as the SCC*mec* cassettes, but also various prophage associated virulence clusters containing toxins and host immune evasion systems (*61–63*), In addition, antibiotic resistance genes in various completed reference genomes that have been sequenced using long-read sequencing technology (PacBio, ONT) in recent years, have uncovered associations of highly mobile regions of integrated plasmids or transposons, associated with hot-spots of recombination along the genome sequence (including in other mobile element regions, such as SCC*mec* elements) (*64–66*).

Molecular techniques based on PCR amplification of a panel of species-specific “housekeeping” genes (multi locus sequence typing, MLST) developed around the same time as the first whole genomes were sequenced (*67*, *68*), have been adopted into an *in silico* approach for whole genome sequencing data, delineating species diversity by clonal lineages or sequence types of *S. aureus* (STs, analogous to variants of SARS-CoV-2), whose larger relationships can be grouped into clonal complexes (CC) some of which are highly promiscuous and have spawned multiple important lineages (*69*). In the last ten years, the increasing availability of population genomic data using whole genome sequencing resolution has led to a revolution in our understanding of the evolution of natural populations of *S. aureus*, not only those that cause disease in healthcare settings (*53*). With the increasing awareness of *S. aureus* lineages and

genetic or genomic traceability, our understanding of the faceted niches in which the pathogen thrives and that have been under less scrutiny than hospitals over the decades has expanded rapidly, including sequencing the animal-adapted *S. aureus* strains and identification of host-switching events leading to establishment of livestock-associated lineages, including the canonical CC398 strains (*70*). Its versatile niche exploitation, and spillover transmission into human hosts (*71*, *72*), has sparked increasing interest in the role of *S.aureus* colonization in wildlife, their transmission potential into human and livestock and the factors that drive the emergence and persistence of such lineages. A similarly under-surveilled niche outside the traditional hospitals associated with infections of the late 20th century were wider communities and human habitation, often remote or neglected in society, including for example remote Indigenous communities (*42*, *73*, *74*) and socioeconomically disadvantaged human populations, such as on the Indian subcontinent and East Asia (*40*, *75*). In these niches, a different epidemiological type of *S. aureus* lineage evolved, one adapted specifically to transmission in the community setting. In this work, we aim to uncover some of the genomic and epidemiological casualties that led to the rapid emergence and displacement of regional community-associated MRSA.

1.2.2. Global emergence of community-associated MRSA lineages

Since the 1990s, antibiotic-resistant community-associated clones without epidemiological links to the healthcare system have emerged around the world, subsequently replacing other regionally prevailing lineages (*53*, *57*). Community-associated MRSA strains tend to be virulent, infect otherwise healthy people, and are frequently exported from the regions in which they emerged (*53*, *76*). Initially, it was not known whether these strains belonged to a single pandemic lineage similar to healthcare associated clones such as ST239 (*77*, *78*). Molecular work uncovered that these community-associated strains belonged to multiple, regionally

restricted clones, which emerged - seemingly convergently - in the 1990s (*53*, *76*). Distinct regional distributions of community-associated lineages are observed in stark contrast to healthcare-associated strains that tend to spread rapidly in local healthcare systems, often following international dissemination (*77*–*79*). The evolutionary and epidemiological trajectory of community-associated lineages is currently not known, although indications are that the prevalence of some lineages and sublineages have declined over the decade, including the North American USA300 clade (*80*).

While generally considered less resistant to antibiotics than healthcare-associated strains, evidence from multiple global and regional whole-genome datasets has suggested that their emergence is associated with the acquisition of specific resistance mutations and mobile elements (*40*, *42*, *50*, *75*, *80*–*84*). For example, the lineage native to Australia (ST93) acquired a singular SCC*mec*-IV element, coinciding with a lineage expansion on the Australian East Coast (*42*). On the Indian subcontinent, the emergent MSSA clade (ST772-A) acquired an chromosomal integration of a multidrug-resistance plasmid, followed by sporadic acquisition of SCC*mec* in the MSSA population, a change of replacement in the fluoroquinolone resistance mutation in *gyrA* and eventually, stable integration of a smaller SCC*mec* variant (*40*). In some lineages but not all, additional non-synonymous mutations specific to emergent clades and associated with virulence and colonisation factors have been detected (*40*, *83*, *85*). We previously assessed ancestral ST772 strains and compared them phenotypically using biofilm formation, growth rate and toxicity assays, but did not detect any significant differences in these (limited) experiments, possibly suggesting that these mutations may be compensatory in effect, given the large fitness costs associated with resistance acquisition (*40*, *86*). It is worth noting that one defining characteristic in addition to resistance acquisition in the USA300 epidemics (ST8-MRSA) in North and South America, has been the acquisition of the ACME and COMER mobile elements, which have been associated with increased skin survival and colonization

(*85*). Recent studies by Gustave *et al.* have shown that local epidemiological processes, such as pollution from mining activities in Colombia, may have contributed to the spread of strains carrying these elements (*87*). Panton-Valentine leukocidin (PVL) has been another strong contender for a mechanistic explanation of why community-associated lineages emerged, as it occurs frequently in these clones and integrated into the ST8-MSSA background just prior to their emergence in the Americas (*85*), but recent genomic studies have now shown that PVL positive MSSA progenitor strains exist for several community-associated lineages (*40*, *42*, *50*, *81*, *83*), putting in question its central epidemiological role in the emergence of community MRSA clades. Overall, canonical mutations and colonization- or virulence associated elements are unlikely to explain the seemingly convergent emergence of global community-associated lineages, given that not all lineages carry these elements, and given the ubiquitous co-occurrence between resistance acquisition and subsequent emergence (*40*, *42*, *50*, *75*, *80–84*). While the association between emergence and resistance acquisition is strong, a mechanistic explanation has been lacking. Gustave *et al*. recently engineered strains of the ST80 and US300 lineages, and showed that a fitness advantage exists in resistant strains in sub-inhibitory antibiotic media, which may be a driving factor in settings with wide access to antimicrobials or environmental contamination, which may contribute to the spread and maintenance of resistance mutations or mobile elements (*88*).

Epidemiological and genomic evidence for historical and ongoing circulation of MSSA progenitor populations exists for nearly all community-associated lineages of interest (*40*, *42*, *50*, *81*, *83*). Strong contemporary evidence comes from the Australian ST93 lineage, whose ancestral MSSA strains continue to circulate amongst remote communities in the Northern Territory (*42*, *89*). In addition, a symplesiomorphic clade of ST8-MSSA progenitor strains has been found circulating in Africa (*82*), having diverged prior to the emergence of the ancestral ST8-MSSA in Europe during the 19th century, which then spread to the Americas where it diverged into the

ST8-USA300 (MRSA) sublineages in the 20th century (*84*, *85*). Local circulation of progenitor MSSA strains in Romania is documented for the European ST1-MRSA sublineage (*50*, *90*, *91*). Emergence of ST80-MRSA in Europe has epidemiological connections to North West Africa through importation of MSSA cases in French legionnaires (*81*, *92*). While few ancestral strains have been sampled, ST772-MRSA-V is thought to have emerged from local MSSA populations in the Bengal Bay area, with the first isolates from 2004 collected in Bangladesh and India, coinciding with the rise of a multidrug-resistant MRSA clade on the Indian subcontinent (*40*). Even less is known about the origins of the ST59 clone, which produced an MRSA epidemic in Taiwan, but had previously diverged into a (largely) MSSA sister clade in the United States (*75*).

Global dissemination of emergent MRSA clades has frequently been linked to travel and family history in their source region (*40*, *42*, *50*, *75*, *80–84*). For example, nearly 60% of isolates included from a global study on the dissemination of the ST772-MRSA-V clone had family contacts or travel history on the Indian subcontinent (*40*). However, to date, community strains tend to cause small-scale outbreaks, consisting of local transmission chains and household clusters failing to become endemic in the community (*40*, *42*, *82*, *93–95*). Some notable exceptions include several USA300 clades (ST8-MRSA-IV genotype) in Colombia, Gabon and France, as well as the Australian (ST93-MRSA-IV genotype) featuring a transmission event into the Māori and Pacific Islander in metropolitan Auckland, New Zealand (NZ) (*82*, *85*, *87*). Additional evidence for successful recruitment arises from molecular surveillance of ST80 and ST1, as well as from genomic surveillance of ST152-MSSA in the Middle East (*81*, *83*, *91*, *96*).

While these data have contributed to a deeper understanding of community-associated lineage emergence, questions remain about the drivers behind these seemingly convergent events in the late 20th century. Increases in the effective population size (N_e) have been observed in some lineages, coinciding with the acquisition of antibiotic resistance but these analyses have

not been conducted for all relevant sequence types (*50*, *75*, *81*, *83*). While historical and contemporary data on MSSA progenitor populations is limited in most lineages (except ST93), we note that these populations tend to be geographically distinct, and that emergence of resistant genotypes occurs rapidly in industrialized host populations, such as on the Indian subcontinent (ST772), the Australian East Coast (ST93), in central Europe (ST1, ST80) and North America (ST8). In addition, it is not clear whether sustained transmission --- characterized by an effective reproduction number (R_e) exceeding a threshold value of one, and remaining above that threshold for a period of time --- has occurred following the emergence and transmission of community-associated strains, and whether drug-resistant strains are capable of becoming endemic following their exportation. While there has been a strong association between resistance acquisition and emergence, and emerging mechanistic explanations, the epidemiological dynamics in these events require further investigation.

1.2.4. MRSA circulation in Papua New Guinea and Northern Australia

Little is known about the genetic diversity of community-associated *S. aureus* sequence types circulating in Oceania (comprising the loosely defined and overlapping regions of Melanesia, Micronesia, Polynesia and Australasia). In the following, a brief overview of the state of knowledge of *S. aureus* epidemics in the South-West Pacific with a focus on Papua New Guinea. While recent genomic surveillance programs have been instigated in the Philippines (where mostly the traditional Pacific sequence type ST30 was detected) (*97*) data from population centers in Oceania (Solomon Island, Papua New Guinea, Vanuatu, Tonga, Fiji) are sparse to non-existent (*98*); if available, studies have largely been conducted using molecular multi-locus sequence typing (*99–101*). Genome-informed studies demonstrated common nosocomial and proto-community-associated lineages circulating in the region, but samples

were small or restricted to particular host sub-populations (especially children) so that general inference about the transmission of community-associated MRSA remains difficult.

Papua New Guinea (PNG) is of immediate interest to northern Australian biosecurity and transmission across borders in the Torres Strait. Previous work by our collaborators has shown that drug-resistant *Mycobacterium tuberculosis* has sporadically been imported from PNG to Queensland through the region (*102*). Almost nothing is known about the incidence and distribution of *S. aureus* in PNG, which has to some degree been neglected in the face of more urgent and severe disease transmission, such as *Plasmodium spp.* and drug-resistant TB. However, molecular data on *S. aureus* is sparse, featuring two recent studies by the Papua New Guinea Institute of Medical Research (PNGIMR). One nasal colonization study uncovered two different sequence types from molecular work (n = 4), one of which (ST1421) which has representative genomes from Australia available at PubMLST, suggesting potential transmission between the countries (*101*). Another study by Izzard Aglua and colleagues (*38*) has conducted microbiological work on an extensive outbreak of osteomyelitis infections in children in Kundiawa (Simbu Province) in the highlands of PNG. It was determined that the outbreak was caused by MRSA strains, but no further genotyping or molecular work was done. Severe osteomyelitis is still occurring in the pediatric population, and current antibiotic guidelines, recommending first line beta-lactams, appear to be largely ineffective. It is also not clear whether the sampled MRSA isolates belong to a mono- or polyclonal outbreak in Kundiwa. Aglua *et al*. recommended to further characterise the genomic diversity and antimicrobial resistance of *S. aureus* strains circulating in the highland provinces, including transmission in children from the nearby town Goroka, Eastern Highlands Province (~ 14,000 population compared with a population of ~ 8,000 in Kundiawa; travel distance is around 8 hours between the towns by car through the highlands; PNGIMR and the reference laboratory are located in Goroka). This is the extent on molecular knowledge about *S. aureus* lineages in Papua New

Guinea; in this work we will sequence the first *S. aureus* genomes from the country, expanding our understanding of lineage evolution and circulation to one of the last remaining places on earth, where genome-informed data on *S .aureus* is unavailable.

Immediately neighboring Papua New Guinea in the Torres Strait is Far North Queensland, a remote and regional part of Australia - roughly the area of Germany or Italy - with a population of around 280,000 and a population density of 1 person/km^2, less than Mongolia or the Western Sahara. It is home to a large Indigenous population (around 10% of the population, and 25% of the state of Queensland) (*39*). Around half of the population is concentrated in Cairns (~ 150,000 people) which also acts as a regional and international travel hub, with frequent flights to Asia and the Pacific (before 2020). Little is currently known about the circulation of MRSA lineages around Cairns, although reviews of laboratory data from the last decade have suggested that some areas are particularly affected (Cooktown, 72% MRSA positivity rate) (*39*). Anecdotally, MRSA infections in the pathology department at Cairns Hospital are so common that antibiograms are often not even needed to identify drug-resistant strains (*pers. comm*. Dr. Simon Smith). Guthridge *et al.* also estimated increased odds of MRSA infections associated with Indigenous status and geographical association with Cairns (which the authors identify as related to a large Indigenous patient cohort). High rates of MRSA in northern Australian healthcare jurisdictions indicate that Cairns and Hinterland and Cape York and Torres Strait are the only regions in which MRSA rates have been increasing (about 4% per year between 2015 - 2017), contrary to all other jurisdictions in Australia (*103*).

Far North Queensland is the northernmost region in the state of Queensland on the East Coast of Australia, which has its own (as it turns out, mistaken) namesake community-associated MRSA lineages named after the state: the Queensland clone (ST93-MRSA-IV) (*104*). Amongst the circulating nosocomial and community-associated sequence types in Australia, the lineage

has received increased attention due to it's rapid emergence as the dominant community-associated clone in Australia . First discovered in Queensland around 2000 (*104*), the clone has disseminated widely on the East Coast and the rest of the continent; it has been noted for its particular virulence, especially its apparent capacity for immune evasion and colonization (*41*, *42*, *89*) Phylogenetic studies by our group and others have unravelled the origins of the lineage, which - somewhat surprisingly - did not originate in Queensland (*42*). Instead, the lineage appears to be native to remote Indigenous communities in the Northern Territory and Western Australia. Due to extensive work with affected communities by Menzies School of Health Research (Charles Darwin University), and the ongoing problems that this clone is causing in remote communities, extensive genomic data on ancestral MSSA strains from the Northern Territory are available - a rarity amongst datasets of community-associated MRSA. Our analysis concluded that the lineage originated amongst Indigenous communities of the Northern Territory, showing patterns of unusual recombination from unknown ancestral *S. aureus* lineages. In the 1990s, the clone then acquired SCC*mec*-IVa in a subpopulation of MSSA strains, and immediately spread to the Australian East Coast. We showed that from there, the clone rapidly expanded across Queensland and other Australian states, and sporadically transmitted overseas including to the United Kingdom (where most our international isolates were from). Our sample also included isolates from community-transmission in the Pacific Islander and Maori population of Auckland, New Zealand. This is noteworthy for two reasons: first, the epidemiological space in which the lineage was able to spread is reminiscent of Indigenous communities in which the clone originated, including domestic overcrowding, a high burden of skin disease, and difficulties in accessing treatment. Second, it is thought that endemic transmission of community-associated lineages following importation overseas is limited (see above), but the ST93-MRSA-IV strains sampled in New Zealand, formed a genetically distinct clade within the lineage phylogeny, indicating that ongoing (sustained) transmission may have occurred in Auckland. However, we did not find any particular genetic

mutations or mobile elements associated with this unusual clade, leading us to hypothesize that host epidemiological factors (in particular socioeconomic conditions) may have contributed to its successful establishment in the Pacific Islander and Maori community. Furthermore, these results suggested that host populations with similar sociodemographic features in the South-Western Pacific, including in Papua New Guinea and Indigenous communities in Far North Queensland, may be susceptible to outbreaks of the Australian lineage (for the publication on this topic, see Appendix 2).

1.3. Advances in methods for applied genomic epidemiology

1.3.1. Bacterial whole genome sequencing on nanopore platforms

Nanopore sequencing - the concept of using engineered membrane proteins to translocate nucleic acids and measure disturbances in membrane-applied voltage, which is then translated into nucleotide space - has been in development for decades (*17*). Recently, the technology has been commercialized by Oxford Nanopore Technologies (*105*), who developed a massively multiplexed array of engineered nanopores (512 - 2048 active pores on the MinION and PromethION arrays) and rapidly improved the ability to sequence nucleic acid fragments. This was supported largely by an early adoptive research and bioinformatics community, who trialled the portable MinION platform for genome-informed infectious disease surveillance during the Ebola outbreak in West Africa (*23*), and then during the ZIka epidemic in South America (*24*), and more recently on a global-scale, implementing decentralized genomic surveillance of SARS-CoV- (*12*, *20*, *21*). Nanopore sequencing has made remarkable advances in the time it took to complete this thesis (which may perhaps not be a good measure of progress). When we started to sequence on the MinION, R7.3 pore versions were just phased out in favour of the R9 architecture and most of the work here was completed on R9.4.1 pores, with R10.3 pore

architectures featuring double read-heads and achieving Q20 values on raw reads about to be release to the community (as of 2021). Not only the pore chemistry has advanced significantly, but also the implementation of native neural-networks for basecalling of which Chiron written by the Coin group was the first example using hybrid recurrent and convolutional architectures (*106*) and the latest iteration of 1D-Time Channel Separable Convolutions and Connectionist Temporal Classification (CTC) in the latest Bonito and Guppy v0.5.0 base-callers, achieving significant improvements in read accuracy. In addition, pair consensus decoding has just been announced to be implemented widely (with modifications to existing library preparation and adaptive sampling of sequences) and is set to achieve Q30 raw read accuracy, making nanopore technology the only long-read (100 kbp+) high accuracy sequencing technology in current applications.

While these recent advances in the ever-shifting field of nanopore technology are promising the widespread adoption of high-accuracy long-reads (*33*, *107*–*110*), in this work we focused on the reality of deploying existing rapid library preparation and nuclease flushes (R9.4.1 with SQ-RBK004 and EXP-WSH003) to further explore applications in setting like regional hospitals (such as in Townsville) or remote reference laboratories (such as the Papua New Guinea Institute of Medical Research). Bacterial genomes pose a huge challenge in this setting, where we cannot expect optimised DNA extraction or library preparation methods, and where rapid, low-cost solutions are required to cover as many samples as possible to enable genome-informed epidemiology of bacterial outbreaks at population scales. In addition to the protocol setup, it is imperative to keep in mind that while nanopore sequencing has recently enjoyed a huge surge from deployment in global SARS-CoV-2 monitoring efforts, bacterial genomes are much larger and cannot be targeted directly from sample (at the moment) but largely rely on culture of pure isolates for nucleic acid extractions. Bacterial genomes are also significantly larger and require sufficient coverage of the chromosome(s) to accurately call

single nucleotide variants from these data and enable downstream SNP-based applications, such as inference of SNP phenotypes like antibiotic resistance or Bayesian phylodynamic modelling to estimate outbreak epidemiological parameters. Thus, we are faced with a combined challenge of large bacterial genomes that need to be sufficiently multiplexed to allow cost-effective genotype surveillance and detailed reconstruction of phylogenetic and -dynamic processes at population scale, while attempting to do this with a protocol that favours rapid and simple nucleic acid extraction and library preparation, significantly compromising read-length and throughput. While we utilise the streaming capacity for rapid genotype prediction in the later chapters, we do not optimise for real-time sequencing capacity, long read lengths and high throughput that the MinION flow cells are capable off, but rather focus here on the ability to generate a minimum requirement of sequencing data in regional hospitals or reference laboratories (or research centers like the AITHM in Townsville) with simple, highly multiplexed sequencing protocols that reduce cost and increase scalability. We will however compensate for this brute force approach by further developing bioinformatics methods based on machine learning variant polishing (*43*) and genomic neighbor typing (*46*), that are capable of operating low coverage nanopore data. We thus aimed to some degree implement realistic conditions where highly skilled and optimised laboratory workflows are not always possible.

1.3.2. Bayesian phylodynamic modelling of bacterial pathogens

Bayesian phylogenetic - and dynamic modelling of disease transmission, which merges mathematical models, such as compartmental SIR or SEIR models with genomic data in the form of phylogenetic trees, which then inform epidemiological parameters of the model including the effective reproduction number, have become critical to outbreak response (*13*, *34*–*37*). Coalescent and birth-death type models have been used extensively in the current SARS-CoV-2 pandemic, and were crucial in the determination of transmission potential of variants like B1.1.7

(Alpha) and B.1.617 (Delta) (*111*, *112*). Birth-death models consider the dynamics of a population forward in time (unlike coalescent approaches) and in the SIR implementation in BEAST2 are capable of infer the effective reproduction number directly from the transmission rate, becoming uninfectious rate and sampling proportion estimates of the model (*30*, *31*, *113*). Given the Bayesian inference framework, uncertain measurements are inherent to the outputs of the model, and provide a complex framework in which to evaluate whole genome sequencing data to conduct genomic epidemiology (in particular reconstruction of dated trees) and phylodynamic modelling of epidemiological parameters for infection control (*37*).

However, their application in bacterial phylogenetics has been limited. This is largely because the timescales at which bacterial evolutionary processes operate is around ten to a hundred times slower than those of viruses (substitution rates on the order of 1e-06 vs. 1e-04). Therefore, even recently emerged bacterial lineages, such as the community-associated sequence types of which we became aware in the 1990s, require longitudinal sampling collections on the order of decades. While it is possible to delineate fine-scale outbreak evolution (on the order of years, in *S. aureus* this translates to around 2-3 SNPs per year) these historical samples are extremely valuable in reducing uncertainty of estimates closer to the phylogenetic root (such as divergence dates) (*37*, *114–116*). Some sampling collections of community-associated *S. aureus* lineages, such as the ST8-MSSA clade, has samples available from the 1950s (*85*), while other lineages (e.g. ST772) include samples going back to the first discovered isolates (in Bangladesh in 2004), making the divergence and recent emergence challenging to process, as few genetic variations had accumulated in the 8 year timespan until the last samples were collected in 2012. Nevertheless, it is all the more remarkable that these lineages have emerged so rapidly and so recently, with only minor changes in their genome, largely from silent and likely negatively or neutrally selected mutations, but also from the insertion of singular mobile elements like the SCC*mec*-IV. We reason here, that phylodynamic

approaches like the birth-death skyline models, are capable of reconstructing the confluence of genomic changes (acquisition of resistance) and effects on the transmission dynamics (epidemic growth and sustained transmission) of genome-resolved *S. aureus* lineages. While coalescent based approaches have been used in *S. aureus* community-associated lineages showing signatures of population expansion concomitant with the emergent MRSA clades (*81–83*, *88*), the application of birth-death skyline models has so far not been considered. Given that transmission occurs predominantly person-to-person (fomites can be a source of infection, albeit not considered predominant) we can track changes in the effective reproduction number with these models, which would allow us to reconstruct changes in transmission dynamics at the emergence and subsequent dissemination of resistant *S. aureus* clades. We would thus be able to take a closer look at the interface of human epidemiological behaviour and pathogen evolution, and jointly investigate potential genomic and epidemiological drivers behind the dissemination and epidemic growth of drug-resistant community-associated MRSA.

1.4. Summary of data chapters and project aims

In this work, we address the sparsity of genomic data available on the circulation of community-associated *S. aureus* lineages in Far North Queensland and Papua New Guinea, as well as the application of lineage-resolved birth-death skyline models and genomic neighbor typing heuristics for bacterial whole genome sequencing on nanopore platforms. We reasoned that emerging nanopore sequencing technology would be capable of decentralized whole genome sequencing, allowing for genome-informed surveillance of pre-eminent lineages circulating in the region. However, several technical challenges had to be addressed, including the adoption of birth-death skyline models (*30*) for lineage-resolved bacterial datasets (*37*, *116*) and overcoming deficits in single-nucleotide polymorphism (SNP) accuracy and precision (*43*)

by adopting machine learning approaches for transmission inference using low-coverage (and low-cost) nanopore sequencing data .

In the second chapter, we used high-resolution Illumina data of a pediatric outbreak in the highlands of Papua New Guinea as well as a “snap-shot” of strains from Far North Queensland communities, to investigate the provenance and evolutionary origins of regionally prevailing *S. aureus* sequence types (synonymous with ‘lineages’ or ‘clones’). Anecdotally, osteomyelitis and skin and soft tissue infections have been common across the neighboring regions, which led us to hypothesize that transmission between Australia and Papua New Guinea is occuring. We then used these data to probe the transmission dynamics at later stages in the “life-cycle” of community-associated lineages (like the Australian ST93 clone) where resistant MRSA clades transmit overseas and establish sustained transmission. We used birth-death skyline models to investigate changes in the effective reproduction number (indicating epidemic growth and transmission, when $R_e > 1$) across the history of other community-associated lineages, and compared and contrasted these dynamics across the community-associated clones. With this approach, we aimed to determine whether specific changes in the reproductive number have occurred in the emergence of these lineages, and whether these changes were associated with genomic changes or other epidemiological processes including, most notably, the acquisition of antimicrobial resistance coinciding with spread in geographically distinct host populations, such as on the Australian East Coast. Ultimately this chapter aimed to uncover the genomic and epidemiological drivers behind the seemingly convergent emergence of community-associated MRSA in the 1990s and determine the regional transmission dynamics of the Australian lineage in remote and vulnerable host populations in the South-Western Pacific.

In the third chapter, we re-sequenced the outbreaks in Far North Queensland and Papua New Guinea using portable nanopore technology and rapid, low-coverage whole genome sequencing

protocols, intended to multiplex sufficient strains onto a single MinION flow cell for cost-effective per-genome sequencing (targeting < AUD $100 per genome). We first used these data in combination with Illumina reference data to design machine learning classifiers that are capable of polishing false SNP calls made by nanopore variant callers. This increased accuracy and precision of the variant calls for outbreak reconstruction within the wider lineage background of the Australian ST93 clone. Hybrid alignments from known lineage-wide population backgrounds and nanopore outbreaks were used to estimate outbreak effective reproduction numbers by applying birth-death skylines models in BEAST2. In the fourth chapter, we then adopted genomic neighbor typing, a heuristic inference of whole genome associated pheno- and genotypes, to conduct outbreak surveillance on the nanopore data, assessing performance against a full species-wide reference database. We tested this approach on various multiplex setups, including on site at the Papua New Guinea Institute of Medical Research in Goroka, on Flongle adapters and by sequencing 48 strains on a single MinION flow cell. As a consequence, we improved bacterial outbreak surveillance with a simple mass-multiplex sequencing protocol for portable sequencing devices with minimal computational resource requirements, allowing for its application in low and middle income countries. Ultimately, these chapters aimed to improve our understanding and application of genomic technologies to infer transmission dynamics and genome evolution of community-associated MRSA in remote populations of northern Australia and Papua New Guinea.

2. Data chapters

2.1. Phylodynamic signatures in the emergence of community-associated MRSA

Eike Steinig[1,2,*], Izzard Aglua[3], Sebastián Duchêne[1], Michael T. Meehan[2], Mition Yoannes[4], Cadhla Firth[2], Jan Jaworski[3], Jimmy Drekore[5], Bohu Urakoko[3], Harry Poka[3], Clive Wurr[6], Eri Ebos[6], David Nangen[6], Elke Müller[7,8], Peter Mulvey[2], Charlene Jackson[9], Anita Blomfeldt[10], Hege Vangstein Aamot[10], Moses Laman[4], Laurens Manning[11,12], Megan Earls[13], David C. Coleman[13], Andrew Greenhill[4,14], Rebecca Ford[4], Marc Stegger[15], Muhammed Ali Syed[16], Bushra Jamil[17], Stefan Monecke[7,18], Ralf Ehricht[7,19], Simon Smith[20], William Pomat[4], Paul Horwood[4,21], Steven Y.C. Tong[1,22,+], Emma McBryde[2,+]

[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]Sir Joseph Nombri Memorial-Kundiawa General Hospital, Kundiawa, Simbu Province, Papua New Guinea, [4]Papua New Guinea Institute of Medical Research, Goroka, Eastern Highlands Province, Papua New Guinea, [5]Simbu Children's Foundation, Kundiawa, Simbu Province, Papua New Guinea, [6]Goroka General Hospital, Surgical Department, Goroka, Eastern Highlands Province, [7]Leibniz Institute of Photonic Technology (IPHT), Jena, Germany, [8]InfectoGnostics Research Campus, Jena, Germany, [9]U.S. National Poultry Research Center, Agricultural Research Service, USDA, Athens, United States, [10]Department of Microbiology and Infection Control, Akershus University Hospital, Lørenskog, Norway, [11]Department of Infectious Diseases, Fiona Stanley Hospital, Murdoch, Western Australia, [12]Medical School, University of Western Australia, Harry Perkins Research Institute, Fiona Stanley Hospital, Murdoch, Western Australia, [13]Microbiology Research Unit, Division of Oral Biosciences, University of Dublin, Trinity College, Dublin, Ireland, [14]Department of Microbiology, Federation University Australia, Ballarat, Australia, [15]Department of Bacteria, Parasites and Fungi, Statens Serum Institut, Copenhagen, Denmark, [16]Department of Microbiology, University of Haripur, Haripur, Pakistan, [17]BJ Micro Lab (SMC Private) Limited, Islamabad, Pakistan, [18]Technical University of Dresden, Dresden, Germany, [19]Institute of Physical Chemistry, Friedrich -Schiller University, Jena, Germany, [20]Cairns Hospital and Hinterland Health Service, Queensland Health, Cairns, Australia, [21]College of Public Health, Medical & Veterinary Sciences, James Cook University, Townsville, Australia, [22]Victorian Infectious Diseases Service, The Royal Melbourne Hospital at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia

Keywords: Community associated MRSA | ST93 | Far North Queensland | Papua New Guinea | Genomic epidemiology | Reproduction number | Birth death skyline | Phylodynamics | Illumina | ONT | Nanopore

* Corresponding authors: eike.steinig@unimelb.edu.au, steven.tong@mh.org.au ; + authors contributed equally

Abstract

Community-associated, methicillin-resistant *Staphylococcus aureus* (MRSA) lineages have emerged in many geographically distinct regions around the world during the past 30 years. Here, we apply consistent phylodynamic methods across multiple community-associated MRSA

lineages to describe and contrast their patterns of emergence and dissemination. We generated whole genome sequencing data for the Australian sequence type (ST) 93-MRSA-IV from remote communities in Far North Queensland and Papua New Guinea, and the Bengal Bay ST772-MRSA-V clone from metropolitan communities in Pakistan. Increases in the effective reproduction number (R_e) and sustained transmission ($R_e > 1$) coincided with spread of progenitor methicillin-susceptible *S. aureus* (MSSA) in remote northern Australia, dissemination of the ST93-MRSA-IV genotype into population centers on the Australian East Coast, and subsequent importation into the highlands of Papua New Guinea and Far North Queensland. Analysis of a ST772-MRSA-V cluster in Pakistan suggests that sustained transmission in the community following importation of resistant genotypes may be more common than previously thought. Applying the same phylodynamic methods to existing lineage datasets, we identified common signatures of epidemic growth in the emergence and epidemiological trajectory of community-associated *S. aureus* lineages from America, Asia, Australasia and Europe. Surges in R_e were observed at the divergence of antibiotic resistant strains, coinciding with their establishment in regional population centers. Epidemic growth was also observed amongst drug-resistant MSSA clades in Africa and northern Australia. Our data suggest that the emergence of community-associated MRSA and MSSA lineages in the late 20th century was driven by a combination of antibiotic resistant genotypes and host epidemiology, leading to abrupt changes in lineage-wide transmission dynamics and sustained transmission in regional population centers.

2.1.1. Introduction

Since the 1990s, antibiotic-resistant community-associated clones without epidemiological links to the healthcare system have emerged around the world, subsequently replacing other regionally prevailing lineages (*53*, *57*). Community-associated MRSA strains tend to be virulent, infect otherwise healthy people, and are frequently exported from the regions in which they

emerged (*53*, *76*). Initially, it was not known whether these strains belonged to a single pandemic lineage similar to healthcare associated clones such as ST239 (*77*, *78*). Molecular work uncovered that these community-associated strains belonged to multiple, regionally restricted clones, which emerged - seemingly convergently - in the 1990s (*53*, *76*). Distinct regional distributions of community-associated lineages are observed in stark contrast to healthcare-associated strains that tend to spread rapidly in local healthcare systems, often following international dissemination (*77*–*79*). The evolutionary and epidemiological trajectory of community-associated lineages is currently not known, although indications are that the prevalence of some lineages and sublineages have declined over the decade, including the North American USA300 clade (*80*).

While generally considered less resistant to antibiotics than healthcare-associated strains, evidence from multiple global and regional whole-genome datasets has suggested that their emergence is associated with the acquisition of specific resistance mutations and mobile elements (*40*, *42*, *50*, *75*, *80*–*84*). For example, the lineage native to Australia (ST93) acquired a singular SCC*mec*-IV element, coinciding with a lineage expansion on the Australian East Coast (*42*). On the Indian subcontinent, the emergent MSSA clade (ST772-A) acquired an chromosomal integration of a multidrug-resistance plasmid, followed by sporadic acquisition of SCC*mec* in the MSSA population, a change of replacement in the fluoroquinolone resistance mutation in *gyrA* and eventually, stable integration of a smaller SCC*mec* variant (*40*). In some lineages but not all, additional non-synonymous mutations specific to emergent clades and associated with virulence and colonisation factors have been detected (*40*, *83*, *85*). We previously assessed ancestral ST772 strains and compared them phenotypically using biofilm formation, growth rate and toxicity assays, but did not detect any significant differences in these (limited) experiments, possibly suggesting that these mutations may be compensatory in effect, given the large fitness costs associated with resistance acquisition (*40*, *86*). It is worth noting

that one defining characteristic in addition to resistance acquisition in the USA300 epidemics (ST8-MRSA) in North and South America, has been the acquisition of the ACME and COMER mobile elements, which have been associated with increased skin survival and colonization (*85*). Recent studies by Gustave *et al.* have shown that local epidemiological processes, such as pollution from mining activities in Colombia, may have contributed to the spread of strains carrying these elements (*87*). Panton-Valentine leukocidin (PVL) has been another strong contender for a mechanistic explanation of why community-associated lineages emerged, as it occurs frequently in these clones and integrated into the ST8-MSSA background just prior to their emergence in the Americas (*85*), but recent genomic studies have now shown that PVL positive MSSA progenitor strains exist for several community-associated lineages (*40*, *42*, *50*, *81*, *83*), putting in question its central epidemiological role in the emergence of community MRSA clades. Overall, canonical mutations and colonization- or virulence associated elements are unlikely to explain the seemingly convergent emergence of global community-associated lineages, given that not all lineages carry these elements, and given the ubiquitous co-occurrence between resistance acquisition and subsequent emergence (*40*, *42*, *50*, *75*, *80*–*84*). While the association between emergence and resistance acquisition is strong, a mechanistic explanation has been lacking. Gustave *et al*. recently engineered strains of the ST80 and US300 lineages, and showed that a fitness advantage exists in resistant strains in sub-inhibitory antibiotic media, which may be a driving factor in settings with wide access to antimicrobials or environmental contamination, which may contribute to the spread and maintenance of resistance mutations or mobile elements (*88*).

Epidemiological and genomic evidence for historical and ongoing circulation of MSSA progenitor populations exists for nearly all community-associated lineages of interest (*40*, *42*, *50*, *81*, *83*). Strong contemporary evidence comes from the Australian ST93 lineage, whose ancestral MSSA strains continue to circulate amongst remote communities in the Northern Territory (*42*,

89). In addition, a symplesiomorphic clade of ST8-MSSA progenitor strains has been found circulating in Africa (*82*), having diverged prior to the emergence of the ancestral ST8-MSSA in Europe during the 19th century, which then spread to the Americas where it diverged into the ST8-USA300 (MRSA) sublineages in the 20th century (*84*, *85*). Local circulation of progenitor MSSA strains in Romania is documented for the European ST1-MRSA sublineage (*50*, *90*, *91*). Emergence of ST80-MRSA in Europe has epidemiological connections to North West Africa through importation of MSSA cases in French legionnaires (*81*, *92*). While few ancestral strains have been sampled, ST772-MRSA-V is thought to have emerged from local MSSA populations in the Bengal Bay area, with the first isolates from 2004 collected in Bangladesh and India, coinciding with the rise of a multidrug-resistant MRSA clade on the Indian subcontinent (*40*). Even less is known about the origins of the ST59 clone, which produced an MRSA epidemic in Taiwan, but had previously diverged into a (largely) MSSA sister clade in the United States (*75*).

Global dissemination of emergent MRSA clades has frequently been linked to travel and family history in their source region (*40*, *42*, *50*, *75*, *80*–*84*). For example, nearly 60% of isolates included from a global study on the dissemination of the ST772-MRSA-V clone had family contacts or travel history on the Indian subcontinent (*40*). However, to date, community strains tend to cause small-scale outbreaks, consisting of local transmission chains and household clusters failing to become endemic in the community (*40*, *42*, *82*, *93*–*95*). Some notable exceptions include several USA300 clades (ST8-MRSA-IV genotype) in Colombia, Gabon and France, as well as the Australian (ST93-MRSA-IV genotype) featuring a transmission event into the Māori and Pacific Islander in metropolitan Auckland, New Zealand (NZ) (*82*, *85*, *87*). Additional evidence for successful recruitment arises from molecular surveillance of ST80 and ST1, as well as from genomic surveillance of ST152-MSSA in the Middle East (*81*, *83*, *91*, *96*).

While these data have contributed to a deeper understanding of community-associated lineage emergence, questions remain about the drivers behind these seemingly convergent events in the late 20th century. Increases in the effective population size (N_e) have been observed in some lineages, coinciding with the acquisition of antibiotic resistance but these analyses have not been conducted for all relevant sequence types (*50*, *75*, *81*, *83*). While historical and contemporary data on MSSA progenitor populations is limited in most lineages (except ST93), we note that these populations tend to be geographically distinct, and that emergence of resistant genotypes occurs rapidly in industrialized host populations, such as on the Indian subcontinent (ST772), the Australian East Coast (ST93), in central Europe (ST1, ST80) and North America (ST8). In addition, it is not clear whether sustained transmission --- characterized by an effective reproduction number (R_e) exceeding a threshold value of one, and remaining above that threshold for a period of time --- has occurred following the emergence and transmission of community-associated strains, and whether drug-resistant strains are capable of becoming endemic following their exportation. While there has been a strong association between resistance acquisition and emergence, and emerging mechanistic explanations, the epidemiological dynamics in these events require further investigation.

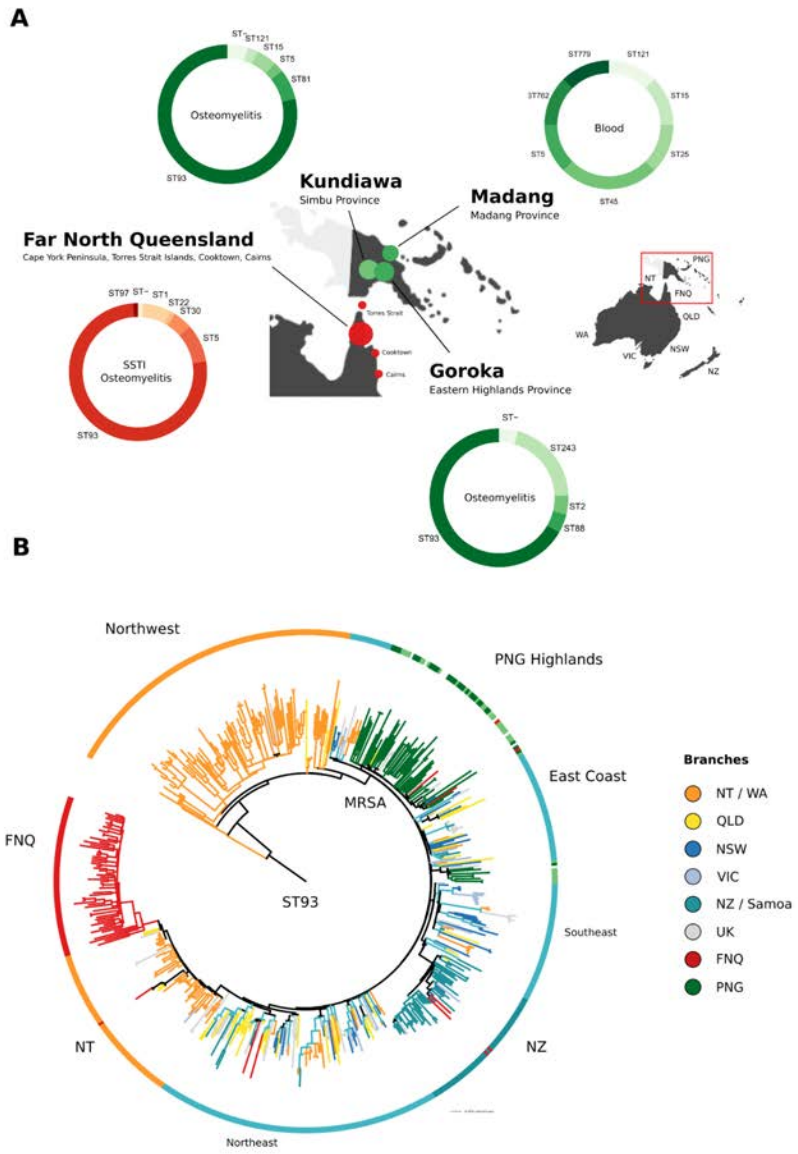

Fig. 1: Genomic epidemiology of *Staphylococcus aureus* outbreak isolates from Papua New Guinea (n = 95) and Far North Queensland (n = 89). (**A)** Map of sampling locations, multilocus sequence types and predominant symptoms of patients (ring annotation) (**B**) Global evolutionary history of the Australian lineage (ST93) showing the rooted maximum-likelihood phylogeny constructed from a non-recombinant core-genome SNP alignment (n = 575) and major regional geographical structure in the evolutionary history of the clone (branch colors). ST93 emerged in remote communities of North West Australia and acquired SCC*mec*-IV, spreading to the Australian East Coast (blues, yellow), remote northern Australian communities (orange, red), the remote highlands of Papua New Guinea (green) and into Auckland communities in New Zealand (seagreen).

2.1.2. Results

We sequenced 187 putative *S. aureus* isolates from remote PNG (2012 - 2018) and FNQ (Torres and Cape / Cairns and Hinterland jurisdictions, collected in 2019) using Illumina short-reads (Fig. 1, Online Supplementary Tables). Genotyping identified the Australian MRSA clone (ST93-MRSA-IV) as the main cause of paediatric osteomyelitis in the highland towns of Kundiawa and Goroka ($n_{Kundiawa}$ = 33/42, n_{Goroka} = 30/35). The remaining isolates from osteomyelitis cases in Kundiawa and Goroka belonged to an assortment of sequence types (ST5, ST25, ST88), single locus variants of ST1247 (n = 1) and ST93 (n = 2), coagulase-negative staphylococci including *S. lugdunensis* (n = 1), *S. delphini* (n = 1) and *Mammaliicoccus sciuri* (n = 1), as well as a neonatal hospital cluster of invasive ST243 (clonal complex 30, n = 9) (Fig. 1A). FNQ isolates sampled in 2019 were largely identified as ST93-MRSA-IV (n_{FNQ} = 68/91) on a background of various other lineages, including one infection with *S. argenteus* (Fig. 1A, Online Supplementary Tables).

ST93-MRSA-IV strains from PNG and FNQ were contextualised within the global sequence diversity of the lineage (ST93, n = 444) to determine strain provenance using a maximum-likelihood (ML) phylogeny constructed from non-recombinant core-genome SNPs (Fig. 1B, n = 575, 6648 SNPs). The resulting tree topology recapitulated our previous analysis of the lineage, confirming its origin from extant MSSA strains circulating in remote Indigenous communities of north-western Australia (*42*). The main divergence event of ST93-MRSA on the Australian East Coast (AEC) coincided with acquisition of SCCmec-IV (Fig. 1B). Isolates from PNG formed a major (n = 55) and minor (n = 8) clade in the ML tree, consisting of mixed strains from Goroka and Kundiawa (Figs. 1A, 1B, green). The major clade contained sporadic isolates sampled in Queensland, FNQ, and New South Wales (n = 3) indicating regional transmission from PNG (Fig. 1B). The FNQ cluster derived from a Northern Territory (NT) clade that itself

appears to have been a re-introduction of ST93-MRSA-IV from Australia's East Coast into the Northern Territory (Fig. 1B). Sporadic isolates sampled in FNQ were imported from other locations, including the North Eastern ST93-MRSA-IV circulation, the NT, as well as NZ and PNG (red branches outside of FNQ cluster in Fig.1B). Sporadic transmission into FNQ most likely occurred through Cairns, which is the regional hub of the area, has an international airport and is frequented by visitors from the region.

Regional transmission dynamics of the ST93-MRSA-IV

We next used fast maximum-likelihood methods (*117*) (Fig. S1, Fig. S2) as well as Bayesian coalescent skyline (*31*) and birth-death skyline (*30*) models for serially (PNG) and contemporaneously sampled isolates (FNQ) in BEAST2 (*113*) to infer time-scaled phylogenies and estimate epidemiological parameters for the ST93-MRSA-IV clone, including changes in R_e and effective population size (N_e) over time (Fig. 2A, Table 1). Previous genomic studies have noted increases in N_e in the emergence of several community lineages (*50*, *75*, *81*, *83*) but data was not available for all lineages (*40*, *42*, *82*) and no studies had previously used birth-death skyline models to investigate changes in R_e. Lineage-wide transmission dynamics of the Australian clone ST93 indicate successive surges in R_e at the divergence of extant MSSA strains in the Northern Territory (NT), at acquisition of SCCmec-IVa and spread on the AEC, and upon recruitment into PNG, NZ and FNQ communities (Fig. 2A). The clone became epidemic ($R_e > 1$) soon after the emergence of an extant MSSA clade in the NT (MRCA = 1990, 95% credible interval, CI: 1988 - 1992), coinciding with the first sample (1991) from the NT in our retrospective collection (Fig. 2A). When the clone was first described in southern Queensland in 2000 (*104*) a resistant clade ST93-MRSA-IV had just established transmission in East Coast population centers (QLD, NSW, VIC) following the acquisition of SCCmec-IV around 1994 (95% CI: 1993 - 1995) (Fig. 1B, Fig. 2C). We estimate that the introduction of ST93-MRSA-IV into PNG occurred in the early 2000s (MRCA = 2000, 95% CI: 1998 – 2003, Fig. 2A) soon after

establishment on the East Coast of Australia. In contrast, introduction of ST93-MRSA-IV into FNQ occurred more recently (MRCA = 2007, 95% CI: 2005 - 2009).

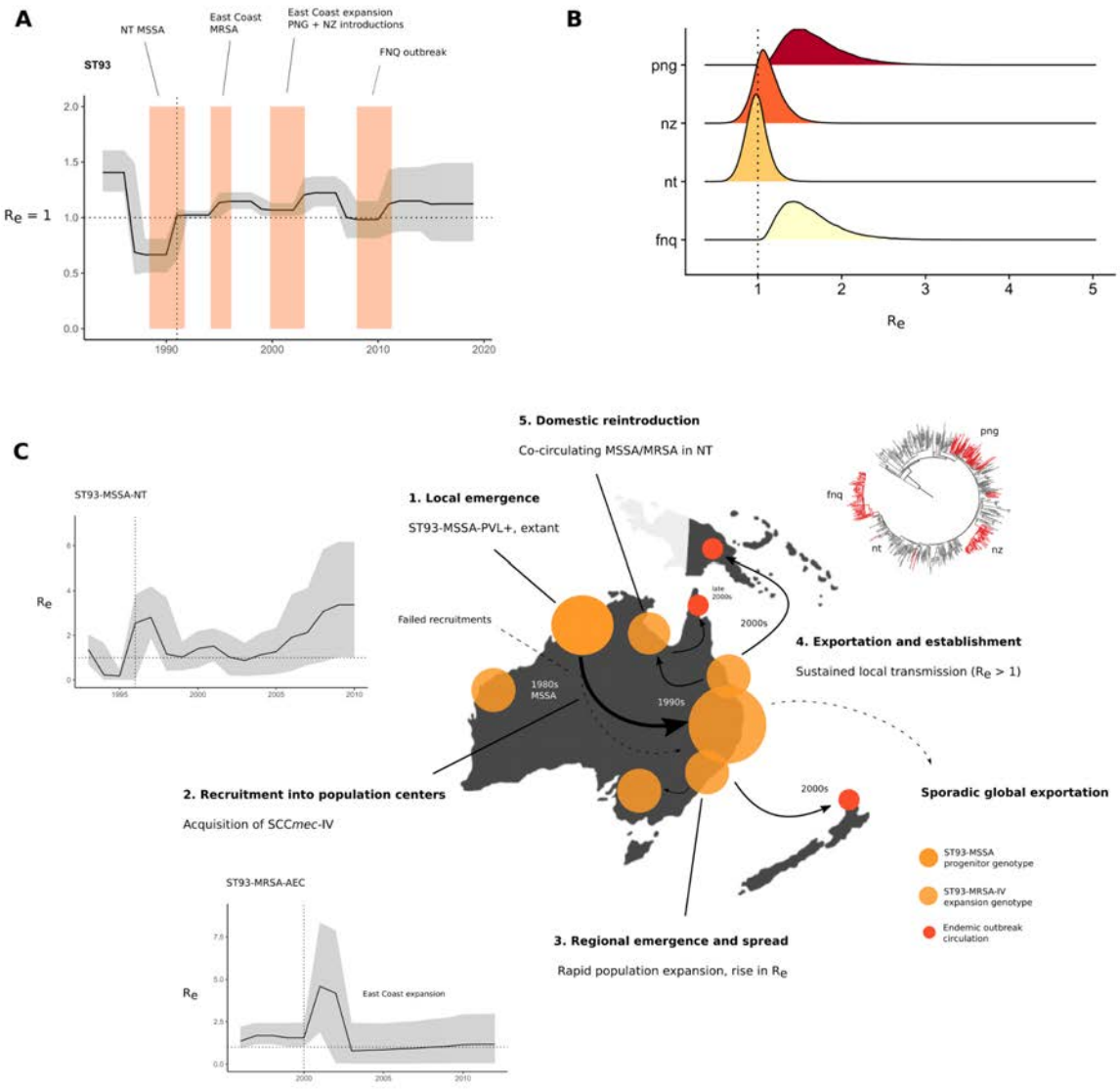

Fig. 2: Phylodynamic signatures and parameter estimates for the Queensland clone (ST93) **(A)** Changes in the effective reproduction number (R_e) over time, showing the 95% credible interval (CI) intervals of the MRCA of clade divergence events of ST93-MSSA and -MRSA (colors) using the birth-death skyline model **(B)** R_e posterior density distributions for introductions in Papua New Guinea (PNG), Far North Queensland (FNQ), New Zealand (NZ) and re-introduction of the MRSA genotype to the Northern Territory (NT) **(C)** Events in the emergence and regional dissemination of the Queensland clone, with maximum-likelihood phylogenies and branch colors indicating major sub-clades and divergence events in the emergence of ST93. Vertical lines in skyline plots indicate the year of first sample from the lineage or clade, horizontal lines indicate the epidemic threshold of R_e = 1.

Birth-death skyline models with fixed lineage-wide substitution rates were additionally applied to regional sublineages and -clades of ST93 (Methods), including the introductions into PNG, FNQ, New Zealand (NZ), and the re-introduction into the Northern Territory (Fig. 2B, Table 1: sublineages). We observed sustained transmission in PNG (R_e = 1.61, 95% CI: 1.13 – 2.40) and FNQ (R_e = 1.55, 95% CI: 1.08 – 2.44). Sustained transmission may have occurred in the Northern Territory re-introduction of the MRSA-IV genotype (R_e = 0.97, 95% CI: 0.72 – 1.25) and the Auckland community cluster (R_e = 1.09, 95% CI: 0.79 – 1.48). Infectious periods ($1/\delta$), the time from acquisition to death / sampling of the strain, were estimated on the order of several years for strains from NZ (3.72 years, 95% CI: 1.12 - 6.10), NT (2.29 years, 95% CI: 0.97 – 3.41), PNG (2.21 years, 95% CI: 0.49 – 5.08) and FNQ (1.06 years, 95% CI: 0.19 - 2.58) (Table S1, Fig. S1). We note that birth-death skyline models assume well-mixed populations and that the comparatively low, lineage-wide median estimate of the infectious period in ST93 (0.427 years, 95% CI: 0.244 – 0.63, Table S1, Fig. S3) was likely a result of applying the model over the early diverging MSSA clade and the MRSA clades on the AEC, where the resulting population structure biased the parameter estimate for the infection period. As we had sufficient sample sizes for these subclades (n_{NT} = 96, n_{AEC} = 278), we applied the birth-death skyline to each clade individually, allowing us to model clade-specific changes in R_e over time (Fig. 2C, insets) as well as using the clade-specific method with fixed lineage-wide clock-rates (Table 1). This produced estimates for the duration of the infectious period consistent with the credible intervals of the outbreak subclades (NT MSSA 1.79 years, 95% CI: 0.77 – 3.76; AEC 1.38 years, 95% CI: 0.47 - 4.86) (Table S1). Stable circulation ($R_e \approx 1$) on the Australian East Coast was observed following a notable spike in R_e shortly after acquisition of SCCmec-IV in the MRCA of the clade (Fig. 2C). In contrast, R_e of ST93-MSSA in the NT has been increasing since around 2003, with credible intervals of $R_e > 1$ suggesting sustained transmission until at least 2011. More recent genomic data on the spread of ST93-MSSA was not available.

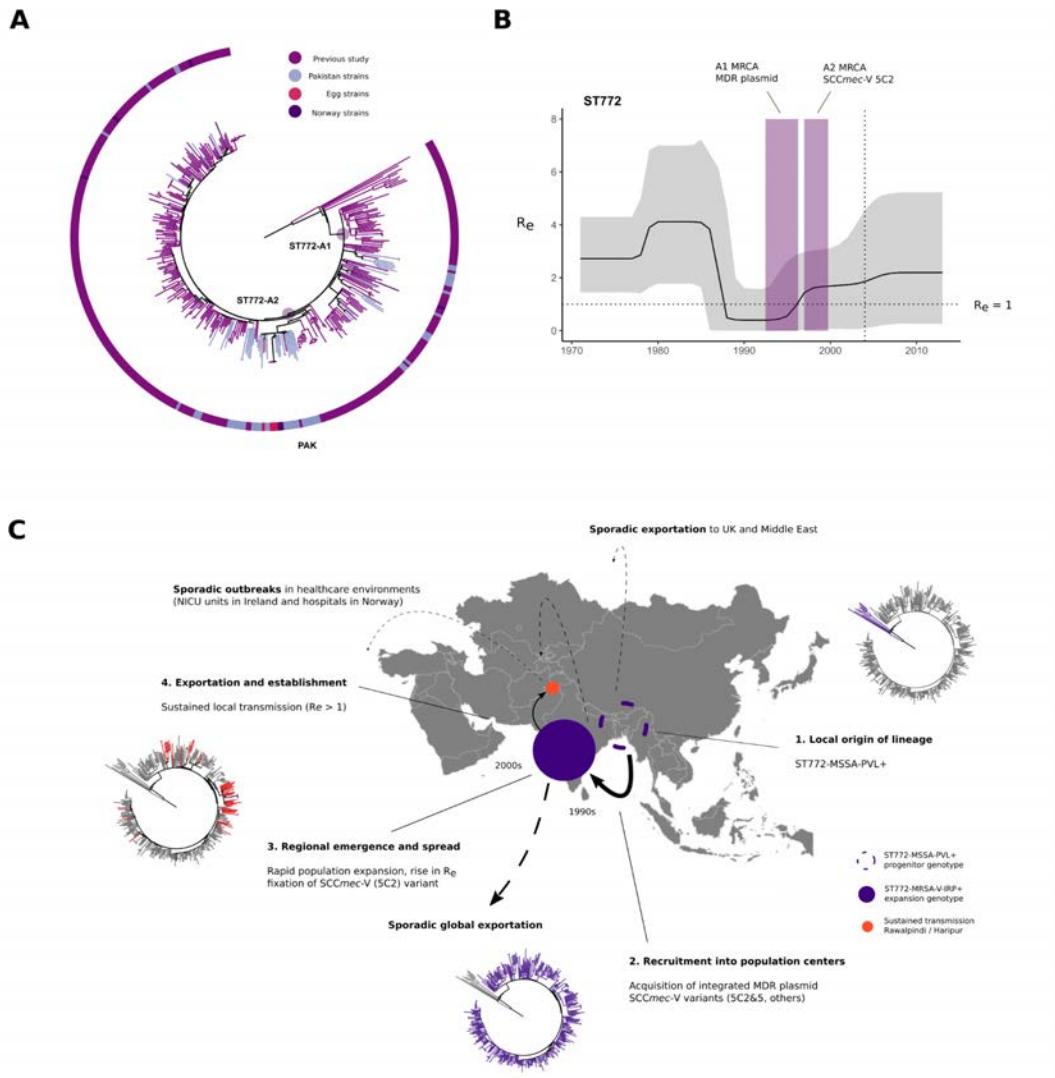

Fig. 2: Genomic epidemiology of the Bengal Bay clone ST772 on the Indian subcontinent. **(A)** Rooted maximum-likelihood phylogeny of ST772 showing new strains (n = 59) from community transmission in Haripur and Rawalpindi (Islamabad metropolitan area). Sporadic importation into Pakistan is evident from singular and small transmission clusters, including a larger community transmission cluster in Rawalpindi (n = 25, PAK), where table-eggs were associated with the community outbreak and indicated additional spread overseas **(B)** Effective reproduction number (R_e) over time; acquisition of the MDR integrated plasmid (MRCA 95% CI colored) is associated lineage-wide epidemic spread (R_e > 1). **(C)** Events in the emergence of drug resistant ST772-MRSA on the Indian subcontinent; branch color in maximum likelihood phylogenies show major subclades and ongoing transmission in Pakistan.

Sustained community transmission of ST772-MRSA-V in Pakistan

We next investigated whether clade-specific signatures of epidemic growth (R_e > 1) could be found in other community-associated MRSA lineages. We had previously reconstructed the detailed (n = 355) evolutionary history of the ST772-MRSA-V clone (*40*), which acquired multiple resistance elements, and emerged in the last two decades on the Indian subcontinent, where it has a become a dominant community-associated lineage (*57*). No other genomic samples were available from these countries with the exception of unreleased ST772-A samples from India (*118*) and a macaque-associated environmental MRSA isolate from Nepal (*119*). We sequenced an additional 59 strains of ST772 from community and hospital sources in the population centers of Rawalpindi and Haripur in Pakistan (*120*), as well as some strains imported into a University hospital in Norway (*93*) (Fig. 3). We found that ST772 was exported into Pakistan on multiple occasions from the background population on the Indian subcontinent (Fig. 3); our sample contained several smaller transmission clusters (n < 8) in line with observations of community spread following international transmission (*40*, *93*) (Fig. 3A). In addition, a larger transmission cluster (n = 25) was established shortly after fixation of SCC*mec*-V (5C2) (2002, 95% CI: 2000 - 2003) in the emergent clade ST772-A2 (Fig. 3A). Application of the birth-death skyline model on the lineage revealed changes in effective reproduction numbers similar to those observed in ST93-MRSA-IV (Fig. 3B). Instead of several pronounced spikes of the reproduction number, its epidemic phase was characterized by a monotonic rise in R_e coinciding with the acquisition of a multidrug resistance-encoding integrated plasmid (*blaZ-aphA3-msrA-mphC-bcrAB*) around 1995 (95% CI: 1992 - 1996). Following a switch in fluoroquinolone resistance mutations in *gyrA* and fixation of the SCC*mec*-V (5C2) variant shortly after its emergence on the Indian subcontinent (1998, 95% CI: 1996 - 1999), a smaller increase in the reproduction number occurred with a delay of several years (Fig. 3B). Estimates for R_e in the Pakistan cluster suggest that importation resulted in

sustained transmission (R_e = 1.36, 95% CI: 1.08 - 1.79). Mirroring the emergence of drug-resistant ST93-MRSA on the Australian East Coast, the Bengal Bay clone emerged in population centers on the Indian subcontinent from a currently unknown MSSA progenitor population and was able to establish sustained transmission after importation into Pakistan (Fig. 3C).

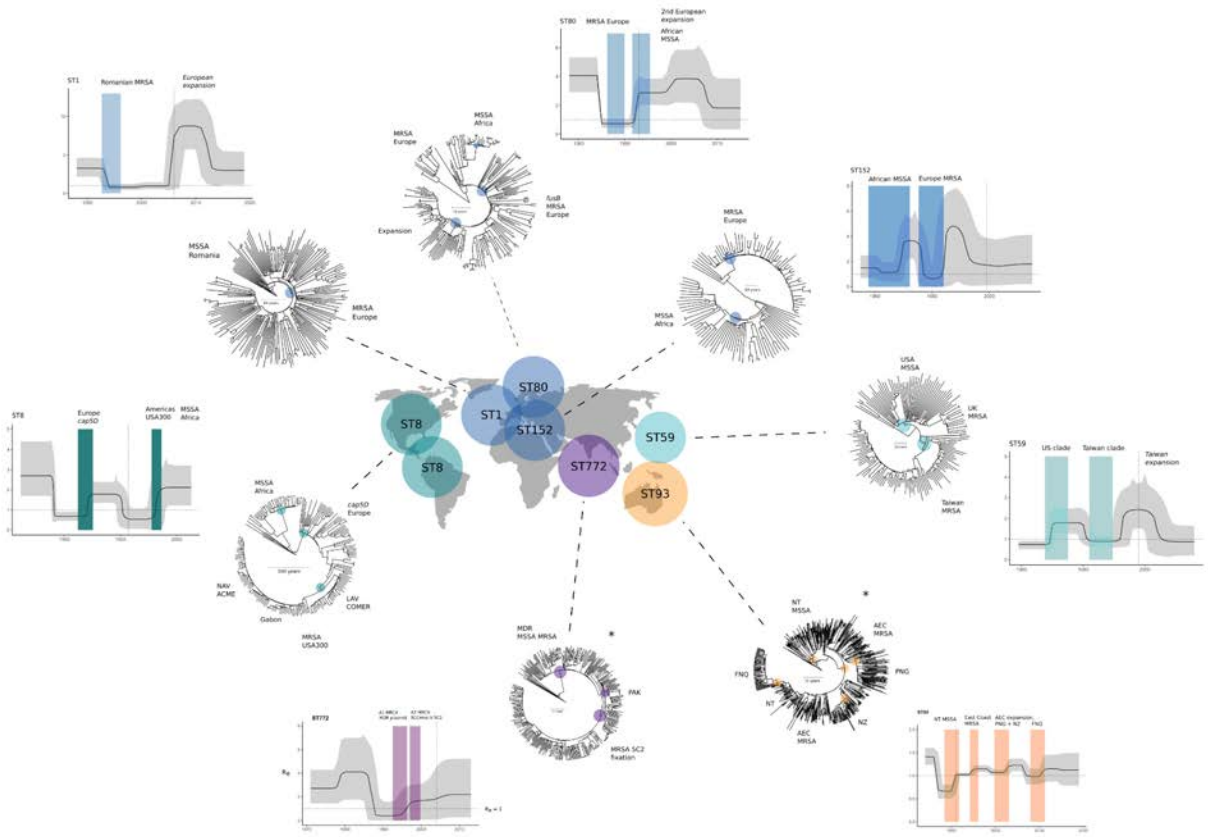

Fig. 2: Bayesian phylogenetic trees and changes in the effective reproduction number (R_e) of global community-associated *Staphylococcus aureus* lineages. The map shows lineages with sufficient (n > 100) genotype-resolved data included in this study, their main geographical regions in which they emerged, and their respective Bayesian maximum clade credibility trees, with colored dots indicating important clade and outbreak divergences (internal nodes). Associated changes in R_e estimated over equally sliced intervals over the sampling period in the birth-death skyline model are shown next to the trees. Trajectories of R_e show the median posterior estimates over time (dark line) and their 95% CI intervals (grey). Colored rectangles in the skyline plots correspond to 95% CI intervals of the clade MRCA indicated by dots in the lineage trees.

Global emergence of community-associated *Staphylococcus aureus*

We next applied the birth-death skyline model to other community-associated MRSA clones, accounting for major lineages that have become dominant community lineages regionally and for which lineage-resolved genomic data were available (n > 100, Fig. 4) (*40*, *42*, *50*, *75*, *80–84*). Short-read sequence data with dates and locations from previous genomic lineage analyses were collected from studies published on the emergence of ST1 (n = 190), ST152 (n = 139) and ST80 (n = 217) in Europe, the US-Taiwan clone ST59 (n = 154) and the European-American ST8 (n = 210, excluding isolates available as assemblies only). Multiple sequence types (ST152, ST8, ST80) included extant MSSA populations circulating in Africa (Online Supplementary Tables, Table 1). Our analysis confirmed signatures of epidemic growth across these lineages, including notable increases in R_e following genomic changes and recruitment into regional host populations (Central Europe, North America, Australian East Coast, India, Taiwan), as well as increases in N_e (effective population sizes of *S. aureus* lineages) noted in previous investigations, coinciding with increases of R_e (Fig. 4, Fig. S1). MRCAs of antibiotic resistant clades in all MSSA and MRSA sublineages were estimated with 95% CI lower bounds between 1972 - 2005, and upper bounds between 1978 - 2009, confirming the seemingly convergent global emergence of resistant community strains in the late 20th century (Table 1). Low estimated sampling proportions suggest that ST8 and ST93 are widespread, consistent with global and regional epidemiological data of these clones; there was less certainty in the predictions for the recently emerged ST1, ST772 and for ST152 (Table 1). High posterior estimates of sampling proportion (σ) in ST80 further suggest that the lineage is in decline in European host populations, although reports indicate potential ongoing circulation in the Middle East (*121*). Overall, median infectious periods varied between lineages with the shortest estimates of several months for ST93 and the longest estimates exceeding ten years in several lineages (Table S1).

Considerable changes in R_e occurred in the ancestral ST8-MSSA genotype at the emergence in European populations in the 19th century, which has been associated with the capsule mutation *cap5D* (*82*) (1860, 95% CI: 1849 - 1871) (Fig. 4). The proto community-associated clone then spread to the Americas and acquired SCC*mec*-IV variants as well as the canonical COMER and ACME elements at the divergence of two regionally distinct epidemics across North America (*84*) and parts of South America (*85*, *88*) which are notable as a combined increase in R_e in the second half of the 20th century (Fig. 4). While data was sparse for ST59-MSSA strains (*75*), elevated reproduction numbers indicate that it became epidemic in the United States in the 1970s and 1980s, followed by the emergence of a resistance-enriched MRSA clade in Taiwan in the late 1970s and its expansion in the 1990s with a delay between the MRCA of resistant strains and the epidemic in Taiwan several years later (Fig. 4, Table 1). Similar delays occurred in European clones ST80-MRSA and ST1-MRSA, which also shared high estimates for their infectious periods (> 10 years). We suspect that ancestral strains circulated for several years in local subpopulations before their emergence across Europe in the 1990s (ST80) (*81*) and 2000s (ST1) (*50*) but a weak temporal signal may contribute to a high degree of uncertainty in ST1 (Fig. S4, 95% CI intervals). Similar to the minor increase in R_e of ST772 after acquisition of the SCC*mec*-V (5C2) variant, a second increase in R_e without notable genomic changes was observed in ST80, suggesting a second shift in transmission dynamics as the MRSA genotype spread across Europe in the early 2000s (Fig. 4). We observed the steepest spikes of reproduction numbers in clones recruiting into European countries (ST1, ST152) where R_e temporarily spiked to > 5 - 8 after initial recruitment into the host population, albeit with large confidence intervals that nevertheless indicate at least epidemic growth in more recent times (Fig. 4). R_e estimates of West African subclades indicated epidemic spread in symplesiomorphic MSSA (ST8-MSSA R_e = 1.54, 95% CI: 1.09 - 2.26; ST152-MSSA R_e = 1.55, 95% CI: 1.10 – 2.25) and an introduction of USA300 (MRSA) in Gabon (ST8-MRSA-Gabon, R_e

= 1.58, 95% CI: 1.11 - 2.34). However, these MSSA clades had acquired mild beta-lactam and other antibiotic resistance before their regional spread, including notable enrichment of *blaZ, dfrG, tetK* and *fosD* in ST8-MSSA, *blaZ* in ST152-MSSA, *blaZ* and *ermC* in ST93-MSSA, as well as occasional acquisition or loss of SCC*mec* in most MSSA clades and sublineages (Fig. S8).

2.1.3. Discussion

In this study, we found a pattern in the emergence of community clones (n_{total} = 1843) associated with the acquisition of antibiotic resistance determinants, which coincide with changes in host-pathogen transmission dynamics (increases in R_e) and lineage population size expansions (N_e) upon recruitment into regional population centers in the 1990s. Increases in R_e exceeding the epidemic threshold ($R_e > 1$) were closely associated with the acquisition of resistance in community-associated MSSA circulating in specific host subpopulations. AMR acquisition was followed with the emergence and sustained transmission of resistant clades in regional population centers, such as the Australian East Coast (ST93), Taiwan (ST59), the Indian subcontinent (ST772) and central Europe (ST1, ST80, ST152). We hypothesize that resistance acquisition enables niche transitions into host populations with distinct socioeconomic structure and population densities, particularly those in urban or industrialized settings. These “spillover” events show patterns in R_e reminiscent of pathogens recruiting into susceptible host populations, where spikes in the effective reproduction numbers are followed by establishment of sustained transmission ($R_e > 1$) or elimination ($R_e < 1$). Sharp increases in R_e of emergent lineages were found concomitant with the MRCA of clades that had obtained AMR elements or mutations (Fig. S8). Fixation of resistance determinants and leveling of R_e following these spikes suggests that epidemiological factors such as widespread antibiotic use in the community, improved access to healthcare services and treatment, public health responses, or environmental antimicrobial contamination, may constitute a new adaptive landscape for the

emerging drug-resistant clade, contributing to its successful dissemination or elimination. This is observed in the persistence of the epidemics ($R_e > 1$) over decades following the initial spikes in emergence, which often coincided with the first available samples of these lineages (Fig. 4, vertical lines in subplots)

Estimates of R_e are susceptible to a number of demographic factors which we could not explicitly model, including access to treatment, host population contact density, and changes in age-specific mixing patterns and others. Signatures in R_e over time inferred from these genomic data therefore combine demographic and epidemiological factors linked to genomic changes and geographical strain attribution in the jointly inferred phylogenies. Our data suggest that the acquisition of multiple locality-specific resistance mutations and mobile genetic elements has driven rapid genotype expansions with notable increases in R_e observed across lineage-wide and clade-specific analyses. SCC*mec*-elements of type IV and V eventually integrated into resistant clades, but cassette genotypes are variable and usually preceded or supplemented by other resistance determinants. For example, the stepwise acquisition of resistance in ST772-MSSA occurred first through the chromosomal integration of a multidrug-resistance plasmid, followed by a shift in the *gyrA* mutation conferring fluoroquinolone resistance and eventual fixation of the short 5C2 variant of SCC*mec*-V (*40*), whereas other lineages such as ST93-MRSA-IV emerged after a singular acquisition event of SCC*mec*-IV(*42*). It is notable that even the successful and sampled MSSA clades in Africa and northern Australia were enriched in resistance determinants, with *blaZ* (ST93-MSSA, ST8-MSSA, ST152-MSSA) and *tetK* (ST8-MSSA) amongst others (Fig. S8).

In support of this "AMR spillover" hypothesis of the emergence of community-associated MRSA strains, Gustave and colleagues (*88*) demonstrated in competition experiments with ST8 (USA300) and ST80 genotypes, that antibiotic-resistant strains expressed a fitness advantage

over wild-type strains on subinhibitory antibiotic media. Presence of low-level antibiotic pressure may therefore be a crucial epidemiological driver in the emergence of resistant clades in host populations that have widespread access to treatment or may not practice effective antibiotic stewardship. However, antibiotic resistance is likely not the only driver for local clade emergence and dissemination. Gustave and colleagues (*87*) recently showed that the mercury-resistance operon located on the COMER element may have driven the dissemination of the USA300 variant in South America on a background of pollution from mining activities. Further data to investigate local competitive fitness dynamics under population antibiotic pressure backgrounds *in vivo* or *in vitro* is required. Complex genotype competition dynamics may arise from environmental coupling at different time-points in the evolution of a lineage, and differential competitive fitness in evolutionary and epidemiological landscapes may play a role in why some community strains successfully recruit into host populations following exportation and fail to become endemic elsewhere.

Local epidemiological patterns in the ST93 lineage phylogeny revealed co-circulating MSSA and MRSA genotypes in the Northern Territory, where a potentially sustained (R_e = 0.97, 95% CI: 0.72 - 1.25) re-introduction of the MRSA genotype eventually spread into communities across Far North Queensland. We further observed that sustained transmission is occurring in symplesiomorphic MSSA populations of ST8 and ST152 in Africa (*82*, *83*), as well as extant ST93-MSSA in northern Australia, particularly amongst Indigenous communities (*42*, *74*, *122*). MSSA clades thus have established sustained transmission in host populations, preceding MRSA clade recruitment into geographically distinct populations (ST8-USA300, ST93-MRSA-IV, ST152-MRSA). It is notable that epidemic signatures were found for resistant MSSA (Northern Territory, Africa) and MRSA (FNQ, PNG, NZ) clades in socioeconomic settings similar to those experienced by many remote Indigenous communities in Australia. These include high burdens

of skin-disease, domestic overcrowding, and poor access to healthcare or other public services (*42*, *74*, *122–124*).

Non-synonymous mutations in factors associated with immune response and skin colonization at the divergence of epidemic MRSA and MSSA have been detected previously in ST772, and ACME and COMER elements in the USA300 clades are implicated in transmission and persistence phenotypes (*85*, *125*, *126*), but it is unclear to what degree these changes have contributed to the emergence, transmission, persistence, or fitness of resistant strains in the presence of other strains or genotypes. Given that acquisition of antibiotic resistance determinants and recruitment into population centers coincides with rapid increases in R_e across all lineages examined in this study, these factors are not likely to explain the rapid change in transmission dynamics we estimated at the divergence of resistant clades (Fig. S8). However, mutations may contribute to ongoing persistence in host populations before and after emergence, or constitute pre-adaptations that support successful transmission in new host populations. Canonical mutations have previously been detected at the divergence of ST772-A and have been associated with colonization ability, which may play a role in compensating for the fitness cost induced by resistance acquisition (*40*). Preliminary phenotypic data from the Bengal Bay clone suggests that there was no significant difference in biofilm formation or growth rate between MSSA and multidrug-resistant MRSA strains (*40*). Further rigorous experiments will need to be conducted to better understand the significance of these mutations in strains preceding the emergence of resistant clades and their interaction with resistance phenotypes.

Our sampling design and models used for the inference of phylodynamic parameters have important limitations. First, we note that uncertainty deriving from incomplete lineage sampling is large, but mitigated by using published collections of metadata-complete and lineage-representative genomes. However, there was a lack of data on ancestral MSSA strains,

a problem pointed out explicitly for ST80 (*81*) and ST59 (*75*), but also relevant to ST772. These effects appeared less severe for ST8, for which there was a wide sampling range going back to 1953 (*82*), and for ST93 (*42*, *89*), which had well-represented MSSA collections from the Northern Territory, and for which the MRCA and origin of the lineage were estimated to have occurred within a year (Fig. S3). For our phylodynamic comparison we addressed sampling bias towards the present by allowing piecewise changes in the sampling proportion consistent with sampling effort for each lineage.

Our study provides phylogenetic and population genomic evidence that community-associated genotypes have emerged in regional host populations following the acquisition of antibiotic resistance. Pre-adaptations for transmission in the ancestral host populations may have contributed to the eventual, epidemic spread of resistant strains in populations with access to antibiotic treatment and healthcare services. However, well-sampled ancestral MSSA genomes are lacking for important community-associated lineages including ST772, ST59 and ST80; deeper sampling and ongoing genome-informed surveillance of these populations will be required to further understand the processes that allow lineages to emerge and become epidemic. It is notable that the seemingly convergent emergence events in the second half of the 20th century --- whose signatures we detect from phylodynamic models across all sequence types --- were caused by epidemics of both, drug-resistant MSSA and MRSA. While antibiotic resistance has facilitated emergence in regional population centers, local evolutionary and host population dynamics have also played a role in the emergence of important community-associated clades, including the potential role of environmental pollution and the COMER element in the dissemination of USA300 in South America (*87*).

Despite the limitations of this study, the phylodynamic estimates observed across sequence types are remarkably consistent with decades of molecular and epidemiological work that has

characterised the global emergence and spread of community-associated S. aureus lineages, including their notable emergence in the 1990s and subsequent establishment across their respective geographical distributions. We provide genomic evidence for sustained transmission of MRSA strains in Pakistan and PNG, as well as for drug-resistant MSSA strains in African countries and northern Australia, where - in addition to the notable enrichment of resistance without stable acquisition of SCC*mec* - sociodemographic host-factors may play an under-appreciated role in the transmission dynamics and epidemic potential of these lineages. Ongoing epidemic transmission of ST93-MSSA is of concern for Indigenous communities in the Northern Territory and there is now evidence for the dissemination of its emergent MRSA genotype beyond the Australian continent. Wider circulation of ST93-MRSA-IV in Papua New Guinea is likely. Our work underlines the importance of considering remote and disadvantaged populations in a domestic and international context. Social and public health inequalities (*73*, *127*) appear to facilitate the emergence and circulation of community-associated pathogens including drug-resistant *S. aureus*.

2.1.4. Materials and Methods

Outbreak sampling and sequencing. We collected isolates from outbreaks in two remote populations in northern Australia and Papua New Guinea (Fig. 1). Isolates associated with paediatric 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 neighbouring Eastern Highlands province town of Goroka. Ultimately these strains were available to us, because the outbreak was identified as unusual and there were concerns that this might be a particularly virulent strain. We supplemented the data with MSSA isolates associated with severe hospital-associated infections and blood cultures in Madang (Madang Province) (n = 8) and Goroka (n = 12). Isolates from communities in Far North Queensland, including metropolitan Cairns, the Cape York Peninsula and the Torres Strait Islands (n = 91),

were a contemporary sample from 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 LB broth at 37°C overnight and stored at -80°C in glycosol and LB. Samples were regrown prior to sequencing, and a single colony was placed into in-house lysis buffer and sequenced at the Doherty Applied Microbial Genomics laboratory using 100 bp paired-end libraries on Illumina HiSeq. Illumina short-read reads from the global lineages included in this study were collected from the European Nucleotide Archive (Online Supplementary Tables).

Genome assembly and variant calling. Illumina data was adapter- and quality- trimmed with *Fastp (128)* before *de novo* assembly with *Shovill* (https://github.com/tseemann/shovill) using downsampling to 100x genome coverage and the *Skesa* assembler (*129*). Assemblies were genotyped with *SCCion* (https://github.com/esteinig/sccion), a wrapper for common tools used in *S. aureus* genotyping from reads or assemblies. These include multilocus sequence (MLST), resistance and virulence factor typing with *mlst* and *abricate* (https://github.com/tseemann) using the ResFinder and VFDB databases (*130*). SCC*mec* types were called using the best *Mash* (*44*) match of the assembled genome against a sketch of the SCC*mec*Finder database (*131*) and confirmed with *mecA* gene typing from the ResFinder database. Antibiotic resistance to twelve common antibiotics were typed with (*132*); this strategy was used for all lineage genomes to confirm or supplement antibiotic resistance determinants as presented in original publications. Strains belonging to ST93 (FNQ, PNG) and ST772 (Pakistan) were extracted and combined with available sequence data from previous studies on the ST772 (n_{total} = 359) and the ST93 (n_{total} = 575) lineages. *Snippy* v.4.6.0 (https://github.com/tseemann/snippy) was used to call core-genome SNPs against the ST93 (6,648 SNPs) or ST772 (7,246 SNPs) reference

genomes JKD6159 (*65*) and DAR4145 (*64*). Alignments were purged of recombination with *Gubbins (133)* using a maximum of five iterations and the GTR+G model in *RAxML-NG (117)*. Quality control, assembly, genotyping, variant calling and maximum likelihood (ML) tree construction, statistical phylodynamic reconstruction and exploratory Bayesian analyses were implemented in Nextflow (*14*) for reproducibility of the workflows (https://github.com/np-core/phybeast). Phylogenetic trees and metadata were visualized with Interactive Tree of Life . All program versions are fixed in the container images used for analysis in this manuscript (Data Availability).

Maximum-likelihood phylogenetics and -dynamics. We used a ML approach with *TreeTime* v0.7.1 (*134*) to obtain a time-scaled phylogenetic tree by fitting a strict molecular clock to the data (using sampling dates in years throughout). Accuracy for an equivalent statistical approach using least-squares dating (*135*) (LSD) is similar to that obtained using more sophisticated Bayesian approaches with the advantage of being computationally less demanding (*114*). As input, we used the phylogenetic tree inferred using ML in *RAXML-NG* after removing recombination with *Gubbins*. The molecular clock was calibrated using the year of sample collection (i.e. heterochronous data) with least-squares optimization to find the root, while accounting for shared ancestry (covariation) and obtaining uncertainty around node ages and evolutionary rates. We also estimated the ML piecewise (skyline) coalescent on the tree using default settings, which provides a baseline estimate of the change in effective population size (N_e) over time. Temporal structure of the data was assessed by conducting a regression of the root-to-tip distances of the ML tree as a function of sampling time and a date-randomisation test on the TreeTime estimates with 100 replicates (*115*, *136*) (Fig. S1). All trees were visualized in Interactive Tree of Life (ITOL) and node-specific divergence dates extracted in Icytree (*137*).

Bayesian phylodynamics and prior configurations. We used the Bayesian coalescent skyline model to estimate changes in effective population size (N_e), and implemented the birth-death skyline from the *bdsky* package (https://github.com/laduplessis/bdskytools) for *BEAST* v2.6 to estimate changes in the effective reproduction number (R_e) (*30*, *31*, *113*). 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 ($\lambda \div \delta$).

For lineage-wide analysis we used all available samples from each sequence type, including MSSA and MRSA clades, but compared model estimates from the entire lineage to distinct clade subsets to mitigate the effect of population structure from well-sampled clones like ST93 (Table 1, Fig. 2B: R_e estimates for ST93-MSSA and -MRSA). Median parameter estimates with 95% CI intervals and Markov chain traces were inspected in *Tracer* to assess convergence (*138*). We implemented Python utility functions to generate XML files and configure priors more conveniently in a standardized form implemented in the *NanoPath* package (https://github.com/esteinig/nanopath). Plots were constructed with scripts (https://github.com/esteinig/nanopath) that use the *bdskytools* package including computation of the 95% highest posterior density (HPD) median intervals (credible intervals, CI) using the Chen and Shao algorithm implemented in the boa package for R (*139*). We chose to present median posterior intervals, since some posterior distributions had long tails in the prior distributions of some parameters (e.g. origin or become uninfectious rates, Fig. S3). Icytree was used to inspect Bayesian maximum clade credibility trees derived from the posterior sample of trees. In the coalescent skyline model for each lineage, we ran exploratory chains of 200 million iterations varying the number of equidistant intervals over the tree height (dimensions, d) specified for the

priors describing the population and estimated interval (dimension) size ($d \in \{2, 4, 8, 16\}$, Fig. S2). As posterior distributions were largely congruent (top rows, Fig. S2), we selected a sufficient number of intervals to model changes in effective population size (N_e) of each lineage over time ($d = 4$ and $d = 8$).

In the birth-death skyline models, priors across lineages were configured as follows: we used a $Gamma(2.0, 4.0)$ prior for the time of origin parameter (T), covering the last hundred years and longer. We chose a $Gamma(2.0, 2.0)$ prior for the reproductive number, covering a range of possible values observed for *S. aureus* sequence types in different settings (*140–143*) which may have occurred over the course of lineage evolution. We configured the reproduction number prior (R_e) to a number of equally sized intervals over the tree; a suitable interval number was selected by running exploratory models for each lineage with 100 million iterations from $d \in \{1 - 10\}$ followed by a comparison of parameters estimates under these configurations (occurence of stable posterior distributions, absence of bi or multi-modal posteriors) (not shown, available in data repository). Because sequence type-specific becoming-non-infectious rates in community-associated *S. aureus* are not well known, either from long-term carriage studies or phylogenetic reconstructions (*94*, *144–146*), we explored a range of prior configurations for the becoming uninfectious rate parameter (δ) including a flat uniform prior $Uniform(1.0, 1.0)$ and a $LogNormal(\mu, 1.0)$prior with $\mu = 0.1$ (10 years infectious period), $\mu = 0.2$ (5 years) and $\mu = 1.0$ (1 year). We chose a $LogNormal(1.0, 1.0)$ prior, as the resulting parameters estimates were coherent (Fig. S5). Lineage-wide sensitivity analysis showed that estimates across lineages were not driven by the prior (Fig. S6). Sampling proportion (ρ) was fixed to zero in the interval ranging from the origin to the first sample (pre-sample period); the remaining time until present (sampling period) was estimated under a flat $Beta(1.0, 1.0)$ prior, accounting for sampling bias towards the present as well as largely unknown estimates of global sampling

proportions across lineages. Final lineage models were run with 500 million iterations on GPUs with the *BEAGLE* library (*147*) under a $GTR + G$ substitution model with four rate categories. We used a strict molecular clock with a $LogNormal(0.0003,\ 0.3)$ rate prior in real space as all lineages 'evolved measurably' (Fig. S1). Models were run until chains were mixed and ESS values reached at least 200, as confirmed in *Tracer*.

Lastly, we ran birth-death skyline models on specific clades within the lineage phylogenies, including the ancestral and symplesiomorphic MSSA populations, the USA300 sublineages, and importations of ST93, ST772 and ST8 (Fig. 4, Fig. S8). For each subset of strains, we extracted the core-genome variant alignment subset, configured the reproduction number prior to a single estimate over the clade (since in outbreak datasets the number of sequences per clade was smaller than 100 and the sampling interval smaller than 10 years) (Fig. S8). Because temporal signal is lost in the clade subsets, we fixed the substitution rate to the lineage-wide estimate in all runs for 100 million iterations for the MCMC with trees sampled every 1000 steps. Runs were quality controlled by assuring that chains mixed and ESS values for all posterior estimates reached at least 200. Since sufficient samples and a wide sampling interval were available to track R_e changes in ST93-MSSA (n = 116 in the Northern Territory) and -MRSA clades (n = 278, Australian East Coast) over time (Fig. 2, inset plots), we explored a stable configuration of the reproductive number prior across equally-spaced intervals ($d \in \{5 - 10\}$) with 200 million iterations of the MCMC (not shown, available in data repository). We explored $Gamma(2.0,\ \theta)$ distributions where $\theta \in \{0.5,\ 1.0,\ 1.5,\ 2.0\}$ in the R_e prior of sublineages and outbreaks. This was to guard against bias towards inferring sustained transmission (R_e > 1) in outbreak models with limited data and temporal signal from subset alignments (representative examples: Fig. S7). Results from the model runs under the conservative$Gamma(2.0,\ 0.5)$ prior assert only minor differences compared to higher configurations (consistently R_e > 1) and the conservative

estimates are presented here (Table 1). We also conducted a sensitivity analysis for all sublineage models where we ran the models under the prior only (Fig. S8, note that under-the-prior posteriors are still influenced by dates, albeit not the genetic data). With the exception of ST93-NT and ST93-NZ we asserted that estimates of R_e were not driven by the prior and dates alone in all sublineages and outbreaks. Full explorative data can be found in the data repository for this study.

2.2. Phylodynamic modelling of bacterial outbreaks using nanopore sequencing

Eike Steinig[1,2,*], Sebastián Duchêne[1], Izzard Aglua[3], Andrew Greenhill[4,5], 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], Moses Laman[4], Laurens Manning[8,9], Cadhla Firth[2], Simon Smith[10], William Pomat[4], Lachlan Coin[1], Steven Y.C. Tong[1,11], Emma McBryde[2,*], Paul Horwood[4,12,*]

[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]Sir Joseph Nombri Memorial-Kundiawa General Hospital, Kundiawa, Simbu Province, Papua New Guinea, [4]Papua New Guinea Institute of Medical Research, Goroka, Eastern Highlands Province, Papua New Guinea, [5]Simbu Children's Foundation, Kundiawa, Simbu Province, Papua New Guinea, [6]Goroka General Hospital, Surgical Department, Goroka, Eastern Highlands Province, [7]Department of Infectious Diseases, Fiona Stanley Hospital, Murdoch, Western Australia, [8]Medical School, University of Western Australia, Harry Perkins Research Institute, Fiona Stanley Hospital, Murdoch, Western Australia, [10]Cairns Hospital and Hinterland Health Service, Queensland Health, Cairns, Australia, [11]College of Public Health, Medical & Veterinary Sciences, James Cook University, Townsville, Australia, [12]Victorian Infectious Diseases Service, The Royal Melbourne Hospital at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia

Keywords: Community associated MRSA | ST93 | Far North Queensland | Papua New Guinea | Genomic epidemiology | Reproduction number | Birth death skyline | Phylodynamics | Illumina | ONT | Nanopore

* Corresponding authors: eike.steinig@unimelb.edu.au; + authors contributed equally

Abstract

Nanopore sequencing and phylodynamic modelling 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 remote communities without access to sequencing infrastructure. Here, we implement Random Forest models for single nucleotide polymorphism (SNP) polishing to estimate divergence and effective reproduction numbers (R_e) of two community-associated, methicillin-resistant *Staphylococcus aureus* (MRSA) outbreaks in remote Far North Queensland and Papua New Guinea (n = 159). Successive barcoded panels of *S. aureus* isolates (2 x 12 per MinION) sequenced at

low-coverage (> 5x - 10x) provided sufficient data to accurately infer assembly genotypes with high recall when compared with Illumina references. *De novo* SNP calling with *Clair* was followed by SNP polishing using intra- and inter-species models trained on *Snippy* reference calls. Models achieved sufficient resolution on ST93 outbreak sequence types (> 70 - 90% accuracy and precision) for phylodynamic modelling from lineage-wide hybrid alignments and birth-death skyline models in *BEAST2*. Our method reproduced phylogenetic topology, geographical source of the outbreaks, and indications of sustained transmission (R_e > 1). We provide Nextflow pipelines that implement SNP polisher training, evaluation, and outbreak alignments, enabling reconstruction of within-lineage transmission dynamics for infection control of bacterial disease outbreaks using nanopore sequencing.

2.2.1. Introduction

Sequence data from infectious disease outbreaks has provided critical information for infection control and inference of pathogen transmission dynamics, for example during the West African Ebola virus epidemic (*23*) and the current SARS-CoV-2 pandemic (*148*). Maximum-likelihood (ML) and Bayesian phylodynamic methods are commonly used to date the emergence of lineages and outbreaks, and to estimate key epidemiological parameters, such as changes in the effective reproduction number over time (R_e) and the most recent common ancestor (MRCA) of an outbreak (*35*, *36*, *111*). Oxford Nanopore Technology (ONT) sequencing has emerged as viable technology for real-time genomic epidemiology, and has been applied across large-scale SARS-CoV-2 surveillance efforts in the United Kingdom and Denmark amongst others (*12*, *112*, *149*, *150*). Moreover, nanopore sequencing devices can be operated in low and middle-income countries where local genomics infrastructure may be lacking or is difficult to access (*24*, *151*), so that a timely outbreak response is not feasible (*11*). This 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 hosts. 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 (*148*, *152*). However, nanopore sequencing for bacterial pathogens, coupled to Bayesian phylodynamic models, have so far not been considered, mainly due the need for sufficiently accurate single nucleotide polymorphism (SNP) calling at bacterial whole genome scales (*37*). SNP calls from high coverage (> 30x) Illumina data is the current standard for accurate SNP calls used in phylogenetic applications, but current generation nanopore SNP calling has suffered from low sequence read accuracy (R9.4.1) and a heavy focus on variant calling in human genomes, with much of the available callers developed specifically for human variants (*32*, *33*). This problem is further aggravated when attempting to sequence cost-effectively, e.g. using low-coverage multiplexed runs (< 5-10x) and simple library preparation with ONT sequencing kits (R9.4.1 pore architecture, SQK-RBK-004 libraries) that can be used in low and middle income countries with large burdens of bacterial disease.

Phylodynamic inference on nanopore platforms is further complicated when (ideally) using an outbreak reference genome 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 (*153*). In addition, on bacterial time-scales (years) little sequence variation will have occured in newly emergent outbreaks, which places an disproportionate emphasis on correctly inferring the few available outbreak-specific polymorphisms. 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 SNP calls, complicating transmission inference from ONT

sequence data, where accuracy and precision are required (*43*). 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 (*114*, *115*, *154*). Due to slower substitution rates in bacteria compared to viruses (*116*), 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, represents a significant barrier to the implementation of phylodynamic modelling 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 e.g. *Neisseria gonorrhoeae* (*43*, *155*) or between species from environmental sources (*156*). Recently a six-strain multiplex protocol for the MinION with genome assembly and determination of phylogenetic relationships to identify outbreaks has been tested for *S. aureus* lineages sampled in Norway. However, it remains unclear whether full within-lineage phylodynamic modelling 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.* (*43*) 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 (*32*). We use *Snippy* Illumina variant profiles as reference data and investigate caller performance across reference genomes and outbreak datasets. We show that Random Forest classifiers sufficiently remove incorrect calls from *Clair* in outbreak isolates with >5x coverage to allow for sequencing of 24

community-associated *S. aureus* isolates per MinION flow cell (n = 181) successfully resolving phylodynamic parameters of two outbreaks of ST93-MRSA-IV in remote Far North Queensland (FNQ) and Papua New Guinea (PNG).

2.2.2. Results

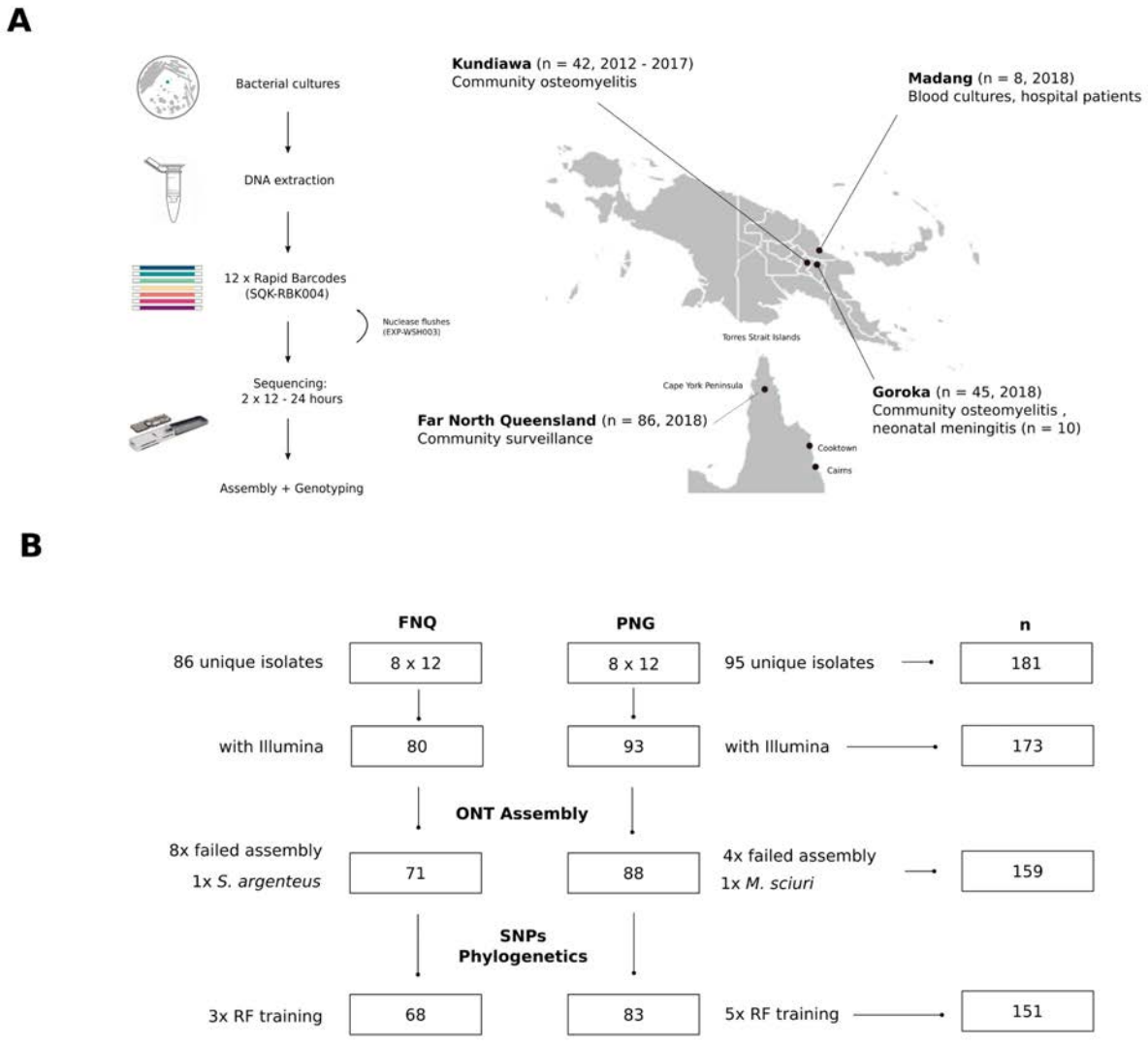

Fig. 1: Sequencing workflow and outbreak sampling locations in northern Australia and Papua New Guinea. **(A)** Isolates were sequenced on 8 flowcells with 24 isolates per flowcell using a sequential nuclease flush protocol. **(B)** Sequenced data was subset to those matching Illumina sequencing of the isolates, assembled and quality controlled.

Several isolates were set aside for independent Random Forest classifier training used in the SNP polishing and phylogenetics pipeline.

We sequenced a total of 181 unique isolates from a paediatric 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 Far North Queensland collected in 2019 (Cairns and Hinterlands, Cape York Peninsula, Torres Strait Islands, processed at Cairns Hospital, n = 86) (Fig. 1, Online Supplementary Tables). Oxford Nanopore Technology (ONT) sequencing was conducted using a minimal, dual-panel barcoding scheme, multiplexing 2 x 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 - 6.47 Gigabases of sequence data per run (< 24 hours) resulting in low - medium coverage per isolate (ST93-JKD6159) (Fig. 2A). We excluded one infection with *S. argenteus* (FNQ) and one co-infection with *Mammaliococcus sciuri* (PNG). Isolates with matching Illumina data were retained to create a high-quality reference dataset for further evaluation of genome assembly and variant calling (n = 159, Fig. 1).

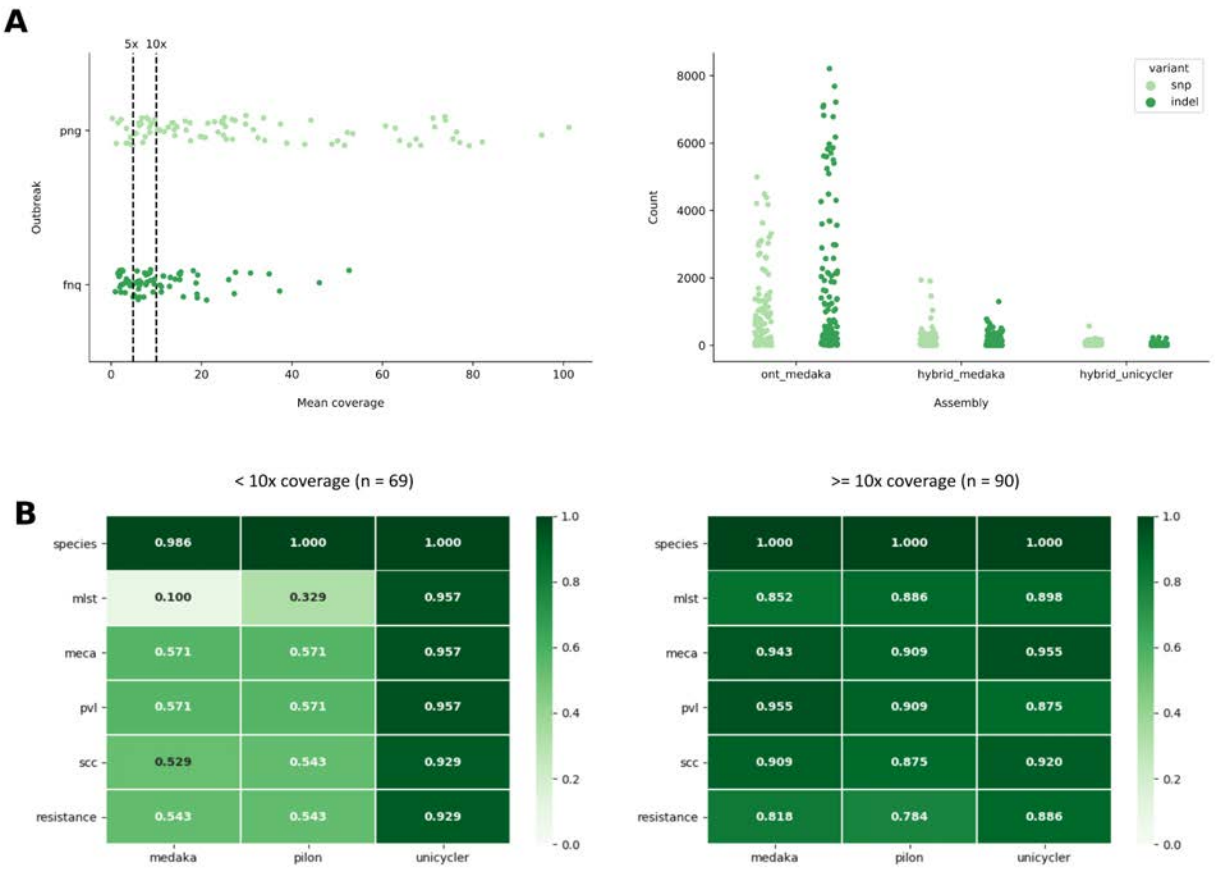

Fig. 2: **(A)** Average genome coverage (R9.4.1, RBK-004) of *Bontio* base called nanopore reads against the JKD6159 (ST93) reference genome (n = 159) where the dashed lines indicate the coverage thresholds chosen to evaluate genotyping (10x) and phylodynamic models (5x) in the Far North Queensland (FNQ) and Papua New Guinea (PNG) outbreaks. SNP (light green) and indel (dark green) 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 *Unicycller* (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 10x coverage threshold.

Genome assembly and genotyping validation

Short-read reference genomes, long-read polished nanopore genomes, and long-read hybrid genomes (*Pilon* (*157*) corrected long-read assemblies, *Unicycler (158)* short read assemblies with long-read correction) were assembled using a standardized Nextflow (*14*) pipeline wrapping *Shovill*, *Flye (159)*, *Medaka* and other components (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* (*129*) or the hybrid assemblies with Unicycler (Online Supplementary Tables), which first assembles short-reads and then scaffolds the assemblies with long reads to generate contiguous whole genome assemblies.

Compared to Illumina reference assemblies, SNP and indels were frequently occuring in low-coverage uncorrected nanopore assemblies (Fig. 2A, right). Errors were considerably reduced in high-coverage isolates leading to assembly identities ranging between 0.9993 and 0.9999 in the *dnadiff* metric (*160*) (Online Supplementary Tables). Recovery of complete chromosomes and *S. aureus* specific genotypes from uncorrected long-read assemblies was sufficient for high- coverage isolates in our collection (Fig. 2B, > 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. 2B). 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. 2B, < 60%, Online Supplementary Tables). Hybrid long-read correction with *Pilon* did not markedly improve genotype recovery in low-coverage isolate; however, recovery improved in the *Unicycler* hybrid assemblies (Fig. 2A, 2B). Lower SCC*mec* subtyping performance was likely due to remaining insertions or deletions from nanopore data impacting on the large cassette chromosomes (> 20kb). *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. 2B). For genome assembly and genotyping, our dual-panel sequencing approach recovers nanopore genotypes in high-coverage isolates (> 10x) although some errors remain, particularly in sequence type calling and SCC*mec* subtyping.

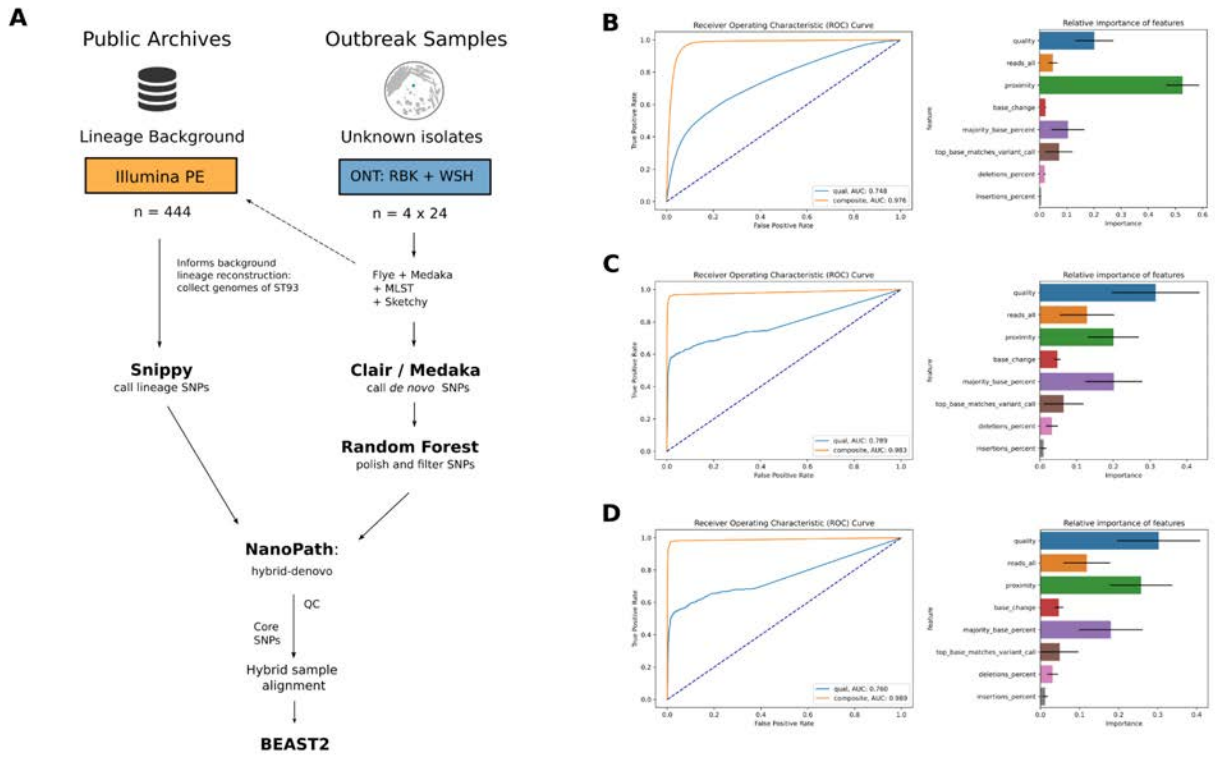

Fig. 3: **(A)** Workflow outlining culture-based protocol for community-associated *Staphylococcus aureus* nanopore sequencing using successive barcode panels on Oxford Nanopore Technology (ONT) MinION flow cells (R9.4.1). MLST typing informs the background population genome collection from a previous study (Illumina). Outbreaks in Papua New Guinea and Far North Queensland 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)** Area under the curve (AUC) scores of quality (blue) or composite (orange) 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 Far North Queensland isolates and **(D)** ST93 from Papua New Guinea with matching Illumina data and Snippy reference calls (all n = 3).

Training and evaluation of Random Forest SNP polishers

Next, we aimed to accurately reconstruct the PNG and FNQ outbreaks within the maximum-likelihood 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, 6616 SNPs) and *Megalodon* which accurately reconstructed the divergence of the PNG clusters from the Australian East-Coast (Fig. S1). 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 (Chapter 2.1). 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 *Escherichiae coli*, *Saccharomyces cerevisiae*, and *H. sapiens* (v1.2.3). While 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%, Fig. S2).

We next adopted SNP polishers using Random Forest classifiers originally developed by Sanderson and colleagues (*43*) to correct nanopore variants in *Neisseria gonorrhoeae* from metagenomic data (Fig. 3, Methods). Each 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, Methods) demonstrated high area under the

curve (AUC) scores (0.976 - 0.989, orange) while models trained on quality features alone showed suboptimal AUC performance (0.748 - 0.760, blue) (Fig. 3B-D).

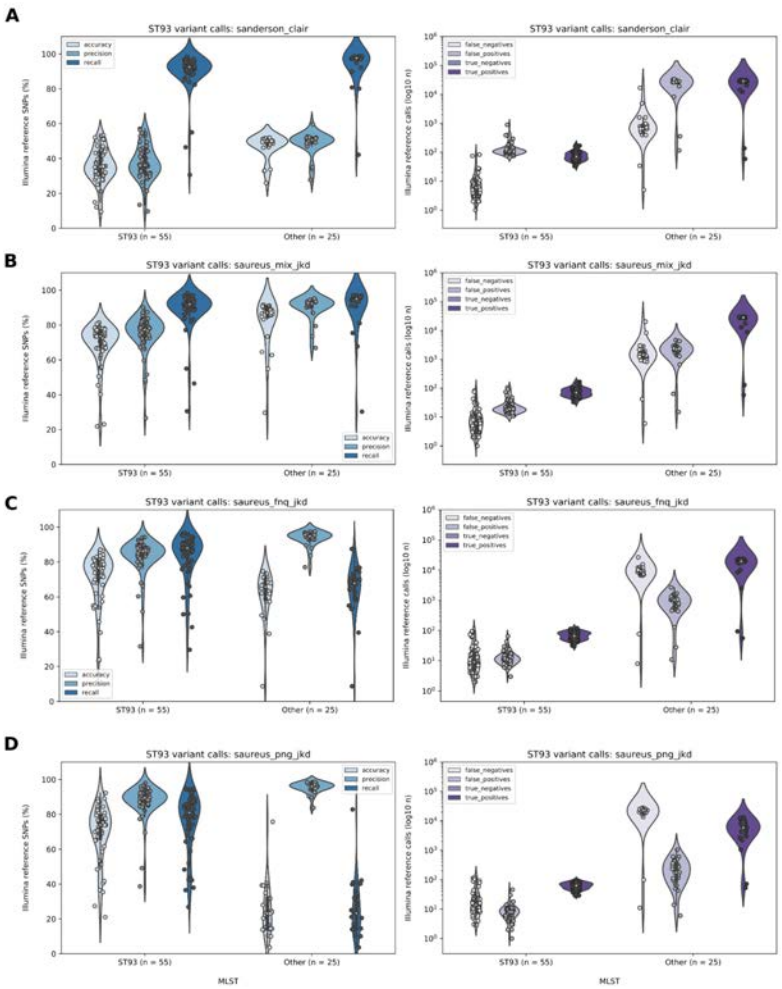

Fig. 4: Trained Random Forest SNP polisher evaluation showing left: accuracy, precision and recall of *Clair* nanopore SNP calls against matching Illumina reference SNPs called with Snippy. Plots are split into ST93 outbreak isolates (inside left) and other sequence types (inside right) from Papua New Guinea (PNG) and Far North Queensland (FNQ) combined. In the right-hand plots the number of false negatives, false positives and true positive SNP calls for the groups is shown on a log-scale. Models were trained on three Illumina matched isolates from between-species **(A)** *Neisseria gonorrhea* from Sanderson et al. within species **(B)** *Staphylococcus 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, > 10x coverage). Outliers in the tails of the distributions are novel multi-locus sequence type variants of ST93.

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 (Figs. 1B). Evaluations indicated that all trained SNP polishers increased accuracy and precision with slight reductions in recall (Fig. 4). However, sub-optimal 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% ± 12.48s accuracy, 75.94% ± 14.56s precision) and other sequence types (81.94% ± 14.56s accuracy, 90.11% ± 6.83s precision). However, despite significant baseline improvement, the inter-species and mixed-sequence type models the number of false positive SNP calls remained in the range of 100s to 1000s (right column, Fig. 4A-B). Training the models with isolates from the same sequence type (ST93, FNQ) slightly improved performance (ST93: 71.69% ± 13.99s accuracy, 83.33% ± 10.42s 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% ± 16.78s accuracy, 87.57% ± 9.83s precision) but incurred a steeper cost to accuracy and recall in non-outbreak 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 - 10s of false SNP calls compared to the *N. gonorrhoeae* and mixed sequence type model (Fig. 4).

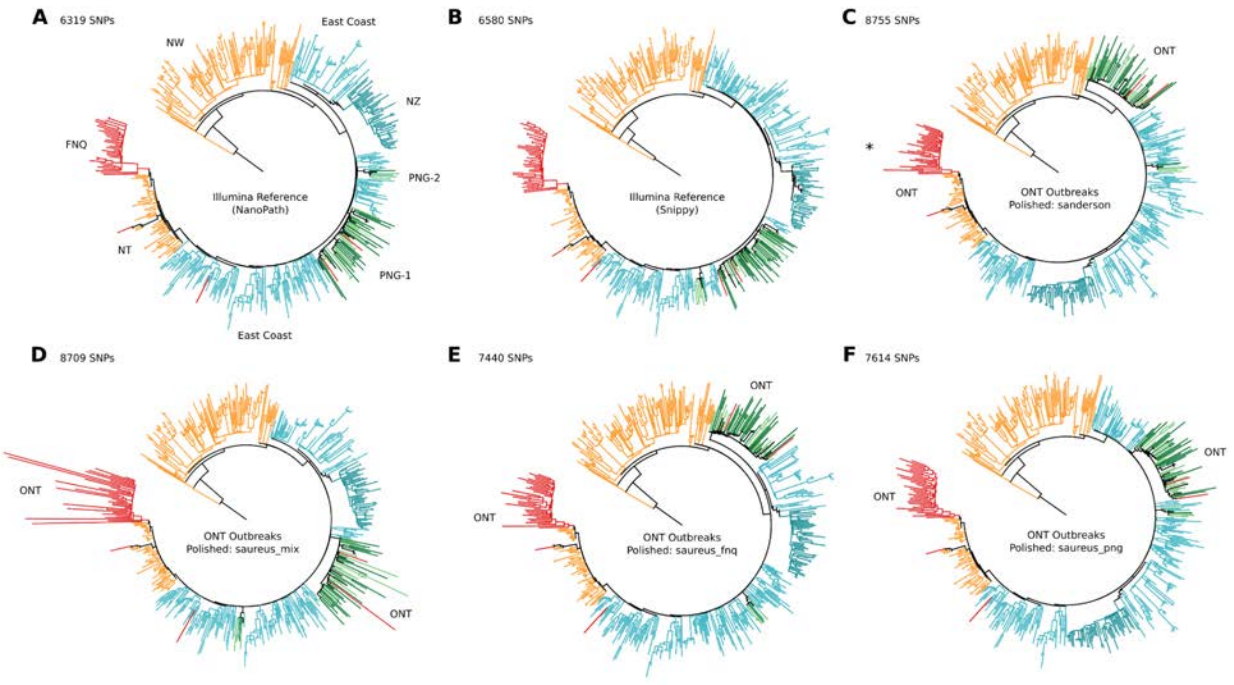

Fig. 5: Comparison of maximum-likelihood phylogenetic topologies of ST93. Illumina reference trees were constructed with *NanoPath* **(A)** and *Snippy* **(B)**. All other trees are hybrid phylogenies including the nanopore data of the outbreaks in Far North Queensland (FNQ) and Papua New Guinea (PNG) (> 5x coverage) within the ST93 background population (Illumina, n = 531) **(B)** after polishing *Clair* SNPs using the trained Random Forest classifiers (**C**: *Neisseria gonorrhoeae*; **D-E**: *S. aureus* mixed and lineage-specific). Asterisk (*) denotes two isolates with excessive branch lengths that were removed for visual clarity (Fig. S6).

Phylodynamic reconstruction using polished *de novo* 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 1x (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 5x coverage (n = 531 / 562) due to low accuracy and precision of these isolates in the SNP polishing step (Fig. 4, Fig. S4). 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).

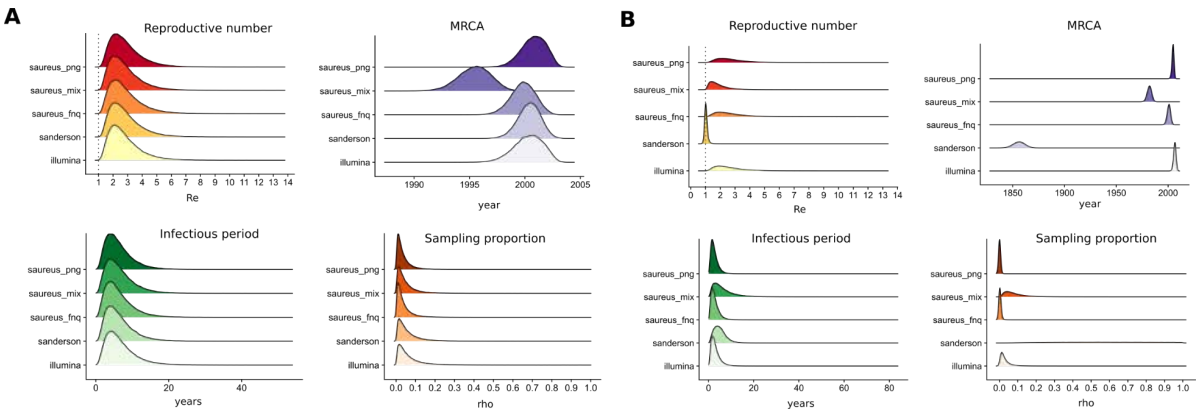

Fig. 6: Posterior distributions of the effective reproduction number (R_e, red), most recent common ancestor of the outbreak (MRCA, purple), infectious period ($1 \div \delta$, green) and sampling proportion (ρ, orange) for the nanopore sequenced outbreak clades in Papua New Guinea (**A**, PNG, n = 56)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

NanoPath's core alignment construction reproduced *Snippy*'s core alignment from Illumina data (6650 SNPs vs. 6662 SNPs, Fig.5A, B). When we called *Clair* SNPs on isolates with > 5x 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). All polished SNPs produced reasonable alignments, where FNQ and PNG polishers produced alignments closest to the Illumina reference (Fig. 5, Supplementary Table 2). 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 (*42*). 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 (Methods) and *Snippy-core* with matching isolates (n = 531, Fig. 5).

All major clades and sub-populations of the background population (North West, East Coast, NT, and NZ) 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 NZ 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 U.K. 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 NT 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 sequence type alignments (Fig. 5, in particular due to two isolates: PNG-36 and PNG-62, Fig. S6). The alignment based on SNPs polished using outbreak sequence type (ST93) isolates were most consistent with the Illumina reference phylogeny of ST93. We note that within-lineage polishing did not require within-outbreak polishers, e.g. 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 (Chapter 2.1). We ran *BEAST2* MCMC chains on the outbreak subsets of the full SNP alignment with

sufficient isolates (n_{PNG-1} = 53; n_{FNQ} = 32) using a fixed substitution rate of the whole-lineage median posterior estimate (3.199×10^{-4}). This was necessary as non-random 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 tMRCA. We note that the models were efficiently run on a standard NVIDIA GTX1080-Ti GPU using *BEAST2* with the *BEAGLE* library at speeds of $<$ 3-4 minutes / million steps in the MCMC (5-7 hours per run and GPU) making timely parameter estimation for outbreak responses feasible on low-cost hardware. On a NVIDIA P100 GPU, walltime decreased to < 50 seconds - 1 minute / million steps in the MCMC, around 1-2 hours 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 MRCA in the mixed polishing model (Fig. 6B, Table S2). More variable posterior estimates were observed in the FNQ clade (Fig. 6), consistent with higher variability in branch lengths as a result of excessive false positive SNP calls retained in low-coverage FNQ isolates (Fig. 5). Nevertheless, when compared to the NanoPath Illumina reference estimates, ST93-polished estimates (saureus_png, saureus_fnq) closely resembled those of the reference, with only minor deviations (Fig. 6, Table S2). Estimates were consistent with full lineage-wide analysis (R_e > 1.5 - 2.0, Chapter 2.1) and we observed robust estimates in an exploration of the R_e prior (Table S2, Fig. S6, S7). 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 (5x - 10x). Finally, we implemented training, evaluation and deployment of SNP polishers for within-lineage transmission modelling in Nextflow.

2.2.3. Discussion

In this study we provide a method for variant polishing and phylodynamic modelling of bacterial whole genome data using low-coverage nanopore sequencing. Previous studies using (high-coverage) nanopore data have evaluated phylogenetic reconstructions on few and distantly related isolates of *Neisseria gonorrhoeae* as well as other bacterial genomes from assembly (*43*, *155*, *156*). A recent pipeline for cluster identification using 6 strains per MinION flow-cell (42 on 7 flowcells) successfully identified clusters in four distinct lineages, using a whole genome assembly based phylogeny (*161*). 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 (*43*) can sufficiently reduce the number of false positive SNP calls from neural-network variant caller *Clair* v.2.1.1 (*32*). Hybrid lineage alignments of ONT sequence and Illumina background data of the outbreak lineage (ST93) can be constructed, and effective reproduction numbers accurately modelled using birth-death skyline models in *BEAST2*.

We evaluated genotype reconstruction against previously sequenced Illumina data (Chapter 2.1) demonstrating the superior quality of hybrid assembly with *Unicycler*. Our genotyping analysis showed that for high coverage isolates (> 10x) genotyping directly from polished nanopore assembly was comparable to hybrid approaches (Fig. 2). We used the most recent models in *Bonito* v0.3.6 for base calling followed by polished long-read assembly or hybrid assembly. With the imminent release of R10.3 pores and associated increases in raw read accuracy (estimated at Q20) we expect that 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 > 5x - 10x coverage. We chose here to

implement a rapid and minimal protocol to evaluate its application in remote reference laboratories, such as at 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 (Chapter 2.1), 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 isolate 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 (Fig. S3), 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 (*43*). Polishing success of *Clair* calls suggest that the features selected here - in particular the proximity and quality features (Fig. 3B-D) - were effective at removing systematic false positive SNP calls, when trained with reference calls against specific reference genomes (e.g. ST93 outbreak genomes against ST93-MRSA-IV JKD6159 reference genome). Systematic error correction is supported by observations that SNP calling did not improve considerably using *Bonito v0.3.6* R9.4.1 DNA models compared to Guppy high accuracy (Fig. S5) and methylation-awar e models (data not shown). SNP polishers therefore appear to exploit systematic errors in the neural networks (trained on human variant calls) when applied to bacterial genomes. It remains to be seen whether re-training *Clair* or *Medaka* neural

networks on *S. aureus* specific signal- and sequence-data would improve species-specific SNP calls.

We demonstrate the utility of our method by sequencing novel isolates of community-associated MRSA from a paediatric osteomyelitis outbreak in the highland towns of Kundiawa and Goroka (Papua New Guinea) and routine surveillance in remote northern Australia (Far North Queensland) (Fig. 1, n = 181). A protocol that minimised cost (without optimisation) allowed us to sequence two 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 (< 5x) whereas even low-medium coverage isolate (>= 5x) produced consistent estimates of the effective reproduction number for the PNG an FNQ clades, when compared to the Illumina reference (Fig. 6). Accurate modelling was possible even with inter-species 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 optimisation 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, May 2021), or short read elimination kits. While we were ultimately unable to use a total of 32 isolates (< 5x coverage) in the phylodynamic estimation, the cost per *S. aureus* genome using the 24x 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). Further optimization would incur small additional cost and can be conducted for bacterial pathogens of interest in sufficiently resourced laboratories. While we chose *S. aureus* as a model organism for this work mainly due to its relatively small genome (2.8 Mbp) and our interest in sequencing the outbreaks in PNG and FNQ, core principles and methods used in this

study are immediately applicable to other bacterial pathogens and all steps of the pipelines are implemented in replicable Nextflow workflows (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 false positive and false negative 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 (*147*) in combination with *BEAST2*. Moderate acceleration on standard hardware (< 5 - 7 hours) and increased acceleration on NVIDIA P100 GPUs (< 2 hours) 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.

2.2.4. Materials and Methods

Outbreak sampling in FNQ and PNG. We collected isolates from outbreaks in two remote populations in northern Australia and Papua New Guinea (Fig. 1). Isolates associated with paediatric 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 neighbouring Eastern Highlands province town of Goroka. We supplemented the data with MSSA isolates associated with severe hospital-associated infections and blood cultures in Madang (Madang Province) (n = 8) and Goroka (n = 12). Isolates from communities in Far North Queensland, 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 LB broth at 37°C overnight and stored at -80°C in glycosol and LB. Illumina short-read data from the ST93 lineage (*42*) included in this study were collected from the European Nucleotide Archive (Online Supplementary Tables).

Nanopore sequencing and basecalling. 2 ml of LB broth was spun down at 5,000 x g for 10 minutes and after removing the supernatant, 50 ul of 0.5 mg / ml lysostaphin were added to the tube and vortexed. Cell lysis was conducted at 37°C for 2 hours with gentle shaking followed by a *proteinase K* digestion for 30 mins. 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 ul of nuclease-free water, quantified on Qubit, and DNA was stored at 4°C until library preparation. Library preparation was done using approx. 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, Fig. S5), 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 x 12 barcoded (SQK-RBK004) isolates per flow cell in two consecutive 18-24 hour 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 (*162*) 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 2910 / 294461 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 200x coverage with *rasusa* v0.3.0 (https://github.com/mbhall88/rasusa) and filtered Q > 7 with minimum read length of 100 bp using *nanoq v0.8.0* (https://github.com/esteinig/nanoq). Fastp v0.20.1 (*128*) 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 (*159*) was used in conjunction with four iterations of *minimap2* v2.17-r941 (*163*) + *Racon* 1.4.20 (*164*) 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 (*132*) (from reads) and *SCCion* v0.4.0 (https://github.com/esteinig/sccion), a wrapper around common assembly-based genotyping tools and databases (*130*, *131*, *165*) for *S. aureus*. We called species, resistance genes, virulence factors, Panton-Valentine leukocidin (PVL), multi-locus 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 (Supplementary Fig. 2). Coverage against the reference genome (ST93:JKD6159) (*65*) 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, 6161 SNPs) and reference Illumina core alignments including the outbreaks in FNQ and PNG isolates (> 5x, n = 531, 6580 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 (*43*) who use the RandomForest classifier from *scikit-learn (166)* with default hyperparameter settings and feature extraction with *pysamstats*. Like the original implementation, we sub-sampled isolates to 2, 5, 10, 20, 50 and 100x 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, ST93) (*saureus_mixed*), one of Far North Queensland 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. gonnorhoeae* 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*, false positive (FP) as ONT SNPs that were not called with *Snippy* }, and false negative (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 (Online Supplementary Tables, Fig. 4).

Hybrid core site outbreak alignments. To contextualise 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: 1x). We first polished ONT SNPs from *Clair* with the trained Random Forest models, including the *N. gonnorhoeae* dataset from Sanderson et al. (*43*). We then created reference alignments of the Illumina data (ST93 background and outbreaks, n = 531, > 5x) 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 the Lewi's ascertainment bias correction for SNP alignments. Trees were rooted on SRR115236 (early isolate from 1992, near the root of the phylogeny, (*42*) and decorated with meta data of sample origin at state level in ITOL (*167*). 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 (3.199×10^{-4}, Chapter 2.1) to account for the loss of temporal signal in the soutbreak 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 (HPD) 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.

2.3. Sketchy: genomic neighbour typing for bacterial outbreak surveillance

Eike Steinig[1,2,*], Tania Duarte[1], Miranda Pitt[1], Izzard Aglua[3], Annika Suttie[3], Christopher Heather[4], Simon Smith[5], William Pomat[3], Paul Horwood[6], Emma McBryde[2], Lachlan Coin[1]

[1]The Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Australia, [2]Australian Institute of Tropical Health and Medicine, James Cook University, Townsville & Cairns, Australia, [3]Papua New Guinea Institute of Medical Research, Goroka, Eastern Highland Province, Papua New Guinea, [4]The Townsville University Hospital, Queensland Health, Townsville, Australia, [5]Cairns Hospital and Hinterland Health Service, Queensland Health, Cairns, Australia, [6]College of Public Health, Medical & Veterinary Sciences, James Cook University, Townsville, Australia

Keywords: ONT | Genomic neighbor typing | MinHash | Outbreaks | Genotyping | *S. aureus* | *K. pneumoniae*

* Corresponding authors: eike.steinig@unimelb.edu.au, lachlan.coin@unimelb.edu.au

Abstract

Nanopore sequencing has been critical for obtaining timely epidemiological and clinical information from bacterial pathogens, but rapid genotype prediction is still challenging. In this study, we explore the use of genomic neighbor typing for large-scale outbreak surveillance of community-associated *Staphylococcus aureus* infections from remote communities of Papua New Guinea (PNG) and Far North Queensland (FNQ). For this purpose, reference databases were scaled to species-wide and lineage-representative whole genome collections from public sources, encompassing the available genetic diversity and genotypes of the target species. We developed *Sketchy*, a variant of genomic neighbor typing that queries reference databases using MinHash. We assessed the performance and limitations of genomic neighbor typing approaches with *Sketchy* on two clinical validation datasets from community-associated MRSA outbreaks in PNG and FNQ with matching Illumina data (n = 158). Our results indicate that *Sketchy* is effective for genotype predictions from less than 1000 reads (sensitivity / specificity 80-90%) --- with some important limitations in database representation and within-lineage genotype predictions. Although genomic neighbor typing has some inherent limitations related

to database representation, heuristic predictions from few reads allowed us to conduct multiplex genotyping experiments in situ at the Papua New Guinea Institute of Medical Research in Goroka, on low-throughput Flongle adapters and using multiple successive libraries on the same flow cell (48 strains on MinION, sensitivity / specificity ≈ 80-90%). Lastly, we used host-depleted sequencing of sputum samples from a cystic fibrosis patient undergoing antimicrobial therapy to confirm the limited resurgence of a pan-susceptible S. aureus strain after treatment.

2.3.1. Introduction

Epidemiological and clinical data on bacterial infections, such as strain provenance, antimicrobial susceptibility and pathogen-specific markers, are valuable targets for decision makers, but their timely inference from genomic data is challenging. Nanopore platforms are particularly suited to point-of-care diagnostics and genomic epidemiology (*161*, *168*), in part due to low acquisition and maintenance cost, as well as high portability. Sequence fragments are streamed from the device and available for immediate analysis on mobile computing devices (*17*, *169*). Fast methods for genome-informed surveillance and genotyping are especially relevant for bacterial diseases, where faithful genome assembly and genotyping of gene markers often require high coverage, sequencing throughput and time. Genomes also cannot be assembled easily from complex metagenomic backgrounds contaminated by host genetic material or from complex microbial communities at low abundance. Exploiting the real-time sequencing capacity of nanopore platforms, a school of streaming methods and diagnostics for bacterial pathogens on nanopore platforms has been developed over the past few years, including pipelines for batch assembly and gene detection approaches (*170*, *171*), novel algorithms for streaming assembly and genotyping (*110*, *172*), and sensitive approaches to antimicrobial resistance prediction, as well as taxonomic identification (*173*, *174*). Clinical

nanopore sequencing studies have demonstrated the feasibility of marker detection and assembly of metagenome-assembled genomes (MAGs) for genotyping, particularly from samples where host nucleic acids are at low abundance, in samples with high bacterial loads, or supported with additional short-read sequencing (*46*, *171*, *175*, *176*). In these studies, genotyping from lower respiratory infections and cystic fibrosis patients was feasible within hours, and particularly efficient when preceded by host nucleic acid depletion protocols (*46*) or enriched by culture-based approaches (*175*).

While metagenome assembly and direct-evidence approaches are improving strain determination in clinical samples, heuristic approaches have been considered for rapid classification of clinical traits, for instance when considering time-critical scenarios like sepsis or septic-shock (*46*). Břinda et al. developed a principle termed 'genomic neighbor typing' in which traits were inferred from a stream of uncorrected nanopore reads based on k-mer matches against a database of whole genomes and the closest associated geno- or phenotype (minimum inhibitory concentrations) (MICs). Using a hierarchical within-lineage score informed by the phylogenetic relationships between isolates in the reference database (RASE), the authors assessed phenotypic resistance over susceptibility, as well as the prediction over its most likely alternative (using a preference score). Lineage detection and antimicrobial susceptibility profiling were extremely fast, often within seconds to minutes of sequencing. While this initial implementation of genomic neighbor typing focused on lineage and susceptibility typing in clinical settings, the authors also demonstrated that sequence type and serotype matching worked in *Streptococcus pneumoniae*. Břinda et al. further suggested that genomic neighbour typing can potentially be used for any typing scheme, for which there are genome-associated features in the database, and may be a useful heuristic in outbreak scenarios. Genomic neighbor typing essentially transforms the problem of observing direct evidence for a genotype with limited data (e.g. reads mapping to genes of interest) into a problem of inferring indirect

evidence for a feature using a previously compiled genome and genotype database (e.g. prediction of resistance based on closest whole genome inference).

A critical component of genomic neighbor typing is sufficient representation of the genomic neighborhood in the reference database. Břinda et al. constructed reference sets from local and national collections to demonstrate the principle of genomic neighbor typing on lineage and phenotype inference in *S. pneumoniae* (n = 789) and *Neisseria gonorrhoeae* isolate and metagenomes (n = 4782). However, for clinical applications and in particular for lineage calling, a small reference database of the globally available sequence space for a pathogen may be insufficiently representative of species-wide lineage diversity and miss important strains that may have entered local epidemiological space. Data from isolates not represented in the databases for *S. pneumoniae* and *N. gonorrhoeae* indicated a considerable decrease in its ability to call lineages and therefore target the correct genomic neighborhood of these strains (*46*). Břinda et al. further discuss the possibility that genomic neighbor typing could be used to rapidly determine any genotype feature, including species-specific virulence markers or associations of strains with previously sequenced outbreaks. Overall, the limitations of genotype-based genomic neighbor typing have not been explored outside their initial application. In particular for outbreak scenarios, large reference databases representing the entirety of species-wide lineage genomes may be required, for example in remote regions where sequencing has not been conducted previously and prior knowledge about circulating strains is absent. We note that due to efficient predictions, genomic neighbor typing could also facilitate massively parallel genotyping by multiplexing strains, including on used floc cells or Flongle adapters on the MinION. Provided that genotypes can be inferred from few reads, it should be feasible to scale genotyping cost-effectively requiring only standard nucleic acid extraction and multiplex sequencing protocols. In combination with the affordability and portability of nanopore sequencing devices, outbreak and genotype surveillance of bacterial

pathogens may be possible in remote regions and under-resourced countries where implementing genomic epidemiology is currently difficult to achieve.

MinHash is a variant of locality-sensitive hashing, initially used for detection of near-duplicate websites or images, that has been extensively used in genomics since its implementation in *Mash (44, 177, 178)*. It was recently tested as a classifier of genetic clusters and ribotype prediction in *Clostridioides difficile* short-read sequencing data using *sourmash (179)*. MinHash distances or containment screening (*44*, *45*) present a highly scalable method to implement a generalised, genomic neighbor typing scheme with reference sketches that may comprise tens of thousands of genomes for comprehensive lineage and genotype representation. Reference sketches from Mash used for uncorrected nanopore reads (k = 15, s = 1000) are small compared to taxonomic databases (MB vs. GB) allowing for applications on mobile devices with limited memory and processing capacity. Genome informed databases can be expanded when new genomes and genotypes become available. As genomic neighbor typing may be able to operate entirely on genotypes, this would allow for the construction of reference sketches from public genome collections, like the European Nucleotide Archive (ENA). In this study, we evaluate genomic neighbor typing against species-wide bacterial pathogen sketches using MinHash techniques by Ondov et al. (*44*). We develop a genomic neighbor typing streaming implementation of Mash using a lineage-resolved ('strain agnostic') database of *Staphylococcus aureus* (n = 38,898) and a a smaller sketch of *Klebsiella pneumoniae* (n = 8,149). We demonstrate the utility and limitations of heuristic genotype prediction on previously unknown outbreak strains from remote northern Australia and Papua New Guinea (n = 159) which consisted of an independent nanopore validation data set for S. aureus with matching Illumina reference data.

2.3.2. Results

We developed *Sketchy*, a Rust client for genomic neighbor typing using min-wise shared hash queries against custom reference sketches in *Mash* (Fig. 1). As species-representative databases are key to a comprehensive genotyping space available for genomic neighbor typing approaches, we surveyed the available short-read data in the European Nucleotide Archive (January 2019, n = 61,598) to obtain candidate genomes for genotyping of S. aureus. Sequence reads were downloaded, quality controlled, filtered by contamination, assembled and genotyped using a pipeline to standardize genotyping results across all isolates (Methods). We included multi-locus sequence type, presence of *mecA*, major SCC*mec* types, Panton Valentine leukocidin (PVL) toxin markers and resistance phenotype profile inferred with *Mykrobe* (12 antibiotics), retaining a total of 38,898 genome assemblies passing our quality filters for database construction. In order to validate the genomic neighbor typing approach for other species, we also collected the available Klebsiella pneumonia genomes from the ENA (July 2019, n = 14,872). We then used Mash to construct the reference sketch (k = 15, s = 1000) and Sketchy to construct the genotype database indices for the client. *S. aureus* sketch size for queries was 162 MB (n = 38,898) and the *K. pneumoniae* (n = 8,149) was 34 MB. In compressed format, reference sketches used 2.6 MB and 1.4 MB disk space respectively for distribution through the client. Monitoring of resource utilisation during screen and stream operations in *Mash* and *Sketchy* suggest that they can be run efficiently on a single CPU and $\approx$ 350 MB of memory, depending on sketch size and extent of compiled reference sketches (Fig. S1).

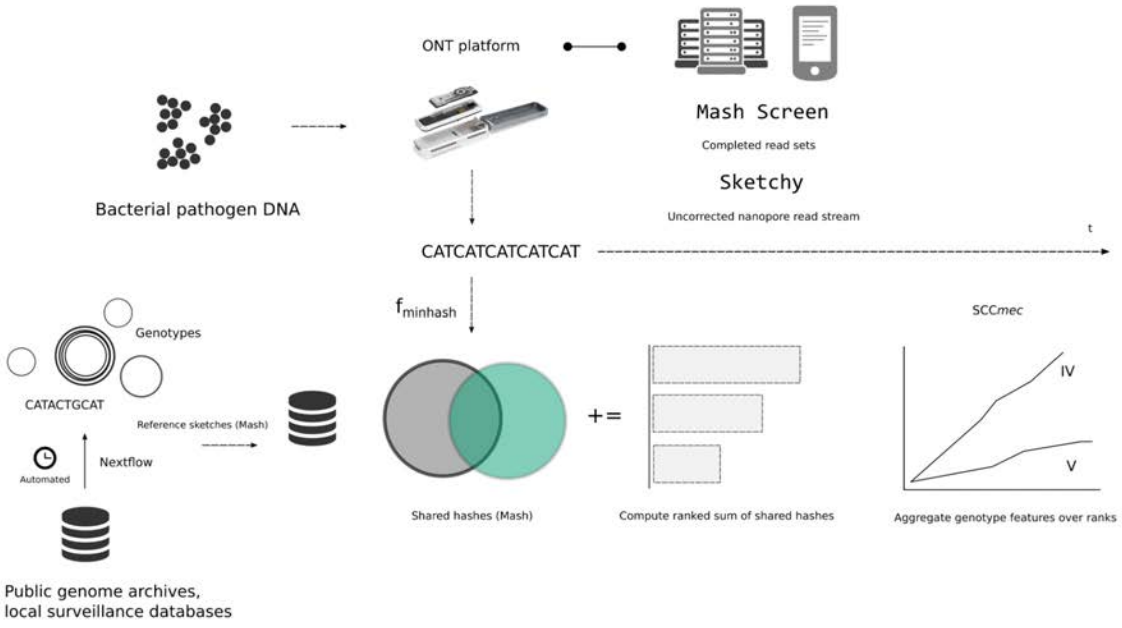

Fig. 1: *Sketchy* client components, showing the inputs of bacterial DNA and a reference sketch constructed from whole genomes derived from public collections (which can be automated and updated). Reference sketches are constructed with *Mash*. Reads are queried against the sketch either using the *Mash* screening algorithm on completed read sets, or the *Sketchy* online algorithm for streaming reads in real-time. In the latter, reads are queried against the reference sketch, and the sums of shared hashes for each genome in the *Sketchy* reference database are first updated and then ranked at each new read by features. In a prediction stage, these scores are then linked to the genotype index associated with the whole genome reference sketch for heuristic genomic neighbor inference of genotype features. Diagnostic plots and preference scores are computed at prediction.

S. aureus and *K. pneumoniae* evaluations

We first explored the initial application of four *Mash*-based methods to conduct genomic neighbor typing from *S. aureus* and *K. pneumoniae* reference sketches: the distance method (*mash dist*), the screen containment method (*mash screen*), screen containment with a winner-take-all strategy (*mash screen -w*), and the *Sketchy* streaming algorithm. All samples in each of the initial test sets (metagenome-extracted or isolated, n = 5) were newly sequenced or had been published after database collection, and therefore were not contained in the reference

sketches. We compared feature predictions by method against the Illumina reference genotypes from each strain at pre-defined read thresholds (Fig. 2). While small differences in accuracy were observed between methods, genotypes were recovered at ≈ 80% accuracy with more than 200 reads (Fig. 2A). Only the Zymo mock community strain, ST9-MSSA (ciprofloxacin, tetracycline and penicillin resistant) *S. aureus* strain was predicted incorrectly in SCCmec-associated features and tetracycline resistance phenotype, which was due to the absence of tetracycline resistant lineage genotypes in the reference database. Diagnostic plots of the ranked sum of shared hashes scores and preference scores indicate confidence in detecting incorrect predictions (orange color, Fig. 2B), featuring low preference scores and close shared hashes scores over the course of the read stream predictions (Fig. 2C). Preference scores near the threshold value (threshold = 0.6) were observed for correct predictions of penicillin and ciprofloxacin resistance (dark seagreen, Fig. 2B), the first of which was immediately distinguishable in the diagnostic plots. In addition, the NOR1 ST772-MRSA strain was intermediately predicted to be ST772-MSSA and showed low preference scores across SCC*mec* related features (including linked gentamicin resistance) and PVL, despite correct predictions and distinguishable scores in the diagnostic plots (Fig. S3). When shared hashes and preference scores are tracked during online prediction with our streaming algorithm, correct genotype predictions are possible much earlier than the chosen endpoints, in some cases even after seconds of sequencing and few reads, including sequence type, penicillin resistance and PVL (Fig. 2C). Predictions for the metagenomic and cultured *K. pneumoniae* strains were less accurate and failures largely related to highly variable serotypes (K- and O-locus sequence types) and complements of resistance genes associated with the carriage of mobile elements (e.g. tetracycline resistance or yersiniabactin), which were not contained in the reference sketch in the specific lineage-genotype combinations. While all lineages were correctly inferred, a considerable number of resistance elements and serotypes in the BAMBI strains were confidently called incorrectly (red color) with the remainder of incorrect calls being correctly

identifiable (orange, Fig. 2B). We additionally sampled read order with replacement at each read thresholds and note that read order may have some impact on classification performance in early predictions (< 500 - 1000 reads), which was slightly more pronounced for the streaming implementation, as expected from the online computation of the sum of shared hashes (Methods, Fig. S4).

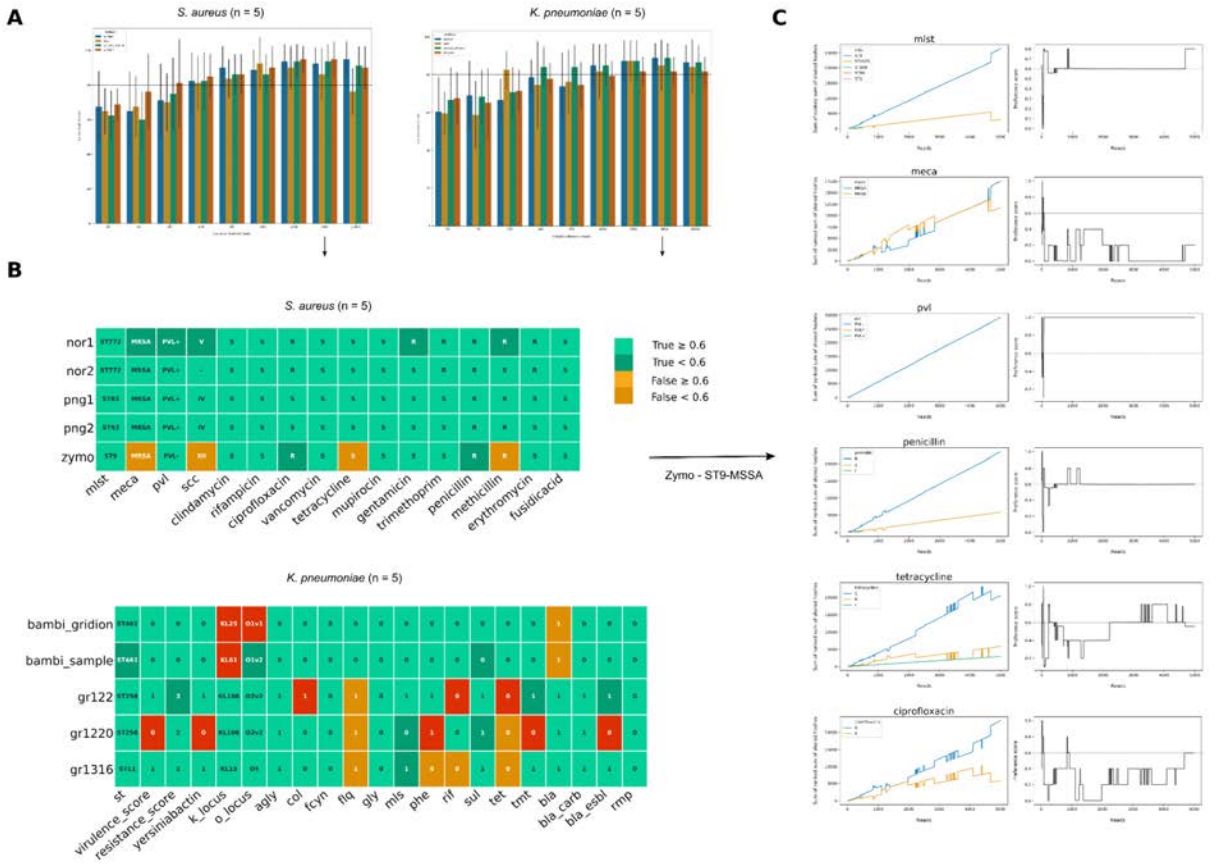

Fig. 2: ***(A)*** Comparison of *Sketchy* genotyping accuracy of the initial test sets of *S. aureus* and *K. pneumoniae* strains (n = 5) at successive read thresholds (20 - 10000). Horizontal line denotes 80% accuracy threshold, each grouped bar represents one of the four methods we compared (dist, screen, screen-w, stream) and error bars are standard deviations of prediction accuracy over the five samples. **(B)** Genotype truth evaluation heatmaps of *S. aureus* and *K. pneumoniae* samples at 5000 reads using *Sketchy* streaming function. Columns with annotations contain the predictions of each of the samples (rows) with color denoting whether predictions were true or false compared to reference Illumina data (green, orange), and whether they had a preference score above a preset threshold (0.6, red shows wrong predictions where preference scores falsely predicted support, and light green correct predictions with strong support). **(C)** Diagnostic plots of the ST9-MSSA reference strain streaming genotype predictions from the Zymo mock microbial community, showing various levels of confidence of ranked sums of sum of shared hashes (per genotype, left) and preference score of the top ranking prediction at each read in the stream.

Genotype surveillance of community-associated outbreaks

We next evaluated *Sketchy* genotyping on two S. aureus outbreaks from remote communities in Papua New Guinea and Far North Queensland (n = 158), which had been sequenced at low-coverage using a dual-library protocol with interspersed nuclease washes (24 strains per MinION flow cell on a total of 8 flowcells) (Fig. 3, Online Supplementary Tables, see also Chapter 2.2). While most isolates belonged to the Australian ST93-MRSA-IV clone (Fig. S5), multiple sequence- and genotypes were recovered so that a relatively diverse within- and between-lineage dataset could be assembled for evaluation. In addition, these were the first S. aureus genomes recovered from Papua New Guinea (*38*), constituting an independent test dataset for genotype and performance evaluation of the *S. aureus* species reference sketch. We compared predicted genotypes against those from Illumina assemblies using the streaming algorithm and genotype predictions at 200 (Table S1) and 1000 reads (Fig. 3, Table 1). Performance metrics indicate an overall accuracy of 93.59% over all features after 1000 reads with expected overestimation of call failures due to linked SCC*mec* features. Binary features (PVL, mecA and resistance phenotypes) showed an overall accuracy of 94.66%, precision of 95.29% and sensitivity of 83.79 % while multi-class features (MLST and SCCmec-type) showed an accuracy of 86.07%, precision of 89.41% and sensitivity of 86.07%. Lower performance was largely driven by failure to genotype the linked SCC*mec* related features (SCC*mec* type, methicillin resistance, *mecA* - in the second, fourth and third last columns respectively, Fig. 3), while all other features showed an accuracy, precision, sensitivity and specificity > 85-90% (MLST, PVL and other resistance phenotypes) (Table 1). Lastly, when preference score thresholds were applied (colors, Fig. 3) false predictions could be reliably identified from low preference scores at the prediction point (orange, with exceptions in red), while most true predictions that did not meet the threshold were often identifiable through the diagnostics plots (data not shown). We also note that lineages were more often misclassified in the PNG data

(Fig. 3, right heatmap). This was reflective of a more diverse dataset than the FNQ outbreak, including several novel MLST variants of the outbreak sequence type (ST93, n = 3) that principally could not be classified, as they were not present in the reference sketch. However, all misclassified lineages in both FNQ and PNG data were either a novel MLST variant and the closest lineage match was correctly identified, or rare lineages not sufficiently captured in the public database we constructed (e.g. ST81 lineage, n = 4, *gmK* variant of ST1). In addition, we picked up on a cluster of ST243 isolates as described previously (Chapter 2.1, Online Supplementary Tables). *Sketchy* was therefore able to pick out the correct genomic neighborhood from a large 'strain-agnostic' reference database constructed from public data, demonstrating its application for novel outbreak scenarios where the pathogen had previously not been sequenced before (e.g. in Papua New Guinea).

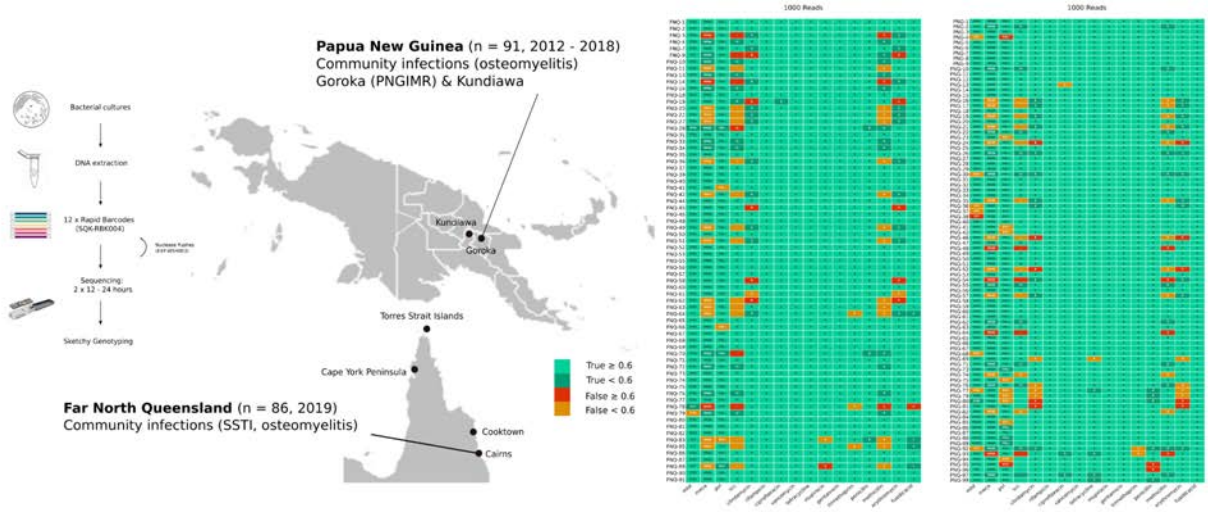

Fig. 2: Community-associated *S. aureus* outbreak validation from remote communities in Papua New Guinea (PNG) and Far North Queensland (FNQ) showing sampling sites of osteomyelitis and skin-and soft tissue (SSTI) infections, including a schematic of the dual-panel barcoding protocol of cultured samples (24 samples per flowcell on the MinION). Genotype truth evaluation compared with reference Illumina genotypes where predictions are true or false (seagreen, orange). Preference scores for each prediction are indicated below the threshold (0.6) and above - an orange color indicates a false prediction that would have no support amd a red color a false prediction that falsely would show support; a light seagreen color indicates a true prediction with strong support, whereas a dark seagreen indicates a true prediction with supprot below the threshold.

In situ genotyping and multiplexing experiments

We next ran a library of strains from the osteomyelitis outbreak in situ at the Papua New Guinea Institute of Medical Research in Goroka (Eastern Highlands). We multiplexed 12 strains onto a MinION flow cell, but ultimately obtained few reads per barcode (288 - 1896) due to malfunctioning laboratory equipment resulting in sub-optimal barcode attachment (Fig. 4A). Despite this challenge, we still obtained sufficient data to identify outbreak genotypes of all samples with *Sketchy* (Fig. 4A). One barcode (513 reads, PNG-69) was misclassified as ST8-MRSA-II instead of ST93-MRSA-IV, but the majority of genotypes had low preference scores and could be identified as misclassifications. One outbreak isolate was misclassified as MSSA instead of MRSA, but preference scores wrongly indicated confidence in the prediction. We next tested the rapid barcoding library protocol (SQK-RBK004) on Flongle adapters. Strains sequenced to ≈ 1000 reads per barcode in 24 hours, which allowed for genotyping with Sketchy (Fig. 4B). Four barcodes in total had misclassifications in SCC*mec* related features and one strain was misclassified as ST8-MSSA. Genotype misclassifications were mostly identifiable from low preference scores, except for the lineage misclassification of ST8 instead of ST93, which perplexingly suggested high preference score support. Finally, we tested a sequential multiplexing protocol, running each library with twelve strains for two hours on a single MinION flow cell, interspersed with nuclease washes (EXP-WSH-003), to assess the capacity for large-scale outbreak surveillance with limited resources. While we planned to sequence all PNG outbreak strains on the flow cell, accumulation of air bubbles on the flow cell prevented us to run additional libraries, but ≈ 900-1000 pores remained after four nuclease washes and libraries (EXP-WSH003). Overall, the four panels - barcoded with the same isolates as in the dual-panel protocol (Fig. 3) - accumulated the majority of misclassifications in SCC*mec* related features, most of which were identified with low preference scores (Fig. 4C). One ST81 strain was misclassified as ST3434 and one novel MLST variant of the outbreak was identified as the

outbreak sequence type (ST93). Over four subpanels (n = 48, 1000 reads), accuracy was 94.01%, with binary labels (see above) at an accuracy of 95.24%, precision of 98.18% and sensitivity of 84.81%, while multilabel features (see above) had an accuracy of 85.41%, precision of 95.83% and sensitivity of 85.41% (Table 2).

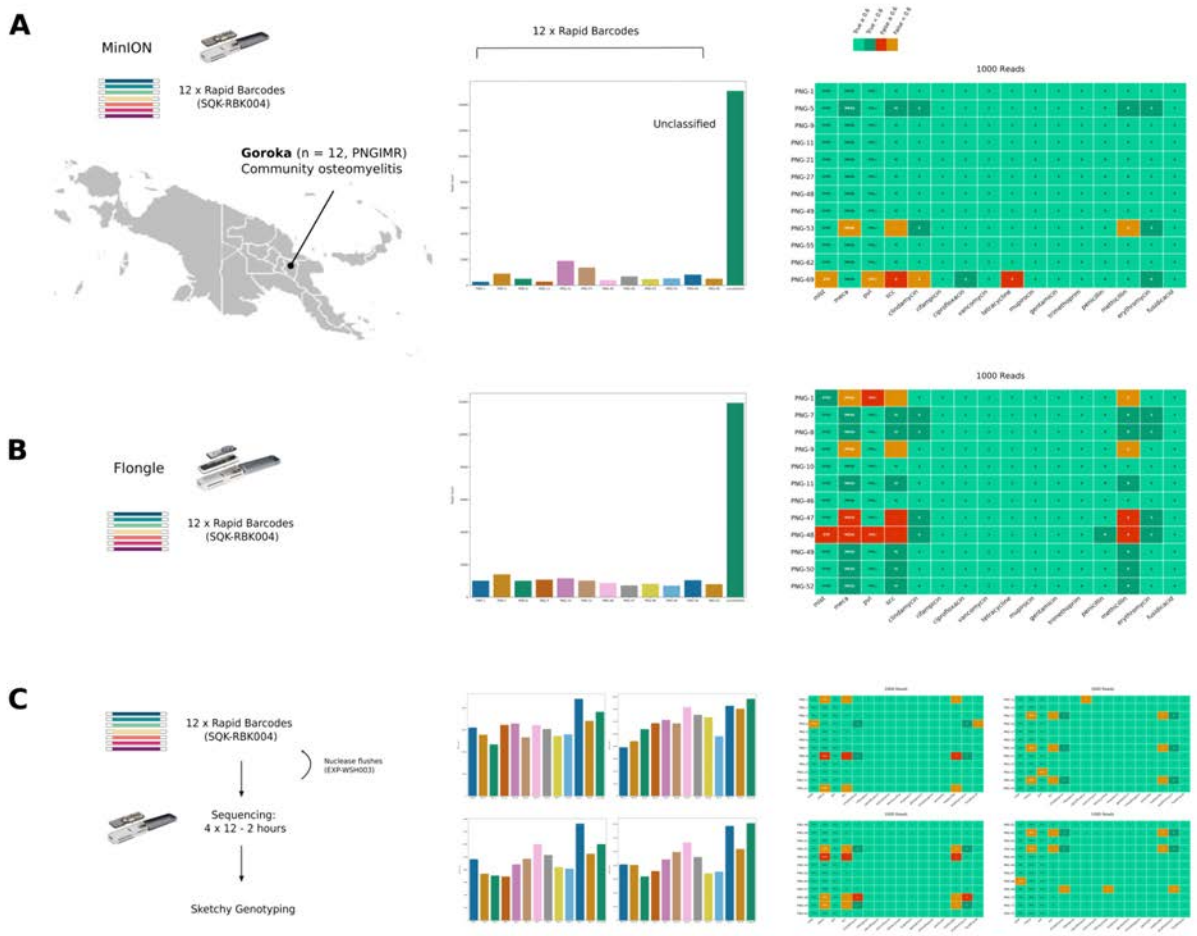

Fig. 4: Multiplexing experiments at the Papua New Guinea Institute for Medical Research **(A)** and extended multiplexing experiments on Flongle **(B)** and MinION (48 strains on four successive panels) **(C)**. Left panel shows a representation of the experiment, middle panel shows the barcode distribution of each sequenced barcoded run (seagreen is unclassified) and right panel shows the genotype truth heatmaps, where colors indicate true or false predictions with strong support (light seagreen, dark red) or low support (dark seagreen, orange). A large number of unclassified barcode reds were due to a malfunctioning instrument during library preparation in **(A)** and due to including all reads above Q5, contributing to the low-quality, unclassified barcodes in the Flongle experiment **(B)**.

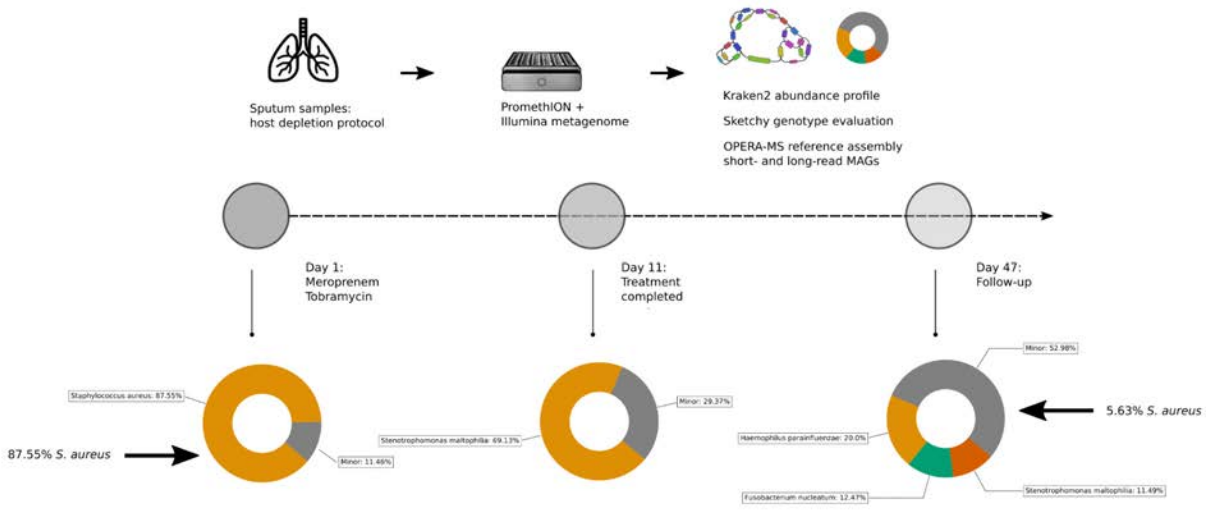

Fig. 5: Hybrid metagenome sequencing and taxon abundance profiling of host-depleted sputum of a cystic fibrosis patient, undergoing antimicrobial therapy with meroprenem and tobramycin over a two week time-frame. Samples were taken at three different timepoints (Fays: 1, 11 and 47) and sequenced after host-DNA depletion on PromethION (Duarte et al. *in preparation*}. Donut plots show abundance profiles from Kraken2 and Bracken from nanopore data, which indicate that *S. aureus* dominated at the start of treatment (87.55%) was replaced by dominant *Stenotrophomonas maltophilia* and complete disappeared after treatment (36 reads total mapping to \textit{S. aureus} on day 11) and resurged to minor abundance in the follow-u sample (5.63%).

Strain resurgence in a cystic fibrosis patient

Lastly, we evaluated the application of *Sketchy* in patient metagenomes, using data generated directly from the sputum of a cystic fibrosis patient undergoing antimicrobial therapy (Fig. 5, Duarte et al. *in preparation*). Intravenous meropenem and tobramycin were started on the 1st day and completed on the 14th day; sputum samples were taken on the 1st, 11th and 48th day. After multiplex sequencing of host-depleted sputum on PromethION (12 barcodes) and validation sequencing on Illumina platforms with metagenome libraries, abundance profiles from nanopore reads for each time point were created with *Kraken* and *Bracken*. Our data shows that *S. aureus* was the dominant member of the sputum community on the 1st day (61,962 reads, 87.55% abundance), had disappeared nearly entirely on the 11th day after treatment (36 reads)

and reappeared on the 48th day at low proportions of the community (2,364 reads = 5.63% abundance) (Fig. 5). We constructed hybrid metagenome-assembled genomes (MAGs) on the first and last sampling point, showing that resurgence after successful treatment was caused by the same strain (ST8-MSSA, pan-susceptible). We then used *Sketchy* to assess whether the strain could be successfully genotyped from nanopore reads without metagenome assembly, particularly at the last time point (day 48), where metagenome assembly was not possible with nanopore data alone. Results show that the correct genotype (ST8-MSSA-PVL, pan-susceptible) was recovered at the terminal sampling points (Fig. 6), including indications for the eradication of the strain at the second time point (ST8, 36 reads) when we assessed diagnostic plots of the streaming function (Fig. S7). In combination with reference genotypes from hybrid MAGSs, this constitutes strong evidence that *Sketchy* is capable of genotyping low-abundance pathogens from metagenomic samples. Our data supports the notion that antimicrobial therapy was successful, leading to limited resurgence of the initially dominant strain after treatment ended.

2.3.3. Discussion

In this study, we explored the possibility of using the heuristic principle of genomic neighbor typing (*46*) for genotype inference using species-wide whole genome reference sketches of *S. aureus* (n = 38,981) and *K. pneumoniae* (n = 8,149). We developed *Sketchy* , a client for genomic neighbor typing through min-wise shared hash queries in Mash, and a genomic neighbor typing algorithm that allows for processing nanopore read streams. We further showed that *Sketchy* can be used for heuristic genotyping of bacterial outbreaks using a dual-panel multiplexing protocol and less than 1000 reads per barcode (MinION). Sketchyeven performs well in difficult sequencing scenarios, which we demonstrated by in situ sequencing in Papua

New Guinea (with failing equipment), barcoding on Flongle adapters and by sequencing 48 strains in a quad-panel multiplex protocol on a single flow cell (MinION). In addition, we showed that the algorithm can be used on metagenomic data, successfully identifying a *S. aureus* strain resurgence after antimicrobial therapy in a cystic fibrosis patient from less than 1000 reads on the PromethION (≈ 2 minutes - 2 hours of sequencing, depending on species abundance at the sampling points).

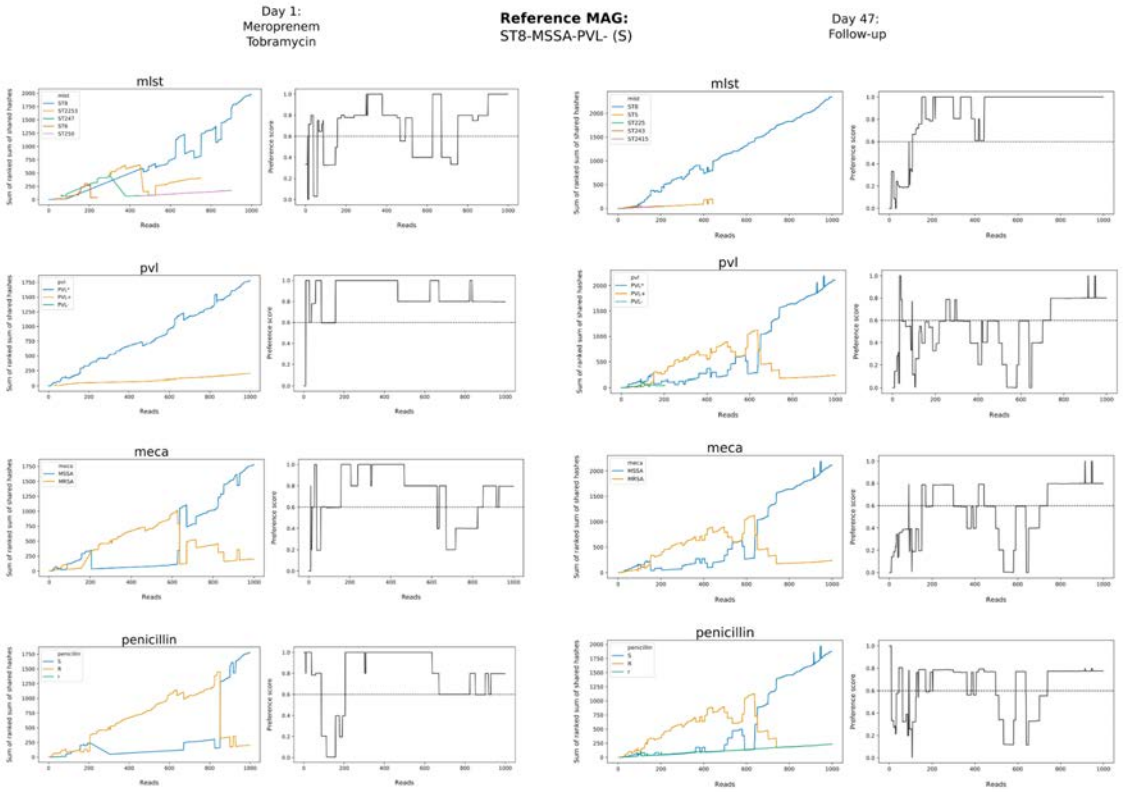

Fig. 6: *Sketchy* evaluation on a maximum of 1000 nanopore reads identified as *S. aureus* from the cystic fibrosis samples on day 1 (start of treatment with meroprenem and tobramycin) and 47 (follow-up). Metagenome-assembled genomes (MAGs) of the *S. aureus* strain could not be constructed from nanoproe data alone, but with matching Illumina data and hybrid-MAG assembly using OPERA-MS, the strain was identified as ST8-MSSA-PVL- (pan-scusceptible). Diagnostic plots of ranked sums of shared hashes and genotype predictions from *Sketchy* predict the correct lineage and genotype, including stable predictions from preference scores (> 0.6, horizontal lines); selected examples of individual genotype diagnostics are shown here (all genotypes can be found in Fig. S5).

Our main focus was to address the issue of database representation raised by Břinda et al. (*46*). Genotype representation in the reference database is a key feature of genomic neighbor typing, because lineages and genotypes not present in the database will inevitably be misclassified, and because insufficient representation of species-wide genomes would make genomic neighbor typing difficult to use in situations, where unknown lineages and genotypes might occur. We attempted to address the limitations of the original, phylogenetically informed approach (RASE) by scaling our reference genome collections to all known, high-quality species genomes at the time (2019) and assessed performance on an independent outbreak dataset of S. aureus infections in Papua New Guinea and Far North Queensland (*38*, *39*). In this context, our extensive nanopore dataset (n = 158) with matching Illumina data constituted a completely independent validation set, as no *S. aureus* genomes from these regions had ever been sequenced before. While the original implementation of genomic neighbor typing successfully demonstrated the concepts of the heuristic, lineage validation failed in several clinical isolates without representative sequence types in the RASE database. While the majority of isolates in this study belonged to the outbreak sequence type (ST93) several other sequence types were identified. By virtue of including all known strains at the time, their background lineages were included during database construction using our automated Nextflow (*14*) pipelines, including a distinct ST243 neonatal outbreak, as described previously (Chapter 2.1). Database construction workflows can parse and update reference sketches using all available genomic data and standardized processing pipelines with quality filters to construct reference assemblies and conduct genotyping from large short-read sequence collections. We note that reference database construction can be automated, so that public archives can be surveyed periodically and new genomes and genotypes continuously integrated into databases. In addition, a large assembly index for all bacterial short-read sequence data in the ENA including extensive quality controls and genotypes has been released recently (n ≈ 600,000) from which additional *Sketchy* genotype indices and reference sketches can be constructed (*180*). Notably, despite the size of

the S. aureus reference sketch (n = 38,981) the uncompressed *Sketchy* database (*k* = 15 and *s* = 1000) was only 162 MB and 2.6 MB compressed. Improvements in min-wise shared hash computation performance, for example with *sourmash (179)* or *ppsketch (181)*, can be made to accommodate even larger genome collections. Further optimization of the genotype representation in the species-wide reference sketches through genotype-informed demultiplexing are also feasible (including balancing of classification performance against the consensus-seeking rank parameter, see Methods).

We further note that our *Klebsiella pneumoniae* reference sketch was small (n = 8,149), and contained insufficient genotypes to account for a large complement of genotypes associated with carriage of mobile genetic elements (including several MDR resistance plasmids) and some highly variable traits like serotypes (*32*, *182*). We also observed a concentration of classification failures in *S. aureus* in the important SCC*mec* associated features. We suggest there are two potential failures associated with genomic neighbor typing and mobile element related features (including many resistance or susceptibility calls): first, inherent to the heuristic principle of turning a geno- (or pheno-) typing problem into a database problem is database representation as pointed out by Břinda et al. (*46*). Despite including all possible *S. aureus* genomes available at the time (2019) we still observed genotyping failures in SCC*mec* related elements on the level of lineages, where either the lineage itself (and therefore the genomic neighborhood) was insufficiently represented (absent or few isolates), and on the level of within-lineage genotype variation, where specific genotypes (such as tetracycline resistance or SCC*mec*) or combinations of genotypes were not available. In principle, these genotypes cannot be correctly inferred, but often a close match is found if relatives of the lineage or clonal complex are otherwise represented. Most notably, these occurred in misclassifications of ST81 isolates from the PNG outbreak (a sequence type variant of ST1) and the nearest sequence type in the clonal complex (CC1:ST1) was called instead. This is likely a result of an imbalanced representation in

the database, where ST81 was included with 19 genomes and the close relative ST1 with 1,089 genomes, outcompeting other closely related genomes in the consensus step of the streaming algorithm, when computing the ranked sum of shared hashes per feature (Methods). In addition, several novel single nucleotide MLST variants of ST93 were typed as ST93, conforming to their closest match in the database. A second problem with typing mobile element related features appears to be inherent to the resolution obtained with our Mash based approach at a database sketch size of $s = 1000$. Despite sufficient representation in the database ST93-MSSA and -MRSA genotypes (within-lineage genotype variants) were often miscalled (Fig. 3). As the lineage recently emerged (on bacterial evolutionary timescales), small genetic distances distinguish the ancestral ST93-MSSA from the genotype acquisition of SCC*mec* in the emergent ST93-MRSA genotype. However, because *Sketchy* databases - unlike RASE - does not use a phylogenetically informed database or method, it appears that resolution and discriminatory power can be low when confronted with a relatively homogenous within-lineage population structure. Larger sketch size ($s = 10000$) resulted in negligible improvements to classification at a cost of disk space and memory size (1.5 GB for S. aureus, 34 MB compressed) and considerably slower queries against the reference sketch (≈ 6- 10x slower, data not shown). Future improvements to the method might incorporate a combined hierarchical approach, where Sketchy is used for inference of the genomic neighborhood in a large species-wide index, and a phylogenetically informed, smaller neighborhood-specific database with the RASE methodology is used for precise placement of genotypes, even within relatively homogenous populations. *Sketchy* could therefore be considered to be complementary to RASE, aiming to support large-scale outbreak surveillance with the limitations of database representation in mind. We should note that the summary metrics estimates presented here (Tables 1, 2, S1) for *S. aureus* are artificially inflating wrong calls, as the SCC*mec* related features (methicillin resistance, *mecA* and cassette type) are inherently linked and measure the same genotype.

Genomic neighbor typing with *Sketchy* should not be used for clinical decision making (i.e. antibiotic prescription) unless a more diverse *S. aureus* (or other species) validation dataset can be assembled and the limitations of the database reference sketches more precisely defined. While the preference score seems to be a reliable indicator for lack of support, we also observed 'switches' of predictions to the correct genotype as soon as sufficient reads were processed (Fig. 6). Read order seems to play a role in predictions less than about a thousand reads (Fig. S5). It should also be noted that in particular the resistance phenotype predictions based on *Mykrobe* were largely correct, a pattern which was driven by the over-representation of outbreak sequence type clones in our dataset, which were susceptible to most antibiotics. Lower performance was achieved in methicillin resistance predictions, associated with difficulties typing SCC*mec* associated genotypes. Because we derive genotypes from other genotype classifications (based on assemblies or reads) it should also be noted that classification with *Sketchy* can only achieve classification performance of the underlying genotyping methods. While we were able to rapidly and correctly track a metagenome-assembled confirmed strain in a patient sample - confirming a minor resurgence and the effectiveness of antimicrobial therapy - we suggest to be cautious with metagenomic data, as multiple strains may be contained in a complex sample and our approach has not been tested on multi-strain mixtures. It is feasible that the ranked sum of shared hashes scores may be able to deconvolute multiple strains in a sample, but in the current version of *Sketchy* a singular prediction of lineages and genotypes is made. Ultimately, we did not focus on clinical utility, but instead demonstrated that genomic neighbor typing is a viable approach for large-scale genotype surveillance in bacterial pathogens and remote outbreak scenarios using community-associated *S. aureus* as a model for outbreak scenarios, and *K. pneumoniae* for validation of applications across other bacteria, highlighting both the strength and limitations of genomic neighbor typing. Genomic neighbor typing - based entirely on other genotyping tools - implements a 'heuristic meta-typing' approach in which any other genotyping tool (as well as

phenotypes and other genome-informed data) can be transformed into a genomic neighbor typing tool, capable of operating on nanopore read streams - albeit trading one problem (rapid, direct evidence for genotypes) for another (database representation).

2.3.4. Materials and Methods

Outbreak sampling and reference sequencing. We collected isolates from outbreaks in two remote populations in northern Australia and Papua New Guinea (Fig. 1). Isolates associated with paediatric 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 neighbouring Eastern Highlands province town of Goroka. We supplemented the data with MSSA isolates associated with severe hospital-associated infections and blood cultures in Madang (Madang Province) (n = 8) and Goroka (n = 12). Isolates from communities in Far North Queensland, including metropolitan Cairns, the Cape York Peninsula and the Torres Strait Islands (n = 91), were a contemporary sample from 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 LB broth at 37°C overnight and stored at -80°C in glycosol and LB. Illumina short-read data from the ST93 lineage (*42*) included in this study were collected from the European Nucleotide Archive (Online Supplementary Tables).

MinION outbreak library preparation and sequencing. 2 ml of LB broth was spun down at 5,000 x g for 10 minutes and after removing the supernatant, 50 ul of 0.5 mg / ml lysostaphin were added to the tube and vortexed. Cell lysis was conducted at 37°C for 2 hours with gentle shaking followed by a *proteinase K* digestion for 30 mins. 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 ul of nuclease-free water, quantified on Qubit, and DNA was stored at 4°C until library preparation. Library preparation was done using approx. 420 ng of DNA and the rapid barcoding kit with 12 barcodes (ONT, SQK-RBK004) as per manufacturer's instructions, with the exception of conducting bead cleanup steps. DNA was quantitated using Qubit 4.0 (Thermo Fisher Scientific), purity determined with a NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific). Basecalling was done using the PyTorch *Bonito* R9.4.1 DNA model, run on a local NVIDIA GTX1080-Ti or a remote cluster of NVIDIA P100 GPUs. Sequence runs were conducted with 2 x 12 barcoded (SQK-RBK004) isolates per flow cell in two consecutive 18-24 hour 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 (162)* and our analysis of a single library panel (FNQ-2) sequenced on a previously used flow cell with a human library. Sequencing runs were managed on two MinIONs and monitored in *MinKNOW* > v20.3.1. In addition, we used extra libraries from the outbreak sequencing to test a faster protocol, in which four libraries were sequenced on the same flow-cell (with intermediate nuclease flushes) with a runtime of 2 hours per library (Table 2).

MinION and Flongle multiplexing experiments. To demonstrate that genotyping is possible on site in Papua New Guinea, we sequenced an additional 12 *S. aureus* outbreak strains at the Papua New Guinea Institute of Medical Research (PNGIMR) in Goroka. We replicated the simple QIAGEN extraction and rapid library sequencing protocol described above, unknowingly using a malfunctioning heat block in the library preparation (SQK-RBK004). Finally, we prepared a multiplex run for a Flongle experiment. *Staphylococcus aureus* glycerol stocks were inoculated in Tryptic soy broth (TSB) and grown overnight at 37°C, 180 rpm. DNA was extracted from 8 ml of overnight culture via pelleting cells at 12,000 rpm for 2 minutes. Cells were resuspended in PrepMan™ Ultra Sample Preparation Reagent (ThermoFisher Scientific) and

Lysing Matrix Y beads (MP Biomedicals). Isolates were incubated at 95°C for 15 minutes and cells further lysed via a TissueLyser LT (Qiagen) at 6.5 m/s for 60 seconds similar to previously described (*102*). Extracts were centrifuged at 13,000 rpm for 10 minutes. Supernatant was removed and mixed with 3M sodium acetate (pH 5.5), ice-cold 100% ethanol (0.3:0.03:0.67 ratio) and DNA was precipitated for 3 hours at -20°C. DNA was pelleted at 13,000 rpm for 15 mins (4°C), washed with 70% ethanol and resuspended in ultrapure water. High-molecular-weight (HMW) DNA was isolated via the MagAttract HMW DNA Kit (Qiagen) as per manufacturer's instructions. Briefly, this included a protein digest with proteinase K for 30 minutes at 56°C (900 rpm) and an RNAse A (0.4mg) treatment for 10 minutes at room temperature. HMW DNA was further purified using Agencourt Ampure XP (Beckman Coulter Australia) beads (1:1 ratio). Libraries were prepared using the ONT Rapid Barcoding (SQK-RBK004) kit with an input of 200ng of HMW DNA for each isolate. The library was sequenced on an ONT Flongle FLO-FLG001 flowcell for 24 hours. All runs in this section were called with *Guppy* v4.6 R9.4.1 DNA high accuracy models (HAC).

Genotyping of reference sketch assemblies. Short-read reference assemblies were constructed for the *S. aureus* strains collected and sequenced in this study. *Fastp (128)* was used to trim adapter and low quality sequences. *Shovill* with *SPAdes* (*183*) (https://github.com/tseemann/shovill) was used to assemble reference genomes. Assemblies were genotyped with *Kleborate (184)* for *K. pneumoniae* and *SCCion* (https://github.com/esteinig/sccion) for *S. aureus* which wraps common, assembly-based genotyping tools including *mlst* and *abricate* (https://github.com/tseemann) with the ResFinder (*130*) and VFDB (*165*) standard databases, as well as *Mash* based SCC*mec* typing against the reference database from SCC*mec*Finder (*44*, *131*). We also used M*ykrobe (132)* on trimmed reads to obtain gold-standard *in silico* resistance profiles for *S. aureus*. Metagenomic reads from the Zymo community (*185*) and from a stool microbiome by Legget et al. (*176*) were classified

with *Kraken2* (default database) (*186*) and extracted for prediction. DNA sequences of multidrug resistant *K. pneumoniae* were collected from Pitt et al. (*182*). Read statistics and genome coverage were assessed with *nanoq* (https://github.com/esteining/nanoq) and CoverM (https://github.com/wwood/CoverM). Average coverage was assessed against outbreak reference strain JKD6159 (*65*) for *S. aureus* and HS11286 for *K. pneumoniae (187)*.

Construction of species-wide reference sketches. For species-wide sketch construction we downloaded all publicly available sequence data for S. aureus and K. pneumoniae from the European Nucleotide Archive. We first obtained paired-end Illumina sequence reads from genomic sources and whole genome sequencing experiments, pre-filtered by removing sequence read files with an estimated coverage below 50x and greater than 700x. This resulted in the collection of 44,626 S. aureus genomes (November 2018) and 15,026 K. pneumoniae (March 2019). Nextflow (*14*) pipelines in the *np-core* collection (https://github.com/np-core) were used for quality control, genome assembly and species-specific molecular typing on a high-performance compute cluster (James Cook University). Raw reads were first trimmed for quality and adapter sequences with *Trimmomatic (188)* and classified taxonomically with Kraken2 (*186*) (read-wise classification at species level). We used a highly conservative quality control strategy by discarding isolates with > 2% read contamination from any species other than the target species (contaminated) and samples that did not have the largest percentage of reads assigned to the target species (misidentified) or had < 80% assigned to the target species (not pure). Read sets passing these filters were assembled with *Skesa* (*129*) as part of *Shovill*, which combines downsampling to 100x coverage with additional pre-filter steps before assembly (https://github.com/tseemann/shovill).

Genotype indices linking each genome to its genotype were created from read and assembly typing as part of the pipeline. In the predictions, we refer to each categorical data that

designates a particular characteristic (gene, susceptibility, allele) as ‘genotype feature’ and each possible state of the feature (e.g. allele, gene, presence or absence, resistant or susceptible) as a 'feature trait'. We used *SCCion* (https://github.com/esteinig/sccion) for *S. aureus* which wraps common, assembly-based genotyping tools including *mlst* and *abricate* (https://github.com/tseemann) with the ResFinder (*130*) and VFDB (*165*) standard databases, as well as *Mash* based SCC*mec* typing against the reference database from SCC*mec*Finder (*44*, *131*). Antimicrobial resistance phenotypes for 12 antibiotics (ciprofloxacin, clindamycin, erythromycin, fusidic acid, gentamicin, methicillin, mupirocin, penicillin, rifampicin, tetracycline, trimethoprim and vancomycin) were inferred directly from reads with Mykrobe v.0.6.1 (*132*). *S. aureus* predictions with Mykrobe are comparable to sensitivity and specificity obtained with culture-based phenotyping methods. For the *K. pneumoniae* genotype index, we used *Kleborate (184)* to define MLST, presence of the yersiniabactin virulence factor (ybt) carried on the *K. pneumoniae* integrated conjugative element (ICEKp), lipopolysaccharide K- and O-locus serotypes, hyper mucoidy inducing genes carried (*rmpA, rmpA2*), 12 antimicrobial classes and 4 subclasses of beta lactam resistance (e.g. Fig 2). In the case of antimicrobials, hyper mucoidy genes and yersiniabactin, multiple alleles or genes encoding the same feature were collapsed into binary presence or absence feature values.

Only genomes passing all pipeline steps and genomes for which lineages (MLST) could be unambiguously identified were retained. This resulted in a total of 8,149 K. pneumoniae and 38,981 S. aureus, genomes used in constructing the reference sketches. MinHash sketches of the assemblies for each species were created using the default sketch size ($s = 1000$) and $k = 15$ which increases sensitivity for noisy uncorrected nanopore reads (*44*). Smaller k-mer sizes used in sketch construction also have the advantage that sketches are stored as 32-bit integer hashes rather than 64-bit integer equivalents, thus reducing the total memory footprint of the sketches.

Sketchy streaming algorithm. *Mash* first queries the reference sketch and computes the number of shared hashes (x_j^i) for each read i and each genome j indexed in the sketch. Sketchy then computes the cumulative sum of shared hashes h_j^i with each index in the reference sketch from the output stream of *Mash* up to and including the terminal read (i'):

$$h_j^i = \sum_{i'=1}^{i} x_j^{i'}$$

Sketchy calculates a ranking r of h^i at each read, where r_j indicates the index which is the j^{th} ranked score, (e.g. $\{r_1 .. r_{10}\}$ are the indices of the top 10 scores). *Sketchy* then calculates an aggregate score by linking the sum of shared hashes h^i to pre-computed genotype features for each genome index in the sketch. For each feature, such as MLST, each genome in the reference sketch is assigned trait t_k (e.g. membership of the k^{th} MLST group) via an assignment matrix M, where $M_{k,j}$ is 0 or 1 indicating whether the j^{th} reference genome has the k^{th} value of the feature. Each column of M has a 1 in precisely one row of this matrix, reflecting the restriction that each genome belongs to exactly one class. *Sketchy* uses the first R ranks to calculate the feature scores (default $R = 10$)

$$g_k^i = \sum_{j=r_1 .. r_R} M_{k,j} * h_{j'}^i$$

which can also be written more compactly as

$$g^{i} = M_{r_1..r_R} * h^{i}_{r_1..r_R}$$

Genotype feature aggregation into the evaluation score thus represents the consensus of each feature trait over other features values at any given read in the sequence in the ranking matches. We adopt the lineage preference score (p) by Břinda et al. (*46*) which at each read measures the relative preference in the cumulative sum of ranked sums of shared hashes of the top ranking feature (f) over the alternative, second ranking feature (u) so that:

$$p = \frac{2f}{(f+u)-1}$$

For prediction and evaluation of our sequence runs, we report the total shared hashes score by feature (t_k) at the read threshold and its preference score below or above an arbitrary threshold (0.6) as described by Břinda et al. Time until prediction was calculated by the difference of the read at which prediction is performed to the starting read of the run. For visual inspection, we plot g^{i} for each feature as competing lines, as well as the preference score values as line crossing the preference score threshold over the course of the read stream (e.g. Fig 2C).

Discussion

In this work, we set out to explore genomic epidemiology and transmission dynamics of community-associated *Staphylococcus aureus* in northern Australia and Papua New Guinea. Our primary aim was to describe the evolutionary history and transmission dynamics of isolates from the remote highland provinces (*38*) and communities in Far North Queensland (*39*). We further reasoned that emerging nanopore sequencing technology would be capable of decentralized whole genome sequencing, allowing for genome-informed surveillance of pre-eminent lineages circulating in the region. However, several technical challenges had to be addressed, including the adoption of birth-death skyline models (*30*) for lineage-resolved bacterial datasets (*37*, *116*) and overcoming deficits in single-nucleotide polymorphism (SNP) accuracy and precision (*43*) adopting machine learning approaches for transmission inference using low-coverage (and low-cost) nanopore sequencing data .

In the first chapter, we used traditional, high-resolution Illumina data to reconstruct the evolutionary history of *S. aureus* strains sampled from remote community-associated outbreaks in Kundiawa (Simbu Province) and Goroka (Eastern Highlands Province) (*38*), as well as from

routine pathology at Cairns hospital, covering the Cairns and Hinterland, the Cape York Peninsula and Torres Strait Islands (*189*). We used whole-genome sequencing data to reconstruct the evolutionary history of the first S. aureus genomes from Papua New Guinea, and discovered that a paediatric osteomyelitis outbreak in the remote Highlands Provinces was caused by the Australian clone ST93-MRSA-IV. Multiple lines of evidence support a wider distribution of ST93 in PNG, including two discernible introductions, sustained and long-term transmission of the outbreak since the early 2000s, occurrence of two MLST allele variants, and a heterogeneous pattern of dissemination in the remote highland towns. It remains unclear to what extent ST93-MRSA-IV has disseminated in PNG. Our data further show that ST93-MRSA-IV is widespread in Far North Queensland and is likely the cause for the increasing rates of MRSA observed in FNQ communities over the last decade. It is unclear what is driving the local persistence and evolution of the ST93-MRSA-IV genotype in PNG, and reservoirs in the community remain to be investigated. Data on antibiotic consumption in Simbu Province or Eastern Highland Province was not available. While antibiotic stewardship may play a role in the dissemination of ST93 in FNQ, and ST772 in the Islamabad-Rawalpindi metropolitan area (*120*), sustained circulation of virulent and transmissible clones in remote settings like PNG may also have been a result of historical transmission opportunities from the Australian East Coast after the emergence of ST93-MRSA-IV, as well as existing strain diversity and competitive interactions in the highlands, even in the absence of widespread antimicrobial consumption. Our work here has contributed to changes in antibiotic treatment guidelines for osteomyelitis at Kundiawa General Hospital, where therapy has now been shifted from oral penicillins and other beta-lactam antibiotics, towards ciprofloxacin, gentamicin and erythromycin.

We applied birth-death skyline models to the ST93 lineage data and demonstrated that increases in the effective reproduction number (R_e > 1) indicating epidemic growth of the population) were associated with expansion of drug-resistant ST93-MSSA in the Northern

Territory, expansion of methicillin resistant ST93-MRSA-IV on the East Coast of Australia, and transmission to Papua New Guinea and into Far North Queensland. This prompted us to investigate whether similar changes could be found in other community-associated lineages. We reasoned that these 'signatures of epidemic growth' - including surges in the effective reproduction number and establishment of sustained transmission (Re > 1 over time) - were indicative of a convergence between the acquisition of resistance and the ability of subsequently resistant genotypes to recruit into a new epidemiological niche and spread in regional population centers such as on the Australian East Coast (ST93), the Indian subcontinent (ST772) or Europe (e.g. ST80, ST152, ST1). Phylogenetic evidence points to ancestral or symplesiomorphic MSSA strains of community-associated lineages circulating in geographically distinct host populations, including in Africa (ST80, ST152), Europe (ST8) and South-Eastern Europe (ST1) and amongst Indigenous communities in remote northern Australia (ST93) before the emergence of resistant clades in geographically close host populations. Using additional samples of the emergent multidrug resistant ST772-MRSA-V clone from Pakistan (n = 59), we show that epidemic growth has likely occurred following introduction into the community around metropolitan Islamabad, associated with a foodborne outbreak of table-eggs and linking to ST772 infections in Norway (*93*). Besides evidence from ST8-MRSA (USA300) introductions to France and Africa in available public data this constituted further evidence (in addition to the ST93 introduction in Papua New Guinea) that resistant, community-associated genotypes are capable of establishing epidemic growth after transmission into a new setting, and apparently establish long-term transmission. Dated phylogenetic trees from the birth-death skyline models indicated that the introduction in Papua New Guinea occurred soon after the emergence of the resistant ST93-MRSA-IV genotype on the Australian East Coast.

Finally we used birth-death skyline models on other available community-associated, lineage-resolved whole genome datasets and showed that surges in the effective reproduction

number coincide with the acquisition of antibiotic resistance in all lineages, although some variation between and delays between the estimated MRCA of the resistant lineage and surges in R_e were observed in some clones (ST1, ST80), indicative of local epidemiological effects that certainly appear to play a role in the emergence of resistant MRSA clones. Estimates from Bayesian phylodynamic models indicate that sustained transmission (R_e > 1) following importation has not only occurred in PNG and FNQ (ST93) but also in Pakistan (ST77-MRSA-V), several African countries (ST152-MSSA, ST8-USA300), South America (ST8-USA300 COMER variant) and Europe (ST8-USA300). It therefore appears that community-associated MRSA lineages are able to establish sustained transmission after dissemination of resistant genotypes. Phylodynamic models predicted lineage- and clade-specific infectious periods $(1/\delta)$ that suggest prolonged durations of infection over several years in concordance with long-term cohort studies with lineage-resolved data. Variation in our model estimates could reflect differential local modes of persistence in the host or community (*140*, *190*), and is susceptible to factors that we were not able to explicitly model, including access to healthcare services and treatment amongst others. It should be noted that the lineage-wide averages of infectious periods may not reflect the considerable heterogeneity in carriage duration that likely arises from the distribution of permanent and transient carriers across the population (*145*, *146*). Our data further suggest that changes in transmission dynamics (R_e) can occur without additional genomic changes either after incursion into a new population (e.g. ST80 European expansion; introductions of ST93 clades into PNG, FNQ and NZ) or following a delay of several years after introduction (e.g. ST772 after SCC*mec*-V (5C2) fixation, ST59 expansion in Taiwan). It is feasible that delayed changes in R_e may be indicative of local competitive interactions with prevailing lineages or changes in human population dynamics (e.g. in healthcare or social policies, travel and immgration policies, opening of markets and borders) that drive further dissemination once established in the host population. In one case, a sharp increase in R_e occurred nearly a decade after the MRCA of the resistant

European ST1-MRSA clade, with the implication that the emerging genotype circulated undetected in South-East Europe - likely in Romania where the first samples originate (*50*, *90*, *91*) - before its emergence across Europe. It remains difficult to disentangle the epidemiological factors that contribute to the estimate of the effective reproduction number (R_e). This includes genotype fitness, local adaptive pressures from antibiotic treatment or environmental contamination, contact networks, age-specific mixing and treatment accessibility, all of which have not been considered in phylodynamic transmission models. Progress is being made on integrating disease incidence from public health datasets with phylodynamic modelling but this addresses uncertainty estimates, rather than providing new insights into the epidemiological drivers interacting with genotypes, ultimately resulting in the emergence of community-associated lineages. The accumulated genomic and epidemiological data, consisting of well documented collections of MSSA and MRSA strains of these lineages, is a valuable resource made possible by an enormous international effort of the *Staphylococcus* research community and decades of collecting strains in freezers. It is imperative that further collaborations and efforts to collect and store strains over the coming years, or transition as quickly as possible to generating real-time data on bacterial pathogens, similar to what we are doing now for SARS-CoV-2 variants and surveillance efforts. In this context it is also critical to simultaneously collect sufficient metadata (at least the date of collection) to inform phylodynamic analysis, and additional geographical information will be useful with integration of mobility and travel data.

Further computational and experimental research is also necessary to untangle the effects of canonical mutations in potential colonization and host immune evasion factors - non-synonymous mutation in relevant genes, including *fbpA*, *tet* and *plc* - which we have found at the divergence of ST772-A (*40*). Affected genes bear resemblance to colonization factors involved in skin and abscess survival though to contribute to the increased transmission

potential of ST8-MRSA-USA300 strains carrying ACME and COMER elements. It is conceivable that these constitute pre-adaptations to increased transmission potential (through mechanisms of cutaneous persistence and host immune evasion) that have evolved in response to favourable conditions for transmission in host populations in which the ancestral community-associated MSSA circulated (and continue to circulate) before the acquisition of resistance and emergence in nearby host population centers. It is clear from epidemiological data on multiple of the lineages analysed here (*40*, *42*, *50*, *75*, *80*–*84*), that ancestral strains circulated in notably similar host sub-populations, namely the poor and marginalized segments of our society, located remotely or geographically distinct from the urban, industrialized population centers in which resistant strains emerged, and without adequate access to the healthcare system, creating conditions favourable to the transmission of community-adapted pathogens.

While we have here successfully unraveled some of the drivers behind the emergence of community-associated lineages, including a likely association with sociodemographic factors and a divide between remote or geographically distinct host sub-populations and urbanised or industrialized population centers in which resistant clades emerged, two other aspects of the "life-cycle" of community-associated *S. aureus* lineages remain perplexing: where did MSSA lineages originate and why did the resistant MSSA and MRSA lineages only emerge, seemingly convergently in the latter half of the 20th century? One hint is given by one of the 'outlier' lienages, one of the few community-associated lineages which has sufficiently representative data to allow their tracing back to its emergence of ST8-MSSA in 19th century Europe (*82*). Capsule mutation (indicative of immune system modulation) enabled its initial emergence in Europe of the 19th century - in a pre-antibiotic era - which shared many sociodemographic dynamics with populations in which community clones emerge today, such as wide-spread poverty, domestic overcrowding and poor hygiene and healthcare, especially in

population-dense increasingly industrialized cities. Its spread across Europe resulted in the eventual emergence of ST8-MRSA-USA300 in the Americas which then successively acquired PVL, another capsule mutations *cap5E*, ACME or COMER and SCC*mec*-IV in the later half of the 20th century. While data is extremely sparse on the ancestral transmission dynamics of the MSSA clades, one perspective to be considered is that they are intricately linked. Supporting observations can be made from the Australia lineage emergence of ST93-MSSA. In our birth-death skyline estimates of the ST93 lineage origin (*T*) the difference to the MRCA (the root branch length) is less than a year. While the parameter is the least stable of the model in general, our estimates show narrow confidence intervals and good support for estimates of *T* in the ST93 lineage, owing to the dense sampling coverage of ancestral (and contemporaneous) ST93-MSSA strains from the Northern Territory. However, such a close origin of the lineage, to the MRCA of the ST93-MSSA progenitor strains circulating in Indigenous communities of remote northern Australia, indicates a rapid establishment and genetic bottleneck, unlikely to have occurred in carriage of long-established human populations in this region, as suggested in our previous work (*42*). In combination with unusual patterns of intra-specific recombination we observed across the ST93 genome, where phylogenetic evidence is indicative of recombination with deeply diverging unsampled *S. aureus* lineages in the species phylogeny, this sudden establishment in the human population may have occurred from a zoonotic spillover, most likely mediated by wild animals. Wildlife can be a source of zoonotic *S. aureus* infection, but is often understudied and overlooked, but potentially highly relevant geographically remote and less urbanised populations, more likely exposed to wildlife-human interfaces. While estimates of R_e closer to the root usually have large confidence intervals due to lack of samples, the estimate for the ST93 lineage has a confident and high estimate ($R_e > 2$) in the 1980s when the lineage originated, followed by a drop in transmission and subsequent emergence of the ST93-MSSA clade in the Northern Territory. This pattern may be consistent with a zoonotic spillover at the origin of the lineage, where the lineage establishes human transmission in remote Indigenous

communities (initial spike) and then shortly after establishes sustained transmission again (R_e > 1) in the remote and isolated communities, before the emergence of the MRSA clade on the Australian East Coast in the 1990s. It is now well established that bi-directional transmission of ST93 can occur between pig-workers and farm-animal (*72*). We have considered that it may be one of the reservoirs for the strain in the Papua New Guinean population, where pigs are highly valued and ubiquitous domestic animals in frequent contact with children, which comprised the vast majority of osteomyelitis cases in Kundiawa and Goroka.

Considering the wide host range of *S. aureus* and frequent spillover events between human and animal populations (*53*, *191*) it is notable that transmission between animals and humans, and isolation from animal hosts have been reported for all lineages studied here, including ST772 from goats in India (*192*), ST1 from wild rooks and domestic mammals in Europe (*91*), ST59 from pigs in rural China and Western Australia (*72*, *193*), ST8 from bovine transmission (*194*) and ST80 and ST152 from domestic animals in North Africa, the Middle East and Europe (*195*). We speculate that the emergence of community-associated MRSA epidemics is the result of a series of niche transition events: wild or domesticated animal-adapted strains first recruit into local human populations where they can become epidemic, exemplified most clearly by the ongoing ST93-MSSA outbreak in northern Australia. In subsequent niche transition events, facilitated by the acquisition of resistance determinants, these strains then emerged in regional (urban and industrialized) host populations governed by a different selective and epidemiological landscape, with access to antibiotic treatment and drugs, more or less effective practice of antibiotic stewardship, environmental pollution, high population densities and changes in host contact patterns or mobility, leaving signatures of epidemic growth in the effective reproduction number upon recruitment into a new adaptive landscape. International travel and migration then facilitated overseas transmission, including sustained transmission of resistant genotypes in both metropolitan (Pakistan) and remote host populations (Papua New

Guinea). In this scenario, the common driver behind the emergence of community-associated lineages - in combination with genetic virulence and resistance determinants - can be found in the level of connectivity at the animal and human interface (increased contact with animal hosts) and at the sociodemographic and -economic interface between different human subpopulations (e.g. increased contact between rural and urban populations, income and housing inequalities); a combination of host and host-niche transitions could therefore have facilitated the epidemic spread of community-associated *S. aureus* strains in an increasingly globalized and industrialized world of the 20th century.

In the second component of this project, we wanted to evaluate whether emerging nanopore sequencing technology can be used for the inference of transmission dynamics and the remote outbreak isolates we had sampled in Papua New Guinea and Far North Queensland. Part of this evaluation was targeted at developing *S. aureus* sequencing capacity for rural and remote laboratories, including at the Townsville University Hospital, a regional hospital with 742 beds and the reference laboratory at the Papua New Guinea institute of Medical Research. As such, we considered the complexity of the workflows involved, particularly when implementing sequencing with routine pathology work in under-staffed laboratories or in remote locations, where training and basic infrastructure like power or functional instruments are often lacking. Ultimately, we exploited some aspects of nanopore technology that allowed us to quickly and efficiently multiplex whole bacterial genomes and cheaper, even faster (2 hour) runs for surveillance using genomic neighbor typing (*46*) including at PNGIMR in Goroka and AITHM in Townsville, where sequencing infrastructure is otherwise not available. However, we decided to forego most throughput and read-length optimisations steps (bead cleanup steps in the rapid libraries, all libraries prepared on heat blocks instead of thermal cyclers, low concentration of DNA for barcodes reduced by up to half of recommended input per barcode, fragmentation from spin columns) and still successfully conducted high multiplexing runs using the successive

barcode panels (24 - 48 strains per MInION). While this resulted in sub-optimal library throughput (≈ 0.4 - 6 GB per 18- 24 hour run) and multiple barcodes that had to be excluded due to low throughput or excessive fragmentation, we confidently assembled and genotyped eight panels of 24 isolates, at an ultimate cost of approximately AUD $50 per genome with extractions (without labour) at low coverage (10 - 15x, Chapter 3) comparable to the cost of high coverage Illumina sequencing (around AUD $120 per isolate as service at the Doherty institute). It is very likely that an optimised (throughput and read-length) MinION run with more than 10 GB will be able to accommodate 96 strains for genomic neighbor typing, and possibly for whole genome assembly and variant calling.

We balanced the un-optimised sequencing strategy with computational methods, which we adopted from work on *Neisseria gonorrhoeae,* where the Oxford group around Sanderson *et al.* demonstrated that Random forest SNP polishers can be used to remove excessive false SNP calls from nanopore-native neural network variant callers *Medaka* and *Clair (32)*; we also adopted work by Karel Brinda and colleagues (*46*), who developed genomic neighbor typing, a rapid nanopore sequencing based heuristic to infer genomic characteristics from a strain-resolved database, We scaled our method to all known strains of *S. aureus* at the time and demonstrating its usefulness (and limitation) on independent validation data from the FNQ and PNG outbreaks. Both methods allowed us to use the cheap genome sequencing protocols at low coverage to infer genotypes at scale (48 strains on a MinION) and to call SNPs on all outbreak isolates from FNQ and PNG. Importantly we then showed that the SNP calls were accurate and precise enough (removing mostly false positive calls when trained against the outbreak reference genome) to use a newly developed hybrid nanopore core variant caller (based on *Snippy* core functions) that incorporates Illumina background data on the lineage (ST93 from our previous study not including the outbreaks in FNQ and PNG) with the nanopore SNP calls from Clair. As sufficient background data on the lineage is necessary in any approach

attempting to reconstruct phylogenetic trees from newly sequenced outbreaks, our results show that we can use only nanopore sequencing data at equivalent or lower cost than Illumina data to reconstruct the outbreaks using ML and Bayesian tree models. Lastly we showed that the Bayesian birth-death models are highly accurate and replicate Illumina reference estimates using outbreak isolates with 5-10x coverage. This allowed us to estimate effective reproduction numbers for FNQ and PNG from nanopore data. While we found evidence for epidemic transmission in the FNQ and PNG outbreaks (and in Pakistan) clusters of ST93-MRSA-IV in Auckland and the Northern Territory seemed to have lower reproduction numbers, and larger estimated sampling proportions, consistent with a decline of these local epidemics. Furthermore, these results indicate that birth-death skyline models may play a role in estimating the effectiveness of infection control measures at the low per genome cost and in locations without access to sequencing infrastructure.

Birth-death skyline models assume that populations are well-mixed, but clear population structure is evident between the MSSA and MRSA strains as well as in other monophyletic clades, such as the introductions of ST93 into PNG, FNQ and NZ. While we attempted to employ the more parameter-rich multitype birth-death models that are inherently capable of accounting for structured populations (*196*), the MCMC chains ultimately did not converge. This was likely due to a combination of large bacterial genome data sets and the parameter-rich model. We therefore employed the birth-death skyline model on monophyletic subsets of the lineage-wide variant alignment, provided sufficient isolates were available, thus reducing the potential impact on lineage-wide estimates arising from excessive population structure in the tree. We further explored a range of realistic configurations on the becoming uninfectious rate prior (1 - 10 years infectious period), for which few data are available from long-term surveillance studies, as well as weighting the reproductive number prior distribution at different levels of transmission (Methods). Prior sensitivity analysis for each lineage and clade confirms

that estimates were largely driven by the available data, rather than by the prior configurations. Further improvements to enable MCMC convergence for phylodynamic models used in large bacterial populations, including multitype birth-death models supporting larger sample sizes with improved numeric stability and Metropolis-coupled MCMC chains (*34*, *197*) will be useful for further investigations into the origins of community-associated S. Aureus. GPU driven models using the *BEAGLE* library for *BEAST2* also significantly accelerated MCMC computations on even moderate GPU hardware and will be essential for running real-time genome-informed surveillance of bacterial pathogens in the future.

In this study we have conducted genome-informed surveillance of community-associated *S. aureus* in remote Far North Queensland and Papua New Guinea, uncovering outbreaks sparked by the Australian ST93 lineage. Using birth-death skyline models we found surges in the reproduction numbers and sustained transmission of drug-resistant community-associated MSSA and MRSA clades across the globe, and unravelled parts the genetic and epidemiological drivers behind the seemingly convergent emergence of community-associated lineages in the late 20th century. We then developed and evaluated SNP calling and polishing for bacterial outbreak isolates and heuristic genotyping algorithms for outbreak surveillance on cheap, easy to implement protocols, that were tested in Papua New Guinea. Our work on the pediatric outbreak in Kundiawa and Goroka has led to changes in antibiotic prescription guidelines at Kundiawa General Hospital and we could show that decentralized sequencing of these outbreak data is even useful in rural and regional settings such as in northern Australia. Ultimately, we could show that nanopore sequencing technology will be a valuable asset for conducting genomic epidemiology and apply advanced epidemiological models of disease transmission to these data. However, while our work advances the technological capacity for bacterial disease surveillance, it also has shown that sociodemographic conditions and human behaviour are perhaps the ultimate drivers behind the emergence of drug-resistant

community-associated *S. aureus* including their potential origins from zoonotic spillovers, their emergence from spillovers into host populations centers, fueled by the acquisition of antibiotic resistance, and their eventual dissemination and ongoing transmission overseas. Large-scale political support, research funding and genome surveillance collaborations (*16*) supporting those domestic populations which are most affected by community-associated disease. Further investigations, particularly at remote human-wildlife interfaces and at rural-urban interfaces, will be required to further investigate the epidemiological and evolutionary processes involved in the origins of community-associated MRSA. Despite advances in tracking and monitoring and even treating bacterial disease in the community, nation states must improve the lives of marginalized people in our communities to address the ultimate drivers of disease emergence, including poverty and wealth distribution, as well as housing and access to healthcare. In a supposedly globalised world, in which diseases can emerge in one country and know no boundaries, it is the responsibility of countries with the means and technology - particularly those from which diseases emerge - to address these inequalities, and to support global disease surveillance and eradication efforts.

Data Availability

All sequence data for this project can be found in the SRA and ENA under project accession number: PRJNA657380. Additional data and all code with tasks implemented as command line utilities to replicate analyses and plots throughout this work (managed mostly through the NanoPath client: https://github.com/np-core/nanopath) - including associated Docker container (https://github.com/np-core/containers) - are available in the GitHub repositories:

Chapter 2.1:

- https://github.com/np-core/np-phybeast

Chapter 2.2:

- https://github.com/np-core/np-signal
- https://github.com/np-core/np-assembly
- https://github.com/np-cor/np-variants

Chapter 2.3:

- https://github.com/esteinig/sketchy

References

1. M. A. Spyrou, K. I. Bos, A. Herbig, J. Krause, Ancient pathogen genomics as an emerging tool for infectious disease research. *Nat. Rev. Genet.* **20**, 323–340 (2019).

2. S. Duchêne, S. Y. W. Ho, A. G. Carmichael, E. C. Holmes, H. Poinar, The Recovery, Interpretation and Use of Ancient Pathogen Genomes. *Curr. Biol.* **30**, R1215–R1231 (2020).

3. J. Davies, D. Davies, Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* **74**, 417–433 (2010).

4. R. I. Aminov, A brief history of the antibiotic era: lessons learned and challenges for the future. *Front. Microbiol.* **1**, 134–134 (2010).

5. K. Gould, Antibiotics: from prehistory to the present day. *J. Antimicrob. Chemother.* **71**, 572–575 (2016).

6. G. Barlow, Clinical challenges in antimicrobial resistance. *Nature Microbiology*. **3**, 258–260 (2018).

7. P. Durão, R. Balbontín, I. Gordo, Evolutionary Mechanisms Shaping the Maintenance of Antibiotic Resistance. *Trends Microbiol.* **26**, 677–691 (2018).

8. S. R. Partridge, S. M. Kwong, N. Firth, S. O. Jensen, Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clin. Microbiol. Rev.* **31**, e00088–17 (2018).

9. E. D. Sonnenburg, J. L. Sonnenburg, The ancestral and industrialized gut microbiota and implications for human health. *Nat. Rev. Microbiol.* **17**, 383–390 (2019).

10. M. C. Wibowo, Z. Yang, M. Borry, A. Hübner, K. D. Huang, B. T. Tierney, S. Zimmerman, F. Barajas-Olmos, C. Contreras-Cubas, H. García-Ortiz, A. Martínez-Hernández, J. M. Luber, P. Kirstahler, T. Blohm, F. E. Smiley, R. Arnold, S. A. Ballal, S. J. Pamp, J. Russ, F. Maixner, O. Rota-Stabelli, N. Segata, K. Reinhard, L. Orozco, C. Warinner, M. Snow, S. LeBlanc, A. D. Kostic, Reconstruction of ancient microbial genomes from the human gut. *Nature*. **594**, 234–239 (2021).

11. J. L. Gardy, N. J. Loman, Towards a genomics-informed, real-time, global pathogen

surveillance system. *Nat. Rev. Genet.* **19**, 9–20 (2018).

12. S. M. Nicholls, R. Poplawski, M. J. Bull, A. Underwood, M. Chapman, K. Abu-Dahab, B. Taylor, B. Jackson, S. Rey, R. Amato, R. Livett, S. Gonçalves, E. M. Harrison, S. J. Peacock, D. M. Aanensen, A. Rambaut, T. R. Connor, N. J. Loman, The COVID-19 Genomics UK (COG-UK) Consortium, MAJORA: Continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. *bioRxiv*. **10.1101/2020.10.06.328328** (2020), doi:10.1101/2020.10.06.328328.

13. J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R. A. Neher, Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. **34**, 4121–4123 (2018).

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

15. J. Köster, S. Rahmann, Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*. **28**, 2520–2522 (2012).

16. S. C. Inzaule, S. K. Tessema, Y. Kebede, A. E. Ogwell Ouma, J. N. Nkengasong, Genomic-informed pathogen surveillance in Africa: opportunities and challenges. *Lancet Infect. Dis.*, S1473–3099(20)30939–7 (2021).

17. D. Deamer, M. Akeson, D. Branton, Three decades of nanopore sequencing. *Nat. Biotechnol.* **34**, 518–524 (2016).

18. T. Seemann, C. R. Lane, N. L. Sherry, S. Duchene, A. Gonçalves da Silva, L. Caly, M. Sait, S. A. Ballard, K. Horan, M. B. Schultz, T. Hoang, M. Easton, S. Dougall, T. P. Stinear, J. Druce, M. Catton, B. Sutton, A. van Diemen, C. Alpren, D. A. Williamson, B. P. Howden, Tracking the COVID-19 pandemic in Australia using genomics. *Nat. Commun.* **11**, 4376 (2020).

19. B. B. Oude Munnink, D. F. Nieuwenhuijse, M. Stein, Á. O'Toole, M. Haverkate, M. Mollers, S. K. Kamga, C. Schapendonk, M. Pronk, P. Lexmond, A. van der Linden, T. Bestebroer, I. Chestakova, R. J. Overmars, S. van Nieuwkoop, R. Molenkamp, A. A. van der Eijk, C. GeurtsvanKessel, H. Vennema, A. Meijer, A. Rambaut, J. van Dissel, R. S. Sikkema, A. Timen, M. Koopmans, G. J. A. P. M. Oudehuis, J. Schinkel, J. Kluytmans, M. Kluytmans-van den Bergh, W. van den Bijllaardt, R. G. Berntvelsen, M. M. L. van Rijen, P. Schneeberger, S. Pas, B. M. Diederen, A. M. C. Bergmans, P. A. V. van der Eijk, J. J. Verweij, A. G. N. Buiting, R. Streefkerk, A. P. Aldenkamp, P. de Man, J. G. M. Koelemal, D. Ong, S. Paltansing, N. Veassen, J. Sleven, L. Bakker, H. Brockhoff, A. Rietveld, F. Slijkerman Megelink, J. Cohen Stuart, A. de Vries, W. van der Reijden, A. Ros, E. Lodder, E. Verspui-van der Eijk, I. Huijskens, E. M. Kraan, M. P. M. van der Linden, S. B. Debast, N. A. Naiemi, A. C. M. Kroes, M. Damen, S. Dinant, S. Lekkerkerk, O. Pontesilli, P. Smit, C. van Tienen, P. C. R. Godschalk, J. van Pelt, A. Ott, C. van der Weijden, H. Wertheim, J. Rahamat-Langendoen, J. Reimerink, R. Bodewes, E. Duizer, B. van der Veer, C. Reusken, S. Lutgens, P. Schneeberger, M. Hermans, P. Wever, A. Leenders, H. ter Waarbeek, C. Hoebe, The Dutch-Covid-19 response team, Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nat. Med.* **26**, 1405–1410 (2020).

20. N. R. Faria, T. A. Mellan, C. Whittaker, I. M. Claro, D. da S. Candido, S. Mishra, M. A. E.

Crispim, F. C. S. Sales, I. Hawryluk, J. T. McCrone, R. J. G. Hulswit, L. A. M. Franco, M. S. Ramundo, J. G. de Jesus, P. S. Andrade, T. M. Coletti, G. M. Ferreira, C. A. M. Silva, E. R. Manuli, R. H. M. Pereira, P. S. Peixoto, M. U. G. Kraemer, N. Gaburo, C. da C. Camilo, H. Hoeltgebaum, W. M. Souza, E. C. Rocha, L. M. de Souza, M. C. de Pinho, L. J. T. Araujo, F. S. V. Malta, A. B. de Lima, J. do P. Silva, D. A. G. Zauli, A. C. de S. Ferreira, R. P. Schnekenberg, D. J. Laydon, P. G. T. Walker, H. M. Schlüter, A. L. P. dos Santos, M. S. Vidal, V. S. Del Caro, R. M. F. Filho, H. M. dos Santos, R. S. Aguiar, J. L. Proença-Modena, B. Nelson, J. A. Hay, M. Monod, X. Miscouridou, H. Coupland, R. Sonabend, M. Vollmer, A. Gandy, C. A. Prete, V. H. Nascimento, M. A. Suchard, T. A. Bowden, S. L. K. Pond, C.-H. Wu, O. Ratmann, N. M. Ferguson, C. Dye, N. J. Loman, P. Lemey, A. Rambaut, N. A. Fraiji, M. do P. S. S. Carvalho, O. G. Pybus, S. Flaxman, S. Bhatt, E. C. Sabino, Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science*. **372**, 815–821 (2021).

21. T. Y. Michaelsen, M. Bennedbæk, L. E. Christiansen, M. S. F. Jørgensen, C. H. Møller, E. A. Sørensen, S. Knutsson, J. Brandt, T. B. N. Jensen, C. Chiche-Lapierre, E. F. Collados, T. Sørensen, C. Petersen, V. Le-Quy, M. Sereika, F. T. Hansen, M. Rasmussen, J. Fonager, S. M. Karst, R. L. Marvig, M. Stegger, R. N. Sieber, R. Skov, R. Legarth, T. G. Krause, A. Fomsgaard, The Danish Covid-19 Genome Consortium (DCGC), M. Albertsen, Introduction and transmission of SARS-CoV-2 B.1.1.7 in Denmark. *medRxiv*, 2021.06.04.21258333 (2021).

22. COVID-19 Genomics UK (COG-UK), An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe*. **1**, e99–e100 (2020).

23. J. Quick, N. J. Loman, S. Duraffour, J. T. Simpson, E. Severi, L. Cowley, J. A. Bore, R. Koundouno, G. Dudas, A. Mikhail, N. Ouédraogo, B. Afrough, A. Bah, J. H. Baum, B. Becker-Ziaja, J.-P. Boettcher, M. Cabeza-Cabrerizo, A. Camino-Sanchez, L. L. Carter, J. Doerrbecker, T. Enkirch, I. G. G. Dorival, N. Hetzelt, J. Hinzmann, T. Holm, L. E. Kafetzopoulou, M. Koropogui, A. Kosgey, E. Kuisma, C. H. Logue, A. Mazzarelli, S. Meisel, M. Mertens, J. Michel, D. Ngabo, K. Nitzsche, E. Pallash, L. V. Patrono, J. Portmann, J. G. Repits, N. Y. Rickett, A. Sachse, K. Singethan, I. Vitoriano, R. L. Yemanaberhan, E. G. Zekeng, R. Trina, A. Bello, A. A. Sall, O. Faye, O. Faye, N. 'faly Magassouba, C. V. Williams, V. Amburgey, L. Winona, E. Davis, J. Gerlach, F. Washington, V. Monteil, M. Jourdain, M. Bererd, A. Camara, H. Somlare, A. Camara, M. Gerard, G. Bado, B. Baillet, D. Delaune, K. Y. Nebie, A. Diarra, Y. Savane, R. B. Pallawo, G. J. Gutierrez, N. Milhano, I. Roger, C. J. Williams, F. Yattara, K. Lewandowski, J. Taylor, P. Rachwal, D. Turner, G. Pollakis, J. A. Hiscox, D. A. Matthews, M. K. O'Shea, A. M. Johnston, D. Wilson, E. Hutley, E. Smit, A. Di Caro, R. Woelfel, K. Stoecker, E. Fleischmann, M. Gabriel, S. A. Weller, L. Koivogui, B. Diallo, S. Keita, A. Rambaut, P. Formenty, S. Gunther, M. W. Carroll, Real-time, portable genome sequencing for Ebola surveillance. *Nature*. **530**, 228–232 (2016).

24. N. R. Faria, J. Quick, I. M. Claro, J. Thézé, J. G. de Jesus, M. Giovanetti, M. U. G. Kraemer, S. C. Hill, A. Black, A. C. da Costa, L. C. Franco, S. P. Silva, C.-H. Wu, J. Raghwani, S. Cauchemez, L. du Plessis, M. P. Verotti, W. K. de Oliveira, E. H. Carmo, G. E. Coelho, A. C. F. S. Santelli, L. C. Vinhal, C. M. Henriques, J. T. Simpson, M. Loose, K. G. Andersen, N. D. Grubaugh, S. Somasekar, C. Y. Chiu, J. E. Muñoz-Medina, C. R. Gonzalez-Bonilla, C. F. Arias, L. L. Lewis-Ximenez, S. A. Baylis, A. O. Chieppe, S. F. Aguiar, C. A. Fernandes, P. S. Lemos, B. L. S. Nascimento, H. A. O. Monteiro, I. C. Siqueira, M. G. de Queiroz, T. R. de Souza, J. F. Bezerra, M. R. Lemos, G. F. Pereira, D. Loudal, L. C. Moura, R. Dhalia, R. F. França, T. Magalhães, E. T. Marques, T. Jaenisch, G. L. Wallau, M. C. de Lima, V.

Nascimento, E. M. de Cerqueira, M. M. de Lima, D. L. Mascarenhas, J. P. M. Neto, A. S. Levin, T. R. Tozetto-Mendoza, S. N. Fonseca, M. C. Mendes-Correa, F. P. Milagres, A. Segurado, E. C. Holmes, A. Rambaut, T. Bedford, M. R. T. Nunes, E. C. Sabino, L. C. J. Alcantara, N. J. Loman, O. G. Pybus, Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature*. **546**, 406–410 (2017).

25. N. R. Faria, M. U. G. Kraemer, S. C. Hill, J. Goes de Jesus, R. S. Aguiar, F. C. M. Iani, J. Xavier, J. Quick, L. du Plessis, S. Dellicour, J. Thézé, R. D. O. Carvalho, G. Baele, C.-H. Wu, P. P. Silveira, M. B. Arruda, M. A. Pereira, G. C. Pereira, J. Lourenço, U. Obolski, L. Abade, T. I. Vasylyeva, M. Giovanetti, D. Yi, D. J. Weiss, G. R. W. Wint, F. M. Shearer, S. Funk, B. Nikolay, V. Fonseca, T. E. R. Adelino, M. A. A. Oliveira, M. V. F. Silva, L. Sacchetto, P. O. Figueiredo, I. M. Rezende, E. M. Mello, R. F. C. Said, D. A. Santos, M. L. Ferraz, M. G. Brito, L. F. Santana, M. T. Menezes, R. M. Brindeiro, A. Tanuri, F. C. P. dos Santos, M. S. Cunha, J. S. Nogueira, I. M. Rocco, A. C. da Costa, S. C. V. Komninakis, V. Azevedo, A. O. Chieppe, E. S. M. Araujo, M. C. L. Mendonça, C. C. dos Santos, C. D. dos Santos, A. M. Mares-Guia, R. M. R. Nogueira, P. C. Sequeira, R. G. Abreu, M. H. O. Garcia, A. L. Abreu, O. Okumoto, E. G. Kroon, C. F. C. de Albuquerque, K. Lewandowski, S. T. Pullan, M. Carroll, T. de Oliveira, E. C. Sabino, R. P. Souza, M. A. Suchard, P. Lemey, G. S. Trindade, B. P. Drumond, A. M. B. Filippis, N. J. Loman, S. Cauchemez, L. C. J. Alcantara, O. G. Pybus, Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science*. **361**, 894–899 (2018).

26. J. Quick, P. Ashton, S. Calus, C. Chatt, S. Gossain, J. Hawker, S. Nair, K. Neal, K. Nye, T. Peters, E. De Pinna, E. Robinson, K. Struthers, M. Webber, A. Catto, T. J. Dallman, P. Hawkey, N. J. Loman, Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol.* **16**, 114 (2015).

27. S. E. Heiden, N.-O. Hübner, J. A. Bohnert, C.-D. Heidecke, A. Kramer, V. Balau, W. Gierer, S. Schaefer, T. Eckmanns, S. Gatermann, E. Eger, S. Guenther, K. Becker, K. Schaufler, A Klebsiella pneumoniae ST307 outbreak clone from Germany demonstrates features of extensive drug resistance, hypermucoviscosity, and enhanced iron acquisition. *Genome Med.* **12**, 113 (2020).

28. X. Didelot, N. J. Croucher, S. D. Bentley, S. R. Harris, D. J. Wilson, Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134 (2018).

29. D. Helekal, A. Ledda, E. Volz, D. Wyllie, X. Didelot, Bayesian inference of clonal expansions in a dated phylogeny. *bioRxiv*, 2021.07.01.450370 (2021).

30. T. Stadler, D. Kühnert, S. Bonhoeffer, A. J. Drummond, Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl. Acad. Sci. U. S. A.* **110**, 228–233 (2013).

31. A. J. Drummond, A. Rambaut, B. Shapiro, O. G. Pybus, Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–1192 (2005).

32. R. Luo, C.-L. Wong, Y.-S. Wong, C.-I. Tang, C.-M. Liu, C.-M. Leung, T.-W. Lam, Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nature Machine Intelligence*. **2**, 220–227 (2020).

33. K. Shafin, T. Pesout, P.-C. Chang, M. Nattestad, A. Kolesnikov, S. Goel, G. Baid, J. M.

Eizenga, K. H. Miga, P. Carnevali, M. Jain, A. Carroll, B. Paten, Haplotype-aware variant calling enables high accuracy in nanopore long-reads using deep neural networks. *bioRxiv*. **10.1101/2021.03.04.433952** (2021), doi:10.1101/2021.03.04.433952.

34. J. Scire, J. Barido-Sottani, D. Kühnert, T. G. Vaughan, T. Stadler, Improved multi-type birth-death phylodynamic inference in BEAST 2. *bioRxiv*. **10.1101/2020.01.06.895532** (2020), doi:10.1101/2020.01.06.895532.

35. T. G. Vaughan, J. Sciré, S. A. Nadeau, T. Stadler, Estimates of outbreak-specific SARS-CoV-2 epidemiological parameters from genomic data. *bioRxiv*. **10.1101/2020.09.12.20193284** (2020), doi:10.1101/2020.09.12.20193284.

36. E. B. Hodcroft, M. Zuber, S. Nadeau, K. H. D. Crawford, J. D. Bloom, D. Veesler, T. G. Vaughan, I. Comas, F. G. Candelas, T. Stadler, R. A. Neher, Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *medRxiv*. **10.1101/2020.10.25.20219063** (2020), doi:10.1101/2020.10.25.20219063.

37. D. J. Ingle, B. P. Howden, S. Duchene, Development of Phylodynamic Methods for Bacterial Pathogens. *Trends Microbiol.* **10.1016/j.tim.2021.02.008** (2021), doi:10.1016/j.tim.2021.02.008.

38. I. Aglua, J. Jaworski, J. Drekore, B. Urakoko, H. Poka, A. Michael, A. Greenhill, Methicillin-Resistant Staphylococcus Aureus in Melanesian Children with Haematogenous Osteomyelitis from the Central Highlands of Papua New Guinea. *Int. J. Pediatr.* **6**, 8361–8370 (2018).

39. I. Guthridge, S. Smith, P. Horne, J. Hanson, Increasing prevalence of methicillin-resistant Staphylococcus aureus in remote Australian communities: implications for patients and clinicians. *Pathology*. **51** (2019), pp. 428–431.

40. E. J. Steinig, S. Duchene, D. Ashley Robinson, S. Monecke, M. Yokoyama, M. Laabei, P. Slickers, P. Andersson, D. Williamson, A. Kearns, R. Goering, E. Dickson, R. Ehricht, M. Ip, M. V. N. O'Sullivan, G. W. Coombs, A. Petersen, G. Brennan, A. C. Shore, D. C. Coleman, A. Pantosti, H. de Lencastre, H. Westh, N. Kobayashi, H. Heffernan, B. Strommenger, F. Layer, S. Weber, H. Aamot, L. Skakni, S. J. Peacock, D. Sarovich, S. Harris, J. Parkhill, R. C. Massey, M. T. G. Holden, S. D. Bentley, S. Y. C. Tong, Evolution and global transmission of a multidrug-resistant, community-associated MRSA lineage from the Indian subcontinent. *Cold Spring Harbor Laboratory* (2019), p. 233395.

41. T. P. Stinear, K. E. Holt, K. Chua, J. Stepnell, K. L. Tuck, G. Coombs, P. F. Harrison, T. Seemann, B. P. Howden, Adaptive change inferred from genomic population analysis of the ST93 epidemic clone of community-associated methicillin-resistant Staphylococcus aureus. *Genome Biol. Evol.* **6**, 366–378 (2014).

42. S. J. van Hal, E. J. Steinig, P. Andersson, M. T. G. Holden, S. R. Harris, G. R. Nimmo, D. A. Williamson, H. Heffernan, S. R. Ritchie, A. M. Kearns, M. J. Ellington, E. Dickson, H. de Lencastre, G. W. Coombs, S. D. Bentley, J. Parkhill, D. C. Holt, P. M. Giffard, S. Y. C. Tong, Global Scale Dissemination of ST93: A Divergent *Staphylococcus aureus* Epidemic Lineage That Has Recently Emerged From Remote Northern Australia. *Front. Microbiol.* **9**, 1453 (2018).

43. N. D. Sanderson, J. Swann, L. Barker, J. Kavanagh, S. Hoosdally, D. Crook, The GonFast

Investigators Group, T. L. Street, D. W. Eyre, High precision Neisseria gonorrhoeae variant and antimicrobial resistance calling from metagenomic Nanopore sequencing. *Genome Res.* (2020), doi:10.1101/gr.262865.120.

44. B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17** (2016), doi:10.1186/s13059-016-0997-x.

45. B. D. Ondov, G. J. Starrett, A. Sappington, A. Kostic, S. Koren, C. B. Buck, A. M. Phillippy, Mash Screen: high-throughput sequence containment estimation for genome discovery. *Genome Biol.* **20**, 232 (2019).

46. K. Břinda, A. Callendrello, K. C. Ma, D. R. MacFadden, T. Charalampous, R. S. Lee, L. Cowley, C. B. Wadsworth, Y. H. Grad, G. Kucherov, J. O'Grady, M. Baym, W. P. Hanage, Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. *Nature microbiology*. **5**, 455–464 (2020).

47. J. J. Bardy, D. S. Sarovich, E. P. Price, E. Steinig, S. Tong, A. Drilling, J. Ou, S. Vreugde, P.-J. Wormald, A. J. Psaltis, Staphylococcus aureus from patients with chronic rhinosinusitis show minimal genetic association between polyp and non-polyp phenotypes. *BMC Ear Nose Throat Disord.* **18**, 16 (2018).

48. D. E. Madden, J. R. Webb, E. J. Steinig, B. J. Currie, E. P. Price, D. S. Sarovich, Taking the next-gen step: Comprehensive antimicrobial resistance detection from Burkholderia pseudomallei. *EBioMedicine*. **63**, 103152 (2021).

49. S. J. van Hal, E. J. Steinig, P. Andersson, M. T. G. Holden, S. R. Harris, G. R. Nimmo, D. A. Williamson, H. Heffernan, S. R. Ritchie, A. M. Kearns, M. J. Ellington, E. Dickson, H. de Lencastre, G. W. Coombs, S. D. Bentley, J. Parkhill, D. C. Holt, P. M. Giffard, S. Y. C. Tong, Global Scale Dissemination of ST93: A Divergent Staphylococcus aureus Epidemic Lineage That Has Recently Emerged From Remote Northern Australia. *Front. Microbiol.* **9**, 1453 (2018).

50. M. Earls, E. J. Steinig, S. Monecke, J. A. C. Samaniego, A. Simbeck, W. Schneider-Brachert, T. Vremeră, O. S. Dorneanu, I. Loncaric, M. Bes, A. Lacoma, A. Wernery, C. P. Ulrich, M. Armengol-Porta, A. Blomfeldt, H. V. Aamot, S. Duchene, M. D. Bartels, R. Ehricht, D. C. Coleman, Exploring the evolution and epidemiology of European CC1-MRSA-IV: tracking a multidrug-resistant community-associated methicillin-resistant *Staphylococcus aureus clone*. *Microbial Genomics*. **in press** (2021).

51. J. L. Guppy, D. B. Jones, S. R. Kjeldsen, A. Le Port, M. S. Khatkar, N. M. Wade, M. J. Sellars, E. J. Steinig, H. W. Raadsma, D. R. Jerry, K. R. Zenger, Development and validation of a RAD-Seq target-capture based genotyping assay for routine application in advanced black tiger shrimp (Penaeus monodon) breeding programs. *BMC Genomics*. **21**, 541 (2020).

52. M. Neuditschko, H. W. Raadsma, M. S. Khatkar, E. Jonas, E. J. Steinig, C. Flury, H. Signer-Hasler, M. Frischknecht, R. von Niederhäusern, T. Leeb, S. Rieder, Identification of key contributors in complex population structures. *PLoS One*. **12**, e0177638 (2017).

53. J. R. Fitzgerald, M. T. G. Holden, Genomics of Natural Populations of Staphylococcus aureus. *Annu. Rev. Microbiol.* **70**, 459–478 (2016).

54. S. Y. C. Tong, J. S. Davis, E. Eichenberger, T. L. Holland, V. G. Fowler Jr, *Staphylococcus aureus* infections: epidemiology, pathophysiology, clinical manifestations, and management. *Clin. Microbiol. Rev.* **28**, 603–661 (2015).

55. B. Krismer, C. Weidenmaier, A. Zipperer, A. Peschel, The commensal lifestyle of Staphylococcus aureus and its interactions with the nasal microbiota. *Nat. Rev. Microbiol.* **15**, 675–687 (2017).

56. X. Du, J. Larsen, M. Li, A. Walter, C. Slavetinsky, A. Both, P. M. Sanchez Carballo, M. Stegger, E. Lehmann, Y. Liu, J. Liu, J. Slavetinsky, K. A. Duda, B. Krismer, S. Heilbronner, C. Weidenmaier, C. Mayer, H. Rohde, V. Winstel, A. Peschel, Staphylococcus epidermidis clones express Staphylococcus aureus-type wall teichoic acid to shift from a commensal to pathogen lifestyle. *Nature Microbiology*. **6**, 757–768 (2021).

57. N. A. Turner, B. K. Sharma-Kuinkel, S. A. Maskarinec, E. M. Eichenberger, P. P. Shah, M. Carugati, T. L. Holland, V. G. Fowler Jr, Methicillin-resistant *Staphylococcus aureus*: an overview of basic and clinical research. *Nat. Rev. Microbiol.* **17**, 203–218 (2019).

58. M. Kuroda, T. Ohta, I. Uchiyama, T. Baba, H. Yuzawa, I. Kobayashi, L. Cui, A. Oguchi, K. Aoki, Y. Nagai, J. Lian, T. Ito, M. Kanamori, H. Matsumaru, A. Maruyama, H. Murakami, A. Hosoyama, Y. Mizutani-Ui, N. K. Takahashi, T. Sawano, R. Inoue, C. Kaito, K. Sekimizu, H. Hirakawa, S. Kuhara, S. Goto, J. Yabuzaki, M. Kanehisa, A. Yamashita, K. Oshima, K. Furuya, C. Yoshino, T. Shiba, M. Hattori, N. Ogasawara, H. Hayashi, K. Hiramatsu, Whole genome sequencing of meticillin-resistant Staphylococcus aureus. *Lancet*. **357**, 1225–1240 (2001).

59. T. Baba, F. Takeuchi, M. Kuroda, H. Yuzawa, K.-I. Aoki, A. Oguchi, Y. Nagai, N. Iwama, K. Asano, T. Naimi, H. Kuroda, L. Cui, K. Yamamoto, K. Hiramatsu, Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet*. **359**, 1819–1827 (2002).

60. E. J. Feil, J. E. Cooper, H. Grundmann, D. A. Robinson, M. C. Enright, T. Berendt, S. J. Peacock, J. M. Smith, M. Murphy, B. G. Spratt, C. E. Moore, N. P. J. Day, How clonal is Staphylococcus aureus? *J. Bacteriol.* **185**, 3307–3316 (2003).

61. S. Jarraud, M. A. Peyrat, A. Lim, A. Tristan, M. Bes, C. Mougel, J. Etienne, F. Vandenesch, M. Bonneville, G. Lina, egc, a highly prevalent operon of enterotoxin gene, forms a putative nursery of superantigens in Staphylococcus aureus. *J. Immunol.* **166**, 669–677 (2001).

62. J. R. Fitzgerald, S. D. Reid, E. Ruotsalainen, T. J. Tripp, M. Liu, R. Cole, P. Kuusela, P. M. Schlievert, A. Järvinen, J. M. Musser, Genome diversification in Staphylococcus aureus: Molecular evolution of a highly variable chromosomal region encoding the Staphylococcal exotoxin-like family of proteins. *Infect. Immun.* **71**, 2827–2838 (2003).

63. N. Malachowa, F. R. DeLeo, Mobile genetic elements of Staphylococcus aureus. *Cell. Mol. Life Sci.* **67**, 3057–3071 (2010).

64. E. J. Steinig, P. Andersson, S. R. Harris, D. S. Sarovich, A. Manoharan, P. Coupland, M. T. G. Holden, J. Parkhill, S. D. Bentley, D. A. Robinson, S. Y. C. Tong, Single-molecule sequencing reveals the molecular basis of multidrug-resistance in ST772 methicillin-resistant *Staphylococcus aureus*. *BMC Genomics*. **16**, 388 (2015).

65. K. Chua, T. Seemann, P. F. Harrison, J. K. Davies, S. J. Coutts, H. Chen, V. Haring, R.

Moore, B. P. Howden, T. P. Stinear, Complete genome sequence of *Staphylococcus aureus* strain JKD6159, a unique Australian clone of ST93-IV community methicillin-resistant *Staphylococcus aureus*. *J. Bacteriol.* **192**, 5556–5557 (2010).

66. R. G. Everitt, X. Didelot, E. M. Batty, R. R. Miller, K. Knox, B. C. Young, R. Bowden, A. Auton, A. Votintseva, H. Larner-Svensson, J. Charlesworth, T. Golubchik, C. L. C. Ip, H. Godwin, R. Fung, T. E. A. Peto, A. S. Walker, D. W. Crook, D. J. Wilson, Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. *Nat. Commun.* **5** (2014).

67. M. C. Enright, N. P. J. Day, C. E. Davies, S. J. Peacock, B. G. Spratt, Multilocus Sequence Typing for Characterization of Methicillin-Resistant and Methicillin-Susceptible Clones of Staphylococcus aureus. *J. Clin. Microbiol.* **38**, 1008–1015 (2000).

68. D. A. Robinson, M. C. Enright, Multilocus sequence typing and the evolution of methicillin-resistant Staphylococcus aureus. *Clin. Microbiol. Infect.* **10**, 92–97 (2004).

69. S. Monecke, G. Coombs, A. C. Shore, D. C. Coleman, P. Akpaka, M. Borg, H. Chow, M. Ip, L. Jatzwauk, D. Jonas, K. Kadlec, A. Kearns, F. Laurent, F. G. O'Brien, J. Pearson, A. Ruppelt, S. Schwarz, E. Scicluna, P. Slickers, H.-L. Tan, S. Weber, R. Ehricht, A field guide to pandemic, epidemic and sporadic clones of methicillin-resistant Staphylococcus aureus. *PLoS One*. **6**, e17936 (2011).

70. L. B. Price, M. Stegger, H. Hasman, M. Aziz, J. Larsen, P. S. Andersen, T. Pearson, A. E. Waters, J. T. Foster, J. Schupp, J. Gillece, E. Driebe, C. M. Liu, B. Springer, I. Zdovc, A. Battisti, A. Franco, J. Zmudzki, S. Schwarz, P. Butaye, E. Jouy, C. Pomba, M. C. Porrero, R. Ruimy, T. C. Smith, D. A. Robinson, J. S. Weese, C. S. Arriola, F. Yu, F. Laurent, P. Keim, R. Skov, F. M. Aarestrup, *Staphylococcus aureus* CC398: host adaptation and emergence of methicillin resistance in livestock. *MBio*. **3** (2012), doi:10.1128/mBio.00305-11.

71. L. B. Price, M. Stegger, H. Hasman, M. Aziz, J. Larsen, P. S. Andersen, T. Pearson, A. E. Waters, J. T. Foster, J. Schupp, J. Gillece, E. Driebe, C. M. Liu, B. Springer, I. Zdovc, A. Battisti, A. Franco, J. Żmudzki, S. Schwarz, P. Butaye, E. Jouy, C. Pomba, M. C. Porrero, R. Ruimy, T. C. Smith, D. A. Robinson, J. S. Weese, C. S. Arriola, F. Yu, F. Laurent, P. Keim, R. Skov, F. M. Aarestrup, Staphylococcus aureus CC398: Host Adaptation and Emergence of Methicillin Resistance in Livestock. *MBio*. **3** (2012).

72. S. Sahibzada, S. Abraham, G. W. Coombs, S. Pang, M. Hernández-Jover, D. Jordan, J. Heller, Transmission of highly virulent community-associated MRSA ST93 and livestock-associated MRSA ST398 between humans and pigs in Australia. *Sci. Rep.* **7**, 5273 (2017).

73. A. C. Bowen, K. Daveson, L. Anderson, S. Y. Tong, An urgent need for antimicrobial stewardship in Indigenous rural and remote primary health care. *Med. J. Aust.* **211**, 9–11.e1 (2019).

74. S. A. J. Harch, E. MacMorran, S. Y. C. Tong, D. C. Holt, J. Wilson, E. Athan, S. Hewagama, High burden of complicated skin and soft tissue infections in the Indigenous population of Central Australia due to dominant Panton Valentine leucocidin clones ST93-MRSA and CC121-MSSA. *BMC Infect. Dis.* **17**, 405 (2017).

75. M. J. Ward, M. Goncheva, E. Richardson, P. R. McAdam, E. Raftis, A. Kearns, R. S. Daum,

M. Z. David, T. L. Lauderdale, G. F. Edwards, G. R. Nimmo, G. W. Coombs, X. Huijsdens, M. E. J. Woolhouse, J. R. Fitzgerald, Identification of source and sink populations for the emergence and global spread of the East-Asia clone of community-associated MRSA. *Genome Biol.* **17**, 160 (2016).

76. S. G. Giulieri, S. Y. C. Tong, D. A. Williamson, Using genomics to understand meticillin- and vancomycin-resistant Staphylococcus aureus infections. *Microb Genom*. **6**, e000324 (2020).

77. S. R. Harris, E. J. Feil, M. T. G. Holden, M. A. Quail, E. K. Nickerson, N. Chantratita, S. Gardete, A. Tavares, N. Day, J. A. Lindsay, J. D. Edgeworth, H. de Lencastre, J. Parkhill, S. J. Peacock, S. D. Bentley, Evolution of MRSA during hospital transmission and intercontinental spread. *Science*. **327**, 469–474 (2010).

78. M. T. G. Holden, L.-Y. Hsu, K. Kurt, L. A. Weinert, A. E. Mather, S. R. Harris, B. Strommenger, F. Layer, W. Witte, H. de Lencastre, R. Skov, H. Westh, H. Zemlicková, G. Coombs, A. M. Kearns, R. L. R. Hill, J. Edgeworth, I. Gould, V. Gant, J. Cooke, G. F. Edwards, P. R. McAdam, K. E. Templeton, A. McCann, Z. Zhou, S. Castillo-Ramírez, E. J. Feil, L. O. Hudson, M. C. Enright, F. Balloux, D. M. Aanensen, B. G. Spratt, J. R. Fitzgerald, J. Parkhill, M. Achtman, S. D. Bentley, U. Nübel, A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. *Genome Res.* **23**, 653–664 (2013).

79. S. Y. C. Tong, M. T. G. Holden, E. K. Nickerson, B. S. Cooper, C. U. Köser, A. Cori, T. Jombart, S. Cauchemez, C. Fraser, V. Wuthiekanun, J. Thaipadungpanit, M. Hongsuwan, N. P. Day, D. Limmathurotsakul, J. Parkhill, S. J. Peacock, Genome sequencing defines phylogeny and spread of methicillin-resistant *Staphylococcus aureus* in a high transmission setting. *Genome Res.* **25**, 111–118 (2015).

80. P. J. Planet, Life after USA300: The rise and fall of a superbug. *J. Infect. Dis.* **215**, S71–S77 (2017).

81. M. Stegger, T. Wirth, P. S. Andersen, R. L. Skov, A. De Grassi, P. M. Simões, A. Tristan, A. Petersen, M. Aziz, K. Kiil, I. Cirković, E. E. Udo, R. del Campo, J. Vuopio-Varkila, N. Ahmad, S. Tokajian, G. Peters, F. Schaumburg, B. Olsson-Liljequist, M. Givskov, E. E. Driebe, H. E. Vigh, A. Shittu, N. Ramdani-Bougessa, J.-P. Rasigade, L. B. Price, F. Vandenesch, A. R. Larsen, F. Laurent, Origin and Evolution of European Community-Acquired Methicillin-Resistant *Staphylococcus aureus*. *MBio*. **5** (2014), doi:10.1128/mBio.01044-14.

82. L. Straußß, M. Stegger, P. E. Akpaka, A. Alabi, S. Breurec, G. Coombs, B. Egyir, A. R. Larsen, F. Laurent, S. Monecke, G. Peters, R. Skov, B. Strommenger, F. Vandenesch, F. Schaumburg, A. Mellmann, Origin, evolution, and global transmission of community-acquired *Staphylococcus aureus* ST8. *Proceedings of the National Academy of Sciences*. **114**, E10596–E10604 (2017).

83. S. Baig, A. Rhod Larsen, P. Martins Simões, F. Laurent, T. B. Johannesen, B. Lilje, A. Tristan, F. Schaumburg, B. Egyir, I. Cirkovic, G. R. Nimmo, I. Spiliopoulou, D. S. Blanc, S. Mernelius, A. E. F. Moen, M. Z. David, P. S. Andersen, M. Stegger, Evolution and Population Dynamics of Clonal Complex 152 Community-Associated Methicillin-Resistant *Staphylococcus aureus*. *mSphere*. **5**, e00226–20 (2020).

84. L. Challagundla, X. Luo, I. A. Tickler, X. Didelot, D. C. Coleman, A. C. Shore, G. W. Coombs, D. O. Sordelli, E. L. Brown, R. Skov, A. R. Larsen, J. Reyes, I. E. Robledo, G. J. Vazquez, R. Rivera, P. D. Fey, K. Stevenson, S.-H. Wang, B. N. Kreiswirth, J. R. Mediavilla, C. A. Arias, P. J. Planet, R. L. Nolan, F. C. Tenover, R. V. Goering, D. A. Robinson, Range Expansion and the Origin of USA300 North American Epidemic Methicillin-Resistant *Staphylococcus aureus*. *MBio*. **9** (2018), doi:10.1128/mBio.02016-17.

85. P. J. Planet, L. Diaz, S.-O. Kolokotronis, A. Narechania, J. Reyes, G. Xing, S. Rincon, H. Smith, D. Panesso, C. Ryan, D. P. Smith, M. Guzman, J. Zurita, R. Sebra, G. Deikus, R. L. Nolan, F. C. Tenover, G. M. Weinstock, D. A. Robinson, C. A. Arias, Parallel Epidemics of Community-Associated Methicillin-Resistant *Staphylococcus aureus* USA300 Infection in North and South America. *J. Infect. Dis.* **212**, 1874–1882 (2015).

86. J. Collins, J. Rudkin, M. Recker, C. Pozzi, J. P. O'Gara, R. C. Massey, Offsetting virulence and antibiotic resistance costs by MRSA. *ISME J.* **4**, 577–584 (2010).

87. C. A. Gustave, J. P. Rasigade, P. Martins-Simões, F. Couzon, C. Bourg, A. Tristan, F. Laurent, T. Wirth, F. Vandenesch, Potential role of Mercury pollutants in the success of Methicillin-Resistant *Staphylococcus aureus* USA300 in Latin America. *bioRxiv*. **10.1101/2020.07.01.150961** (2020), doi:10.1101/2020.07.01.150961.

88. C.-A. Gustave, A. Tristan, P. Martins-Simões, M. Stegger, Y. Benito, P. S. Andersen, M. Bes, T. Le Hir, B. A. Diep, A.-C. Uhlemann, P. Glaser, F. Laurent, T. Wirth, F. Vandenesch, Demographic fluctuation of community-acquired antibiotic-resistant *Staphylococcus aureus* lineages: potential role of flimsy antibiotic exposure. *ISME J.* **12**, 1879–1894 (2018).

89. G. W. Coombs, R. V. Goering, K. Y. L. Chua, S. Monecke, B. P. Howden, T. P. Stinear, R. Ehricht, F. G. O'Brien, K. J. Christiansen, The molecular epidemiology of the highly virulent ST93 Australian community *Staphylococcus aureus* strain. *PLoS One*. **7**, e43037 (2012).

90. M. R. Earls, P. M. Kinnevey, G. I. Brennan, A. Lazaris, M. Skally, B. O'Connell, H. Humphreys, A. C. Shore, D. C. Coleman, The recent emergence in hospitals of multidrug-resistant community-associated sequence type 1 and spa type t127 methicillin-resistant *Staphylococcus aureus* investigated by whole-genome sequencing: implications for screening. *PLoS One*. **12**, e0175542 (2017).

91. M. R. Earls, A. C. Shore, G. I. Brennan, A. Simbeck, W. Schneider-Brachert, T. Vremeră, O. S. Dorneanu, P. Slickers, R. Ehricht, S. Monecke, D. C. Coleman, A novel multidrug-resistant PVL-negative CC1-MRSA-IV clone emerging in Ireland and Germany likely originated in South-Eastern Europe. *Infect. Genet. Evol.* **69**, 117–126 (2019).

92. S. M. Edslev, H. Westh, P. S. Andersen, R. Skov, N. Kobayashi, M. D. Bartels, F. Vandenesch, A. Petersen, P. Worning, A. R. Larsen, M. Stegger, Identification of a PVL-negative SCC*mec*-IVa sublineage of the methicillin-resistant *Staphylococcus aureus* CC80 lineage: understanding the clonal origin of CA-MRSA. *Clin. Microbiol. Infect.* **24**, 273–278 (2018).

93. A. Blomfeldt, K. W. Larssen, A. Moghen, C. Gabrielsen, P. Elstrøm, H. V. Aamot, S. B. Jørgensen, Emerging multidrug-resistant Bengal Bay clone ST772-MRSA-V in Norway: molecular epidemiology 2004-2014. *Eur. J. Clin. Microbiol. Infect. Dis.* **36**, 1911–1921 (2017).

94. M. T. Alam, T. D. Read, R. A. Petit 3rd, S. Boyle-Vavra, L. G. Miller, S. J. Eells, R. S. Daum, M. Z. David, Transmission and microevolution of USA300 MRSA in U.S. households: evidence from whole-genome sequencing. *MBio*. **6**, e00054 (2015).

95. A.-C. Uhlemann, J. Dordel, J. R. Knox, K. E. Raven, J. Parkhill, M. T. G. Holden, S. J. Peacock, F. D. Lowy, Molecular tracing of the emergence, diversification, and transmission of *S. aureus* sequence type 8 in a New York community. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 6738–6743 (2014).

96. H. H. Harastani, S. T. Tokajian, Community-associated methicillin-resistant *Staphylococcus aureus* clonal complex 80 type IV (CC80-MRSA-IV) isolated from the Middle East: a heterogeneous expanding clonal lineage. *PLoS One*. **9**, e103715 (2014).

97. M. L. Masim, S. Argimón, H. O. Espiritu, M. A. Magbanua, M. L. Lagrada, A. M. Olorosa, V. Cohen, J. M. Gayeta, B. Jeffrey, K. Abudahab, C. M. Hufano, S. B. Sia, M. T. G. Holden, J. Stelling, D. M. Aanensen, C. C. Carlos, on B. of the Philippines Antimicrobial Resistance Surveillance Program, Genomic Surveillance of Methicillin-Resistant Staphylococcus aureus in the Philippines from 2013-2014. *bioRxiv* (2020), doi:10.1101/2020.03.19.998401.

98. S. R. Ritchie, M. G. Thomas, P. B. Rainey, The Genetic Structure of Staphylococcus aureus Populations from the Southwest Pacific. *PLoS One*. **9**, e100300 (2014).

99. A. Jenney, D. Holt, R. Ritika, P. Southwell, S. Pravin, E. Buadromo, J. Carapetis, S. Tong, A. Steer, The clinical and molecular epidemiology of Staphylococcus aureus infections in Fiji. *BMC Infect. Dis.* **14**, 160 (2014).

100. N. D. Foxlee, N. Townell, L. McIver, C. L. Lau, Antibiotic Resistance in Pacific Island Countries and Territories: A Systematic Scoping Review. *Antibiotics (Basel)*. **8** (2019), doi:10.3390/antibiotics8010029.

101. M. Laman, A. Greenhill, G. W. Coombs, O. Robinson, J. Pearson, T. M. E. Davis, L. Manning, Methicillin-resistant *Staphylococcus aureus* in Papua New Guinea: a community nasal colonization prevalence study. *Trans. R. Soc. Trop. Med. Hyg.* **111**, 360–362 (2017).

102. A. Bainomugisa, S. Pandey, E. Donnan, G. Simpson, J. Foster, E. Lavu, S. Hiasihri, E. McBryde, R. Moke, S. Vincent, V. Sintchenko, B. Marais, L. J. M. Coin, C. Coulter, Cross-Border Movement of Highly Drug-Resistant Mycobacterium tuberculosis from Papua New Guinea to Australia through Torres Strait Protected Zone, 2010–2015. *Emerging Infectious Disease journal*. **25**, 406 (2019).

103. T. M. Wozniak, W. Cuningham, S. Buchanan, S. Coulter, R. W. Baird, G. R. Nimmo, C. C. Blyth, S. Y. C. Tong, B. J. Currie, A. P. Ralph, Geospatial epidemiology of *Staphylococcus aureus* in a tropical setting: an enabling digital surveillance platform. *Sci. Rep.* **10**, 13169 (2020).

104. W. J. Munckhof, J. Schooneveldt, G. W. Coombs, J. Hoare, G. R. Nimmo, Emergence of community-acquired methicillin-resistant Staphylococcus aureus (MRSA) infection in Queensland, Australia. *Int. J. Infect. Dis.* **7**, 259–264 (2003).

105. M. Jain, H. E. Olsen, B. Paten, M. Akeson, The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).

106. H. Teng, M. D. Cao, M. B. Hall, T. Duarte, S. Wang, L. J. M. Coin, Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *Gigascience*. **7**, giy037–giy037 (2018).

107. M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O'Grady, H. E. Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan, T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, M. Loose, Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).

108. M. Loose, S. Malla, M. Stout, Real-time selective sequencing using nanopore technology. *Nat. Methods*. **13**, 751–754 (2016).

109. A. C. Rand, M. Jain, J. M. Eizenga, A. Musselman-Brown, H. E. Olsen, M. Akeson, B. Paten, Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* (2017), doi:10.1038/nmeth.4189.

110. M. D. Cao, S. H. Nguyen, D. Ganesamoorthy, A. G. Elliott, M. A. Cooper, L. J. M. Coin, Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nat. Commun.* **8**, 14515 (2017).

111. E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg, W. R. Hinsley, D. J. Laydon, G. Dabrera, Á. O'Toole, R. Amato, M. Ragonnet-Cronin, I. Harrison, B. Jackson, C. V. Ariani, O. Boyd, N. J. Loman, J. T. McCrone, S. Gonçalves, D. Jorgensen, R. Myers, V. Hill, D. K. Jackson, K. Gaythorpe, N. Groves, J. Sillitoe, D. P. Kwiatkowski, S. Flaxman, O. Ratmann, S. Bhatt, S. Hopkins, A. Gandy, A. Rambaut, N. M. Ferguson, The COVID-19 Genomics UK (COG-UK) consortium, Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. *bioRxiv*. **10.1101/2020.12.30.20249034** (2021), doi:10.1101/2020.12.30.20249034.

112. L. du Plessis, J. T. McCrone, A. E. Zarebski, V. Hill, C. Ruis, B. Gutierrez, J. Raghwani, J. Ashworth, R. Colquhoun, T. R. Connor, N. R. Faria, B. Jackson, N. J. Loman, Á. O'Toole, S. M. Nicholls, K. V. Parag, E. Scher, T. I. Vasylyeva, E. M. Volz, A. Watts, I. I. Bogoch, K. Khan, COVID-19 Genomics UK (COG-UK) Consortium, D. M. Aanensen, M. U. G. Kraemer, A. Rambaut, O. G. Pybus, Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science*. **371**, 708–712 (2021).

113. R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. De Maio, M. Matschiner, F. K. Mendes, N. F. Müller, H. A. Ogilvie, L. du Plessis, A. Popinga, A. Rambaut, D. Rasmussen, I. Siveroni, M. A. Suchard, C.-H. Wu, D. Xie, C. Zhang, T. Stadler, A. J. Drummond, BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).

114. S. Duchêne, J. L. Geoghegan, E. C. Holmes, S. Y. W. Ho, Estimating evolutionary rates using time-structured data: a general comparison of phylogenetic methods. *Bioinformatics*. **32**, 3375–3379 (2016).

115. S. Duchene, P. Lemey, T. Stadler, S. Y. W. Ho, D. A. Duchene, V. Dhanasekaran, G. Baele, Bayesian Evaluation of Temporal Signal in Measurably Evolving Populations. *Mol. Biol. Evol.* **37**, 3363–3379 (2020).

116. S. Duchêne, K. E. Holt, F.-X. Weill, S. Le Hello, J. Hawkey, D. J. Edwards, M. Fourment, E. C. Holmes, Genome-scale rates of evolutionary change in bacteria. *Microb Genom*. **2**, e000094 (2016).

117. A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*. **35**, 4453–4455 (2019).

118. Y. Bakthavatchalam, K. Vasudevan, S. Rao, S. Varughese, P. Rupali, M. Gina, M. Zervos, J. Peter, B. Veeraraghavan, Genomic Portrait of Community-Associated Methicillin-Resistant *Staphylococcus aureus* ST772-SCC*mec*-V Lineage From India. *bioRxiv*. **10.21203/rs.3.rs-141678/1** (2021), doi:10.21203/rs.3.rs-141678/v1.

119. M. C. Roberts, P. R. Joshi, S. Monecke, R. Ehricht, E. Müller, D. Gawlik, S. Paudel, M. Acharya, S. Bhattarai, S. Pokharel, R. Tuladhar, M. K. Chalise, R. C. Kyes, MRSA Strains in Nepalese Rhesus Macaques (Macaca mulatta) and Their Environment. *Front. Microbiol.* **10**, 2505 (2019).

120. M. A. Syed, S. H. H. Shah, Y. Sherafzal, S. Shafi-Ur-Rehman, M. A. Khan, J. B. Barrett, T. A. Woodley, B. Jamil, S. A. Abbasi, C. R. Jackson, Detection and Molecular Characterization of Methicillin-Resistant *Staphylococcus aureus* from Table Eggs in Haripur, Pakistan. *Foodborne Pathog. Dis.* **15**, 86–93 (2018).

121. C. Montelongo, C. R. Mores, C. Putonti, A. J. Wolfe, A. Abouelfetouh, Phylogenomic study of *Staphylococcus aureus* and *Staphylococcus haemolyticus* clinical isolates from Egypt. *bioRxiv*. **10.1101/2021.05.01.442118** (2021), doi:10.1101/2021.05.01.442118.

122. L. Thomas, A. C. Bowen, M. Ly, C. Connors, R. Andrews, S. Y. C. Tong, Burden of skin disease in two remote primary healthcare centres in northern and central Australia. *Intern. Med. J.* **49**, 396–399 (2019).

123. D. A. Williamson, G. W. Coombs, G. R. Nimmo, *Staphylococcus aureus* “Down Under”: contemporary epidemiology of *S. aureus* in Australia, New Zealand, and the South West Pacific. *Clin. Microbiol. Infect.* **20**, 597–604 (2014).

124. A. C. Bowen, A. Mahé, R. J. Hay, R. M. Andrews, A. C. Steer, S. Y. C. Tong, J. R. Carapetis, The Global Epidemiology of Impetigo: A Systematic Review of the Population Prevalence of Impetigo and Pyoderma. *PLoS One*. **10**, e0136789 (2015).

125. B. A. Diep, G. G. Stone, L. Basuino, C. J. Graber, A. Miller, S.-A. des Etages, A. Jones, A. M. Palazzolo-Ballance, F. Perdreau-Remington, G. F. Sensabaugh, F. R. DeLeo, H. F. Chambers, The Arginine Catabolic Mobile Element and Staphylococcal Chromosomal Cassette *mec* Linkage: Convergence of Virulence and Resistance in the USA300 Clone of Methicillin-Resistant *Staphylococcus aureus*. *J. Infect. Dis.* **197**, 1523–1530 (2008).

126. P. J. Planet, S. J. LaRussa, A. Dana, H. Smith, A. Xu, C. Ryan, A.-C. Uhlemann, S. Boundy, J. Goldberg, A. Narechania, R. Kulkarni, A. J. Ratner, J. A. Geoghegan, S.-O. Kolokotronis, A. Prince, Emergence of the epidemic methicillin-resistant *Staphylococcus aureus* strain USA300 coincides with horizontal transfer of the arginine catabolic mobile element and *speG*-mediated adaptations for survival on skin. *MBio*. **4**, e00889–13 (2013).

127. C. J. Bond, D. Singh, More than a refresh required for closing the gap of Indigenous

health inequality. *Med. J. Aust.* **212**, 198–199.e1 (2020).

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

129. A. Souvorov, R. Agarwala, D. J. Lipman, SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biol.* **19**, 153 (2018).

130. E. Zankari, H. Hasman, S. Cosentino, M. Vestergaard, S. Rasmussen, O. Lund, F. M. Aarestrup, M. V. Larsen, Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640–2644 (2012).

131. H. Kaya, H. Hasman, J. Larsen, M. Stegger, T. B. Johannesen, R. L. Allesøe, C. K. Lemvigh, F. M. Aarestrup, O. Lund, A. R. Larsen, SCC*mec*Finder, a Web-Based Tool for Typing of Staphylococcal Cassette Chromosome *mec* in *Staphylococcus aureus* Using Whole-Genome Sequence Data. *mSphere*. **3**, e00612–17 (2018).

132. M. Hunt, P. Bradley, S. G. Lapierre, S. Heys, M. Thomsit, M. B. Hall, K. M. Malone, P. Wintringer, T. M. Walker, D. M. Cirillo, I. Comas, M. R. Farhat, P. Fowler, J. Gardy, N. Ismail, T. A. Kohl, V. Mathys, M. Merker, S. Niemann, S. V. Omar, V. Sintchenko, G. Smith, D. Soolingen, P. Supply, S. Tahseen, M. Wilcox, I. Arandjelovic, T. E. A. Peto, D. W. Crook, Z. Iqbal, Antibiotic resistance prediction for Mycobacterium tuberculosis from genome sequence data with Mykrobe [version 1; peer review: 2 approved, 1 approved with reservations]. *Wellcome Open Research*. **4** (2019), doi:10.12688/wellcomeopenres.15603.1.

133. N. J. Croucher, A. J. Page, T. R. Connor, A. J. Delaney, J. A. Keane, S. D. Bentley, J. Parkhill, S. R. Harris, Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).

134. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol*. **4**, vex042 (2018).

135. T.-H. To, M. Jung, S. Lycett, O. Gascuel, Fast Dating Using Least-Squares Criteria and Algorithms. *Syst. Biol.* **65**, 82–97 (2016).

136. S. Duchêne, D. Duchêne, E. C. Holmes, S. Y. W. Ho, The Performance of the Date-Randomization Test in Phylogenetic Analyses of Time-Structured Virus Data. *Mol. Biol. Evol.* **32**, 1895–1906 (2015).

137. T. G. Vaughan, IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics*. **33**, 2392–2394 (2017).

138. C. Ramsden, E. C. Holmes, M. A. Charleston, Hantavirus evolution in relation to its rodent and insectivore hosts: no evidence for codivergence. *Mol. Biol. Evol.* **26**, 143–153 (2009).

139. B. J. Smith, boa: an R package for MCMC output convergence assessment and posterior inference. *J. Stat. Softw.* **21**, 1–37 (2007).

140. F. Di Ruscio, G. Guzzetta, J. V. Bjørnholt, T. M. Leegaard, A. E. F. Moen, S. Merler, B. Freiesleben de Blasio, Quantifying the transmission dynamics of MRSA in the community

and healthcare settings in a low-prevalence country. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 14599–14605 (2019).

141. E. M. C. D'Agata, G. F. Webb, M. A. Horn, R. C. Moellering Jr, S. Ruan, Modelling the invasion of community-acquired methicillin-resistant *Staphylococcus aureus* into hospitals. *Clin. Infect. Dis.* **48**, 274–284 (2009).

142. B. S. Cooper, T. Kypraios, R. Batra, D. Wyncoll, O. Tosas, J. D. Edgeworth, Quantifying type-specific reproduction numbers for nosocomial pathogens: evidence for heightened transmission of an Asian sequence type 239 MRSA clone. *PLoS Comput. Biol.* **8**, e1002454 (2012).

143. M. Prosperi, N. Veras, T. Azarian, M. Rathore, D. Nolan, K. Rand, R. L. Cook, J. Johnson, J. G. Morris Jr, M. Salemi, Molecular epidemiology of community-associated methicillin-resistant *Staphylococcus aureus* in the genomic era: a cross-sectional study. *Sci. Rep.* **3**, 1902 (2013).

144. N. C. Gordon, B. Pichon, T. Golubchik, D. J. Wilson, J. Paul, D. S. Blanc, K. Cole, J. Collins, N. Cortes, M. Cubbon, F. K. Gould, P. J. Jenks, M. Llewelyn, J. Q. Nash, J. M. Orendi, K. Paranthaman, J. R. Price, L. Senn, H. L. Thomas, S. Wyllie, D. W. Crook, T. E. A. Peto, A. S. Walker, A. M. Kearns, Whole-Genome Sequencing Reveals the Contribution of Long-Term Carriers in *Staphylococcus aureus* Outbreak Investigation. *J. Clin. Microbiol.* **55**, 2188–2197 (2017).

145. G. Muthukrishnan, R. P. Lamers, A. Ellis, V. Paramanandam, A. B. Persaud, S. Tafur, C. L. Parkinson, A. M. Cole, Longitudinal genetic analyses of *Staphylococcus aureus* nasal carriage dynamics in a diverse population. *BMC Infect. Dis.* **13**, 221 (2013).

146. A. Scanvic, L. Denic, S. Gaillon, P. Giry, A. Andremont, J. C. Lucet, Duration of colonization by methicillin-resistant *Staphylococcus aureus* after hospital discharge and risk factors for prolonged carriage. *Clin. Infect. Dis.* **32**, 1393–1398 (2001).

147. D. L. Ayres, M. P. Cummings, G. Baele, A. E. Darling, P. O. Lewis, D. L. Swofford, J. P. Huelsenbeck, P. Lemey, A. Rambaut, M. A. Suchard, BEAGLE 3: Improved Performance, Scaling, and Usability for a High-Performance Computing Library for Statistical Phylogenetics. *Syst. Biol.* **68**, 1052–1061 (2019).

148. R. A. Bull, T. N. Adikari, J. M. Ferguson, J. M. Hammond, I. Stevanovski, A. G. Beukers, Z. Naing, M. Yeang, A. Verich, H. Gamaarachchi, K. W. Kim, F. Luciani, S. Stelzer-Braid, J.-S. Eden, W. D. Rawlinson, S. J. van Hal, I. W. Deveson, Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat. Commun.* **11**, 6272 (2020).

149. A. da Silva Filipe, J. G. Shepherd, T. Williams, J. Hughes, E. Aranday-Cortes, P. Asamaphan, S. Ashraf, C. Balcazar, K. Brunker, A. Campbell, S. Carmichael, C. Davis, R. Dewar, M. D. Gallagher, R. Gunson, V. Hill, A. Ho, B. Jackson, E. James, N. Jesudason, N. Johnson, E. C. McWilliam Leitch, K. Li, A. MacLean, D. Mair, D. A. McAllister, J. T. McCrone, S. E. McDonald, M. P. McHugh, A. K. Morris, J. Nichols, M. Niebel, K. Nomikou, R. J. Orton, Á. O'Toole, M. Palmarini, B. J. Parcell, Y. A. Parr, A. Rambaut, S. Rooke, S. Shaaban, R. Shah, J. B. Singer, K. Smollett, I. Starinskij, L. Tong, V. B. Sreenu, E. Wastnedge, COVID-19 Genomics UK (COG-UK) Consortium, M. T. G. Holden, D. L. Robertson, K. Templeton, E. C. Thomson, Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland. *Nat Microbiol*. **6**,

112–122 (2021).

150. A. S. Hammer, M. L. Quaade, T. B. Rasmussen, J. Fonager, M. Rasmussen, K. Mundbjerg, L. Lohse, B. Strandbygaard, C. S. Jørgensen, A. Alfaro-Núñez, M. W. Rosenstierne, A. Boklund, T. Halasa, A. Fomsgaard, G. J. Belsham, A. Bøtner, SARS-CoV-2 Transmission between Mink (*Neovison vison*) and Humans, Denmark. *Emerg. Infect. Dis.* **27**, 547–551 (2021).

151. M. Giovanetti, N. R. Faria, J. Lourenço, J. Goes de Jesus, J. Xavier, I. M. Claro, M. U. G. Kraemer, V. Fonseca, S. Dellicour, J. Thézé, F. da Silva Salles, T. Gräf, P. P. Silveira, V. A. do Nascimento, V. Costa de Souza, F. C. de Melo Iani, E. A. Castilho-Martins, L. N. Cruz, G. Wallau, A. Fabri, F. Levy, J. Quick, V. de Azevedo, R. S. Aguiar, T. de Oliveira, C. Bôtto de Menezes, M. da Costa Castilho, T. M. Terra, M. Souza da Silva, A. M. Bispo de Filippis, A. Luiz de Abreu, W. K. Oliveira, J. Croda, C. F. Campelo de Albuquerque, M. R. T. Nunes, E. C. Sabino, N. Loman, F. G. Naveca, O. G. Pybus, L. C. Alcantara, Genomic and Epidemiological Surveillance of Zika Virus in the Amazon Region. *Cell Rep.* **30**, 2275–2283.e7 (2020).

152. J. Quick, N. D. Grubaugh, S. T. Pullan, I. M. Claro, A. D. Smith, K. Gangavarapu, G. Oliveira, R. Robles-Sikisaka, T. F. Rogers, N. A. Beutler, D. R. Burton, L. L. Lewis-Ximenez, J. G. de Jesus, M. Giovanetti, S. C. Hill, A. Black, T. Bedford, M. W. Carroll, M. Nunes, L. C. Alcantara Jr, E. C. Sabino, S. A. Baylis, N. R. Faria, M. Loose, J. T. Simpson, O. G. Pybus, K. G. Andersen, N. J. Loman, Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* **12**, 1261–1276 (2017).

153. C. L. Gorrie, A. G. Da Silva, D. J. Ingle, C. Higgs, T. Seemann, T. P. Stinear, D. A. Williamson, J. C. Kwong, M. Lindsay Grayson, N. L. Sherry, B. P. Howden, Systematic analysis of key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria. *bioRxiv*. **10.1101/2020.09.24.310821** (2020), doi:10.1101/2020.09.24.310821.

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

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

156. L. Urban, A. Holzer, J. J. Baronas, M. B. Hall, P. Braeuninger-Weimer, M. J. Scherm, D. J. Kunz, S. N. Perera, D. E. Martin-Herranz, E. T. Tipper, S. J. Salter, M. R. Stammnitz, Freshwater monitoring by nanopore sequencing. *Elife*. **10**, e61504 (2021).

157. B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, A. M. Earl, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. **9**, e112963 (2014).

158. R. R. Wick, L. M. Judd, C. L. Gorrie, K. E. Holt, Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595

(2017).

159. M. Kolmogorov, J. Yuan, Y. Lin, P. A. Pevzner, Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

160. G. Marçais, A. L. Delcher, A. M. Phillippy, R. Coston, S. L. Salzberg, A. Zimin, MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).

161. F. A. Ferreira, K. Helmersen, T. Visnovska, S. B. Jørgensen, H. V. Aamot, Rapid nanopore-based DNA sequencing protocol of antibiotic-resistant bacteria for use in surveillance and outbreak investigation. *Microb Genom*. **7** (2021), doi:10.1099/mgen.0.000557.

162. S. Lipworth, H. Pickford, N. Sanderson, K. K. Chau, J. Kavanagh, L. Barker, A. Vaughan, J. Swann, M. Andersson, K. Jeffery, M. Morgan, T. E. A. Peto, D. W. Crook, N. Stoesser, A. S. Walker, Optimized use of Oxford Nanopore flowcells for hybrid assemblies. *Microb Genom*. **6** (2020), doi:10.1099/mgen.0.000453.

163. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. **34**, 3094–3100 (2018).

164. R. Vaser, I. Sović, N. Nagarajan, M. Šikić, Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).

165. L. Chen, D. Zheng, B. Liu, J. Yang, Q. Jin, VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. *Nucleic Acids Res.* **44**, D694–7 (2016).

166. F. Pedregosa, G. Varoquaux, A. Gramfort, Scikit-learn: Machine learning in Python. *of machine Learning …*. **12**, 2825–2830 (2011).

167. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).

168. K. L. Wyres, T. N. T. Nguyen, M. M. C. Lam, L. M. Judd, N. van Vinh Chau, D. A. B. Dance, M. Ip, A. Karkey, C. L. Ling, T. Miliya, P. N. Newton, N. P. H. Lan, A. Sengduangphachanh, P. Turner, B. Veeraraghavan, P. V. Vinh, M. Vongsouvath, N. R. Thomson, S. Baker, K. E. Holt, Genomic surveillance for hypervirulence and multi-drug resistance in invasive Klebsiella pneumoniae from South and Southeast Asia. *Genome Med.* **12**, 11 (2020).

169. H. Samarakoon, S. Punchihewa, A. Senanayake, J. M. Hammond, I. Stevanovski, J. M. Ferguson, R. Ragel, H. Gamaarachchi, I. W. Deveson, Genopo: a nanopore sequencing analysis toolkit for portable Android devices. *Communications Biology*. **3**, 538 (2020).

170. R. M. Leggett, D. Heavens, M. Caccamo, M. D. Clark, R. P. Davey, NanoOK: multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. *Bioinformatics*. **32**, 142–144 (2015).

171. E. L. Moss, D. G. Maghini, A. S. Bhatt, Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. Biotechnol.* **38**, 701–707 (2020).

172. S. H. Nguyen, M. D. Cao, L. J. M. Coin, Real-time resolution of short-read assembly

graph using ONT long reads. *PLoS Comput. Biol.* **17**, 1–18 (2021).

173. M. D. Cao, D. Ganesamoorthy, A. G. Elliott, H. Zhang, M. A. Cooper, L. J. M. Coin, Streaming algorithms for identification pathogens and antibiotic resistance potential from real-time MinION™ sequencing. *Gigascience*. **5** (2016), doi:10.1186/s13742-016-0137-2.

174. A. T. Dilthey, C. Jain, S. Koren, A. M. Phillippy, Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nat. Commun.* **10**, 3066 (2019).

175. F. J. Whelan, B. Waddell, S. A. Syed, S. Shekarriz, H. R. Rabin, M. D. Parkins, M. G. Surette, Culture-enriched metagenomic sequencing enables in-depth profiling of the cystic fibrosis lung microbiota. *Nature Microbiology*. **5**, 379–390 (2020).

176. R. M. Leggett, C. Alcon-Giner, D. Heavens, S. Caim, T. C. Brook, M. Kujawska, S. Martin, N. Peel, H. Acford-Palmer, L. Hoyles, P. Clarke, L. J. Hall, M. D. Clark, Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens. *Nature Microbiology*. **5**, 430–442 (2020).

177. A. Z. Broder, in *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)* (1997), pp. 21–29.

178. Z. Andrei, I. Broder, Filtering Near-Duplicate Documents, COM'00: Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching (2000).

179. N. T. Pierce, L. Irber, T. Reiter, P. Brooks, C. T. Brown, Large-scale sequence comparisons with sourmash. *F1000Res.* **8**, 1006–1006 (2019).

180. G. A. Blackwell, M. Hunt, K. M. Malone, L. Lima, G. Horesh, B. T. F. Alako, N. R. Thomson, Z. Iqbal, Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *bioRxiv* (2021), doi:10.1101/2021.03.02.433662.

181. J. A. Lees, S. R. Harris, G. Tonkin-Hill, R. A. Gladstone, S. W. Lo, J. N. Weiser, J. Corander, S. D. Bentley, N. J. Croucher, Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* **29**, 304–316 (2019).

182. M. E. Pitt, S. H. Nguyen, T. P. S. Duarte, H. Teng, M. A. T. Blaskovich, M. A. Cooper, L. J. M. Coin, Evaluating the genome and resistome of extensively drug-resistant Klebsiella pneumoniae using native DNA and RNA Nanopore sequencing. *Gigascience*. **9** (2020), doi:10.1093/gigascience/giaa002.

183. A. Prjibelski, D. Antipov, D. Meleshko, A. Lapidus, A. Korobeynikov, Using SPAdes De Novo Assembler. *Curr. Protoc. Bioinformatics*. **70**, e102 (2020).

184. M. M. C. Lam, R. R. Wick, S. C. Watts, L. T. Cerdeira, K. L. Wyres, K. E. Holt, Genomic surveillance framework and global population structure for Klebsiella pneumoniae. *bioRxiv* (2021), doi:10.1101/2020.12.14.422303.

185. S. M. Nicholls, J. C. Quick, S. Tang, N. J. Loman, Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience*. **8** (2019).

186. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

187. P. Liu, P. Li, X. Jiang, D. Bi, Y. Xie, C. Tai, Z. Deng, K. Rajakumar, H.-Y. Ou, Complete genome sequence of Klebsiella pneumoniae subsp. pneumoniae HS11286, a multidrug-resistant strain isolated from human sputum. *J. Bacteriol.* **194**, 1841–1842 (2012).

188. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 2114–2120 (2014).

189. I. Guthridge, S. Smith, P. Horne, J. Hanson, Increasing prevalence of methicillin-resistant *Staphylococcus aureus* in remote Australian communities: implications for patients and clinicians. *Pathogen*. **51**, 428–431 (2019).

190. R. L. Mork, P. G. Hogan, C. E. Muenks, M. G. Boyle, R. M. Thompson, M. L. Sullivan, J. J. Morelli, J. Seigel, R. C. Orscheln, J. Bubeck Wardenburg, S. J. Gehlert, C.-A. D. Burnham, A. Rzhetsky, S. A. Fritz, Longitudinal, strain-specific *Staphylococcus aureus* introduction and transmission events in households of children with community-associated meticillin-resistant *S. aureus* skin and soft tissue infection: a prospective cohort study. *Lancet Infect. Dis.* **20**, 188–198 (2020).

191. M. Matuszewska, G. G. R. Murray, E. M. Harrison, M. A. Holmes, L. A. Weinert, The Evolutionary Genomics of Host Specificity in Staphylococcus aureus. *Trends Microbiol.* **28**, 465–477 (2020).

192. R. Venkatvasan, P. X. Antony, H. K. Mukhopadhyay, V. Jayalakshmi, V. M. Vivek Srinivas, J. Thanislass, S. Stephen, Characterization of methicillin - Resistant Staphylococcus aureus from goats and their relationship to goat handlers using multi-locus sequence typing (MLST). *Small Rumin. Res.* **186**, 106097 (2020).

193. Z. Bi, C. Sun, S. Börjesson, B. Chen, X. Ji, B. Berglund, M. Wang, M. Nilsson, H. Yin, Q. Sun, A. Hulth, Y. Wang, C. Wu, Z. Bi, L. E. Nilsson, Identical genotypes of community-associated MRSA (ST59) and livestock-associated MRSA (ST9) in humans and pigs in rural China. *Zoonoses Public Health*. **65**, 367–371 (2018).

194. O. Sakwinska, M. Giddey, M. Moreillon, D. Morisset, A. Waldvogel, P. Moreillon, Staphylococcus aureus Host Range and Human-Bovine Host Shift. *Appl. Environ. Microbiol.* **77**, 5908 (2011).

195. A. Agabou, Z. Ouchenane, C. Ngba Essebe, S. Khemissi, M. T. E. Chehboub, I. B. Chehboub, A. Sotto, C. Dunyach-Remy, J.-P. Lavigne, Emergence of nasal carriage of ST80 and ST152 PVL+ Staphylococcus aureus isolates from livestock in Algeria. *Toxins* . **9**, 303 (2017).

196. D. Kühnert, T. Stadler, T. G. Vaughan, A. J. Drummond, Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data. *Mol. Biol. Evol.* **33**, 2102–2116 (2016).

197. N. F. Müller, R. R. Bouckaert, Adaptive Metropolis-coupled MCMC for BEAST 2. *PeerJ*. **8**, e9473 (2020).

Appendix 1: Preprints

Chapter 2.1: Phylodynamic signatures in the emergence of CA-MRSA

Attached pages (1st document)

Chapter 2.2: Phylodynamic modelling of bacterial outbreaks using nanopore sequencing

Attached pages (2nd document)

Chapter 2.3: Sketchy - genomic neighbor typing for bacterial outbreak surveillance

Attached pages (3rd document)

Appendix 2: Supplementary publications

Infectious disease genomics

Staphylococcus aureus from patients with chronic rhinosinusitis show minimal genetic association between polyp and non-polyp phenotypes

Attached pages (4th document)

Taking the next-gen step: comprehensive antimicrobial resistance detection from *Burkholderia pseudomallei*

Attached pages (5th document)

Global scale dissemination of ST93: a divergent Staphylococcus aureus epidemic lineage that has recently emerged from remote Northern Australia

Attached pages (6th document)

Exploring the evolution and epidemiology of European CC1-MRSA-IV: tracking a multidrug-resistant community-associated methicillin-resistant *Staphylococcus aureus* clone

Attached pages (7th document)

Bioinformatics

Nanoq: minimal but speedy quality control of nanopore reads in Rust
Attached pages (8th document)

Population genomics

Development and validation of a RAD-Seq target-capture based genotyping assay for routine application in advanced black tiger shrimp (Penaeus monodon) breeding programs
Attached pages (9th document)

Identification of key contributors in complex population structures
Attached pages (10th document)