Short- and long-read based resolution of complete bacterial genomes with applications in outbreak analysis and tracking of resistance genes

Inaugural dissertation

for the attainment of the title of doctor
in the Faculty of Mathematics and Natural Sciences
at the Heinrich Heine University Düsseldorf

presented by

Sebastian Alexander Fuchs
from Nordhorn

Düsseldorf, Mai 2024

from the Institute of Medical Microbiology and Hospital Hygiene
at the Heinrich Heine University Düsseldorf

Published by permission of the Faculty of Mathematics and Natural Sciences
at the Heinrich Heine University Düsseldorf

Supervisor: Herr Prof. Dr. Alexander T. Dilthey

Co-Supervisor: Herr Prof. Dr. Gunnar W. Klau

Date of the oral examination: 06.11.2024

Statement

I declare under oath that I have produced my thesis independently and without any undue assistance by third parties under consideration of the “Principles for the Safeguarding of Good Scientific Practice at Heinrich Heine University Düsseldorf”.

Düsseldorf, Mai 2024

Sebastian A. Fuchs

List of publications

1. Alexander T. Dilthey, Sebastian A. Meyer, Achim J. Kaasch.
   Ultraplexing: Increasing the efficiency of long-read sequencing for hybrid assembly with k-mer-based multiplexing
   Published in: Genome Biology 21, Article number: 68 (2020)

2. Carolina Silva Nodari, Sebastian Alexander Fuchs, Kyriaki Xanthopoulou, Rodrigo Cayô, Harald Seifert, Ana Cristina Gales, Alexander Dilthey, Paul G. Higgins.
   pmrCAB Recombination Events among Colistin-Susceptible and -Resistant *Acinetobacter baumannii* Clinical Isolates Belonging to International Clone 7
   Published in: ASM Journals, mSphere, Vol. 6, No. 6 (2021)

3. Sebastian A. Fuchs, Lisanna Hülse, Teresa Tamayo, Susanne Kolbe-Busch, Klaus Pfeffer, Alexander T. Dilthey.
   NanoCore: Core-genome-based bacterial genomic surveillance and outbreak detection in healthcare facilities from Nanopore and Illumina data
   Submitted to: ASM Journals, mSystems (December 2023)

Abstract

Research in bacterial pathogen genomics has witnessed significant advancements in sequencing technologies. In the realm of bacterial genomics, our work addresses both bioinformatic challenges, as well as cost and labour constraints, associated with the use of novel long-read sequencing technologies within three different projects.

First, introducing Ultraplexing, we present a method that substantially reduces per-sample sequencing costs and hands-on time in Nanopore sequencing for hybrid assembly. Ultraplexing eliminates the need for molecular barcoding bs bioinformatically determining which specific sequenced isolate a long-read belongs to; this is done by comparing each long-read to the k-mer spectrum of the sequenced isolates, measured using Illumina data. This method holds promise for large-scale bacterial genome projects that utilize hybrid assembly strategies, enabling considerable savings without compromising assembly quality. These advantages are enabled by the possibility to multiplex at least 100 isolates together, representing roughly fourfold increase of isolates possible at the time of publication, thus also reducing hands-on time in the lab by a factor of four.

Second, shifting focus to the hospital associated pathogen *Acinetobacter baumannii*, we investigate genome plasticity and horizontal gene transfer mechanisms in the context of transmission of colistin resistance elements. Through short- and long-read sequencing and creation of hybrid assemblies, we identify two probable recombination events in the *pmrCAB* operon, which confers colistin resistance. Our findings highlight the role of homologous recombination and shed light on the possible contribution of mobile genetic elements to this phenomenon in *A. baumannii*. This study contributes to the understanding of antibiotic resistance dynamics in clinical isolates of *A. baumannii*, specifically those belonging to International Clone 7.

Third, expanding the scope to genomic pathogen surveillance in healthcare facilities, we introduce NanoCore, a user-friendly method developed for Nanopore-based outbreak surveillance and investigation. NanoCore enables the determination and visualization of cgMLST-like sample distances directly from raw Nanopore reads by mapping Nanopore data to a core genome reference, variant-calling and calculating distances from the results, thus offering a fast and flexible solution. Validated on methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant *Enterococcus faecium* (VRE) datasets, NanoCore demonstrates high accuracy, producing results quasi-identical to those of current gold-standard tools with an average difference of 0.75 alleles for MRSA and 0.81 alleles for VRE in Nanopore-only-mode and 3.44 and 1.95 alleles respectively in hybrid-mode (measured in closely related isolates). The computational efficiency, open-source availability, and user-friendly installation via bioconda make NanoCore a valuable tool for effective bacterial pathogen surveillance in healthcare settings.

In conclusion, the work presented in this thesis spans the development of methods for hybrid genome assembly, long-read-based genomic surveillance and the investigation of the transmission of antibiotic resistance elements. The presented work demonstrates the potential of combining data generated by different sequencing technologies for bacterial genomics, as well as the potential of bioinformatics methods development for emerging sequencing technologies.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor Prof. Dr. Alexander T. Dilthey for his support and guidance during my time as a PhD student. Our meetings and countless conversations afforded me to find my way in the field of bioinformatics and next generation sequencing and to find the right direction for my different projects and this thesis.

Also, I would like to thank my mentor Prof. Dr. Gunnar W. Klau for his support and insight into the more algorithmic parts of bioinformatics, as well as for the possibility to have someone to mediate between me and my supervisor, even if this was never necessary.

Further, I want to thank my co-supervisor Prof. Dr. med. Klaus Pfeffer not only for giving me the opportunity to do my PhD in his institute, but also for giving me his honest opinion on projects and results whenever I needed it.

I would like to thank Prof. Dr. med. Achim J. Kaasch for being my former supervisor and his support during that time, as well as for giving me insight into the more medical-based side of microbes and bioinformatics.

I also want to thank all colleagues and former colleagues from my workgroup for keeping up my mood and mental health by providing a great working environment and for their support with different projects or topics whenever I needed an open ear. I would like to especially express my thank to the following scientists, even if some of them are not part of my workgroup, but collaborators from other workgroups: Anna Rommerskirchen and Dr. Sebastian Scharf, who are members of my office and always help with a second opinion if I get stuck on a particular difficult thought. Lisanna Hülse, who is a former member of my office and produces all the sequencing data I need for my different projects. Dr. Torsten Houwaart, who is a former member of my workgroup and always helped me with some programming languages, about which he had a better understanding than me. Johannes Ptok, who guides me whenever I have new issues with our servers, in which he has more insight.

Next, I want to thank the collaboration partners, whom I did not mention yet and who worked with me on different projects and collaborated in publishing the manuscripts that are the basis for my thesis. Namely and in the same order as listed in the publications: Dr. Carolina S. Nodari,

Dr. Kyriaki Xanthopoulou, Dr. Rodrigo Cayô, Prof. Dr. Harald Seifert, Prof. Dr. Ana C. Gales, Dr. Paul G. Higgins, Dr. Teresa Tamayo, and Dr. Susanne Kolbe-Busch.

I would like to voice a special thank you to my parents Gerhard and Martina. They always supported me and paved the way for me to become the person I am today. I owe them more than I can put into words.

Finally, from the bottom of my heart, I would like to express my gratitude to my wonderful wife Sabrina. Her love and support helped me, when I struggled with my research and her motivation kept me on track, whenever I threatened to stray too far. Especially in the last few years, during phases which just seemed to be overloaded with brainwork, she managed to keep me at balance. Thank you for being by my side!

Contents

List of figures and tables

List of abbreviations

BAM	Binary alignment map
BLAST	Basic local alignment search tool
cgMLST	Core genome multi-locus-sequence-typing
DNA	Deoxyribonucleic acid
HGT	Horizontal gene transfer
ICEs	Integrative and conjugative elements
InDel	Insertion or deletion
IS	Insertion sequence
MGE	Mobile genetic element
MLST	Multi-locus-sequence-typing
MRSA	Methicillin-resistant Staphylococcus aureus
NGS	Next generation sequencing
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PCR	Polymerase chain rection
SAM	Sequence alignment map
SNP	Single-nucleotide-polymorphism
VCF	Variant call format
VRE	Vancomycin-resistant Enterococcus
WGS	Whole-genome sequencing

Chapter 1: Introduction

DNA sequencing, the determination of genomic information, originated in 1977 with the introduction of Sanger sequencing by Frederick Sanger and colleagues (1). Subsequently, diverse sequencing methods, such as Illumina (2), Pacific Bioscience (3) and Oxford Nanopore (4) have been developed and refined up to the present day. Each method possesses unique advantages and disadvantages, encompassing factors like speed, costs, fragment length, and method complexity (5). While at first sequencing was limited by both sequencing costs and the length of the sequencable fragments, technological advancements of methods and sequencers permitted the efficient sequencing of whole genomes and large numbers of isolates (6).

Precise knowledge of an organism's genome enables the classification of carried genes, identification of other genomic motifs and pinpointing single nucleotide polymorphisms. Furthermore, the availability of high-quality sequencing data or a resolved genome is an important requisite for many bioinformatical applications such as algorithms for base-calling, mapping, assembling or variant-calling (7,8). Today, it is common practice in research or hospital settings to search for resistance genes based on specific sequences or analyse potential outbreaks using genomic similarities (9,10), enabling for example applications like epidemiology characterizations and outbreak investigations (11). Progress in genomics also established novel fields, like genome-based personalized healthcare (12).

Bacterial genomes, particularly significant in these contexts, are often carriers of clinically relevant diseases affecting patient treatment, which become even more problematic, if said genomes also contain resistance genes (13–15). Consequently, it is crucial to interrogate bacterial genomes to our fullest capacity.

Despite their apparent simplicity compared to eukaryotic DNA, bacterial genomes, clinically speaking, have great potential to negatively influence patient treatment. They for example normally have shorter genomes (1 to 10 mb compared to 3 to 5.000 mb), a smaller gene number (below 10.000 compared to often above 10.000) and shorter intergenic regions (below 100 bp compared to often above 100 kb) (16). However, despite seeming easier to analyse, bioinformatically speaking, the unravelling of bacterial genomes can be fairly intricate, due to genomic complexities like repeats or gene duplications.

Within this dissertation two central topics tare addressed: First, the comprehensive cost-effective resolution and assembly of large numbers of bacterial genomes; second, the application and development of methods for analysing bacterial genome sequencing data with respect to tracking infection chains and the spread of antibiotic resistance elements.

Given that both inquiries are profoundly influenced by high-throughput sequencing developments, the scope of recent changes in sequencing technologies will be given particular consideration.

This dissertation makes significant contributions by enhancing existing methods or introducing novel applications through the synergistic use of short-read and long-read sequencing methods (Illumina and Oxford Nanopore Technologies) in diverse hybrid approaches:

In the first publication titled “Ultraplexing: Increasing the efficiency of long-read sequencing for hybrid assembly with k-mer-based multiplexing” (17), we aimed at a solution for cost-efficient sequencing despite the constraints imposed by the limited number of barcodes available for Nanopore sequencing. Thus, a novel method was developed to enable barcode-less pooling of isolates for Nanopore sequencing. For the assignment of sequencing data to corresponding samples this method utilizes k-mer statistics based on inter-sample genetic variability of the produced reads in comparison to barcoded Illumina sequencing data. This increases the number of isolates that can be sequenced together on the Nanopore platform at least by a factor of 4 of what would be possible with barcodes, thereby improving sequencing speed and decreasing sequencing costs.

The second publication “*pmrCAB* Recombination Events among Colistin-Susceptible and -Resistant *Acinetobacter baumannii* Clinical Isolates Belonging to International Clone 7” (18), elucidates the genomic structure of *A. baumannii* isolates, particularly around the *pmrCAB* operon which plays an important role in the acquisition of colistin resistance. This was achieved by combining Illumina short-read and Oxford Nanopore Technologies long-read sequencing data in a hybrid assembly to create fully resolved genomes instead of highly fragmented genomes on Illumina basis alone and by the calculation of k-mer statistics for said assemblies, thus enabling the detailed analysis of said operon and the surrounding genomic region to identify possible recombination events.

In a third yet unpublished manuscript titled “NanoCore: Core-genome-based bacterial genomic surveillance and outbreak detection in healthcare facilities from Nanopore and Illumina data”, a novel method is introduced to enable whole-genome Nanopore-based outbreak investigations and expedite them in comparison to established whole genome cgMLST methods, which are solely Illumina-based. Utilizing the speed and improved error rate of present-day Oxford Nanopore sequencing, this open-source method provides a faster alternative to established Illumina-based gold-standard methods, while retaining a similar level of accuracy, and a more accurate alternative to existing Nanopore-based methods, which are limited to an analysis of small sets of housekeeping genes.

These projects were undertaken to address technical limitations of the Nanopore sequencing technology and to enable and demonstrate the utilization of Nanopore sequencing for novel applications.

Initially, our focus was on addressing a significant and surmountable issue related to the limited availability of barcodes for standard Oxford Nanopore Technologies sequencing (19), thereby targeting a concern inherent to the sequencing process itself. Given that multiplexing numerous samples with Illumina sequencing poses no challenges due to the availability of a sufficient number of barcodes (e.g. 386 barcodes for a standard NexteraXT library), we leveraged small sequence difference between isolates, measured using the Illumina technology, to determine which sample a long-read belongs to.

Subsequently, our aim was to determine sequence homologies and potential recombination events. For this purpose, we processed the generated sequencing data into fully resolved assemblies, which highlights the capabilities of Nanopore sequencing in a use-case situated in the initial phase following sequencing. Assemblies based solely on Illumina data often suffer from pronounced fragmentation, making it challenging to resolve repetitive regions or establish the gene order (20,21). To overcome this limitation, we conducted comprehensive assemblies using both Illumina and Nanopore data. This approach demonstrated the feasibility of uncovering regions of horizontal gene transfer in crucial resistance genes. These identified regions could potentially contribute to the transmission of resistances to susceptible strains.

Our final objective centered on exploring the potential of Nanopore long-reads for conducting outbreak investigations, in comparison to established Illumina-based methods, an application involving several steps (or consecutively applied tools) beyond sequencing. This has been attempted in various approaches, each with its own limitations. The inherent difficulty lied in the higher error rate of Nanopore compared to Illumina sequencing (22,23). However, with the improved performance (for example read accuracy, read length and number of InDels) of present-day Oxford Nanopore technology (24), we achieved results on both Nanopore and Illumina basis that closely resemble those obtained using the established cgMLST method Ridom SeqSphere+ (10), which compares Illumina data to a core genome, while reducing the time required for that analysis.

Chapter 2: Background

The upcoming chapter delves into various biological and bioinformatic subjects and mechanisms essential for comprehending the projects elucidated in this thesis.

2.1 The structure of bacterial genomes

DNA, the carrier of the genetic information in all living beings, consists of four nucleotide bases: adenine, cytosine, guanine, and thymine (Figure 2.1 A). These bases form specific sequences known as genes, translating into distinct amino acid chains crucial for the functioning of every organism (25). In contrast to human DNA, comprising approximately 6.3 gigabases distributed across 23 pairs of chromosomes in the form of diploid double-stranded helixes (Figure 2.1 B) (16,26), bacterial genomes are normally characterized by a single chromosome. This haploid double-stranded helix is often arranged in a circular structure, with sizes ranging from approximately 0.6 to 8.0 megabases (Figure 2.1 C). Additionally, bacteria frequently harbour one or multiple plasmids – circular, extrachromosomal DNA molecules. These plasmids typically vary in size from a few kilobases to several hundred kilobases and often carry genes providing a selective advantage to the organism, such as antibiotic resistances (27).

2.1.1 Small scale variation in bacterial genomes

Another important term in the field of bacterial genomes is nucleotide mutation: Errors that occur during the DNA replication of organisms (40). Various types of mutations exist, including single nucleotide polymorphisms (alteration of only one base), insertions and deletions of nucleotides, as well as the copying or relocation of sequence elements to alternative positions (41,42). Such variations can contribute to diseases or enhance the genetic diversity of a species (43). In the context of this thesis “small variations” can also be used to investigate the relatedness of closely-related isolates of the same species.

2.1.2 Variation in gene content and mechanisms of gene transfer

The pangenome, encompassing the entirety of all genes in a bacterial organism, is divided into two primary components: the core genome and the accessory genome. The core genome comprises genes that are essential for metabolism and are present in all strains of a particular species. In contrast, the accessory genome consists of genes that are absent in at least one of the known strains of that species. (28,29). The composition of the core and accessory genomes exhibits considerable variation not only between different species (Table 1) but also within a given species.

Table 2.1: Variation in gene content in six clinically relevant bacteria.

Bacterial species	Pangenome size (genes)*	Core genome size (genes)*	Accessory genome size (genes)*	Source
Acinetobacter baumannii	4062	2390	1672	(30,31)
Enterococcus faecium	3915	1423	2492	(30,32)
Escherichia coli	4401	2513	1888	(30,33)
Klebsiella pneumoniae	5500	2358	3142	(30,34)
Pseudomonas aeruginosa	9786	3867	5919	(30,35)
Staphylococcus aureus	4360	1861	2499	(30,36)

*Pan, core, and accessory genome values can also vary within the same species, depending on which study is considered.

Those genes are often disseminated among bacteria through horizontal gene transfer (HGT), a non-sexual exchange of genetic information originating from the chromosome or plasmids. HGT stands as the primary mechanism for acquiring antibiotic resistances (37). Possible recombination events that can lead to HGT are transformation (the uptake and incorporation of genetic material from the bacteria's surroundings), conjugation (the transfer of genetic material between bacterial cells through a direct connection) and transduction (the introduction of foreign genetic material into a bacterium by viruses or viral vectors). Typically, these events are facilitated by mobile genetic elements (MGE), for example ICEs (which contain the conjugation machinery) or IS elements (the smallest autonomous unit). These are small segments of DNA sequence inside the genome capable of movement within the genome or transfer between genomes of different organisms (38,39).

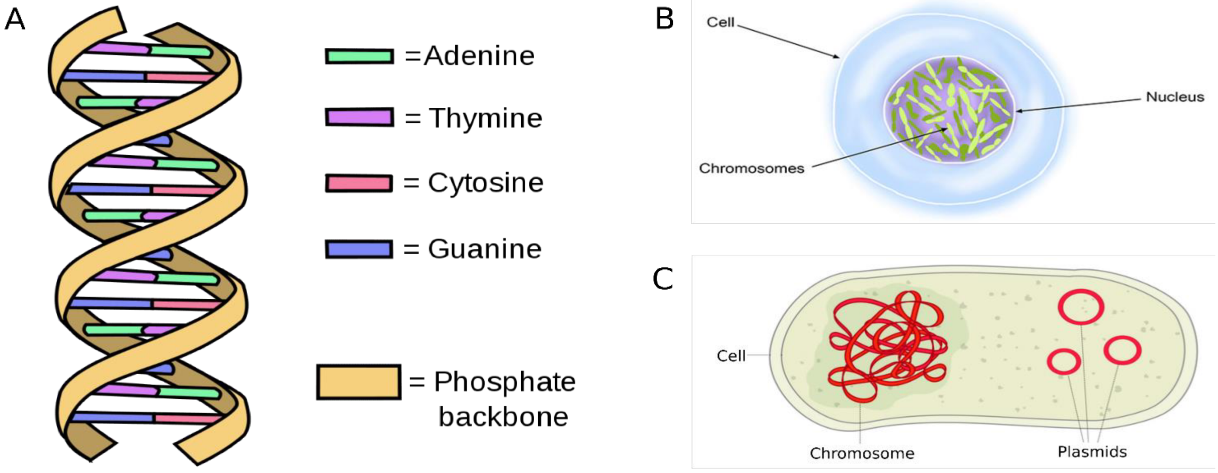


Figure 2.1: DNA in humans and bacteria. **A**. The structure of the DNA helix and the four bases. **B**. A human cell and the genetic information as 23 chromosomes within the nucleus. **C**. A bacteria and the genetic information as a chromosome and multiple plasmids within the cell.
Figure adapted from (44–46).

2.2 Sequencing bacterial genomes and the processing of the produced data

The first sequencing technique, known as Sanger sequencing (1), was developed in 1977 by Frederick Sanger and colleagues. Subsequent to this milestone, various methods for the determination of genomic information were developed, each exhibiting distinct advantages and areas of application (5). Generally, sequencing methods aim to identify the sequence of nucleotide bases within DNA and transform it into a machine-readable format, known as reads, which can be subjected to further processing through various methodologies. Before the sequencing process commences, DNA must undergo preparatory steps in the laboratory (47). This typically involves cultivating bacterial isolates overnight, extracting DNA from the cultured isolates, and adjusting the DNA concentration to adhere to the specifications of the desired sequencing protocol.

Sequencing can be conducted either on a single-isolate basis (48), where each isolate is individually cultivated and then sequenced individually; using barcodes in combination with other isolates, which is called multiplexing; or on a metagenomic basis (49), where a potentially diverse pool of organisms present in a particular community is sequenced collectively without prior sorting.

2.2.1 Gold-standard sequencing methods

Currently applied sequencing methodologies include Illumina (2), Pacific Biosciences (3) and Oxford Nanopore Technologies (4).

Illumina, developed in 1994 by Bruno Canard and Simon Sarfati (2) stands out as a leading short-read sequencing technique that has rapidly become a dominant force in the field, exhibiting a wide range of applications. Its primary strength lies in its high basecalling accuracy, exceeding 99.9% (50). This property ensures tasks such as genotyping and variant-calling can be performed with a high degree of confidence. However, the method's limitation arises from the short fragment length, often resulting in fragmented genome assemblies when carrying out *de novo* assembly (21).

Briefly, the Illumina sequencing process (Figure 2.2) involves fragmenting the DNA, attaching adapters to the fragments, and loading the adapter-fragment pairs onto a flow cell. Subsequently, the fragments undergo multiple rounds of duplication through bridge amplification PCR. Modified fluorescent nucleotides are then attached to the fragments by a DNA polymerase. Following each polymerase step, a photograph is captured, with the added nucleotide identified by its unique wavelength.

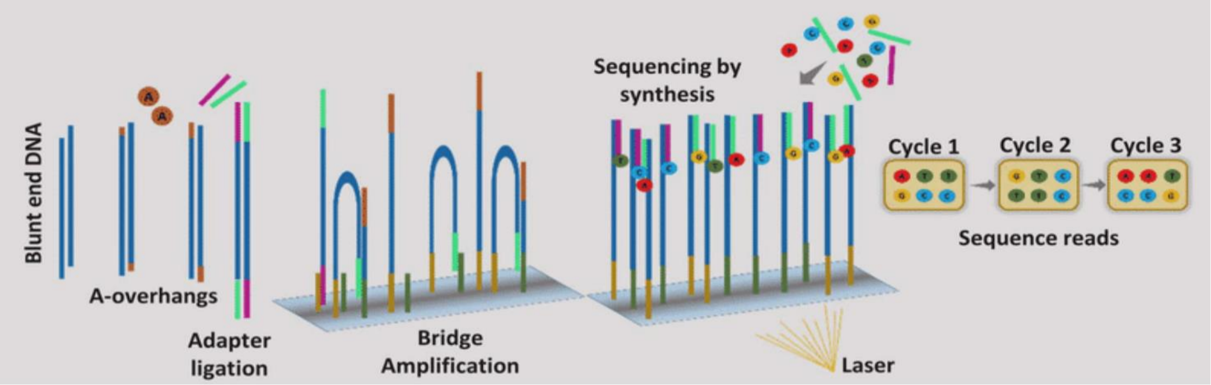


Figure 2.2: Illumina Sequencing. The basic steps of the Illumina sequencing workflow from the fragmentation of the DNA to the identification of the attached nucleotides.
Figure adapted from (51).

Oxford Nanopore Technologies (ONT) is a long-read sequencing technique conceptualized by David

Deamer in 1989 (52). Its practical implementation came to fruition in 2005 when Gordon Sanghera, Spike Willcocks, and Hagan Bayley founded Oxford Nanopore. Notable advantages of ONT include a rapid data generation and the capability to sequence long fragments of DNA (53). These attributes make tasks such as whole-genome assembly or the non-fragmented sequencing of larger sequences for applications such as the resolution of repeat-rich regions more manageable or possible at all, compared to short-read methods. However, an early challenge was the high error rate associated with Nanopore sequencing, particularly affecting tasks like genotyping that are susceptible to sequencing errors. Substantial progress has been achieved in this area in recent years. Although the current basecalling accuracy exceeds 99%, it still falls slightly behind that of Illumina sequencing. Nevertheless, ONT can now produce results comparable to Illumina-based methods (24,54).

In essence, the ONT sequencing process (Figure 2.3) involves fragmenting the DNA and attaching adapters to the fragments. A motor protein is then affixed to a fragment, guiding it towards and through a nanopore. As each nucleotide passes through the pore, the current within the pore undergoes changes, and the passing nucleotide is identified by its unique current alteration.

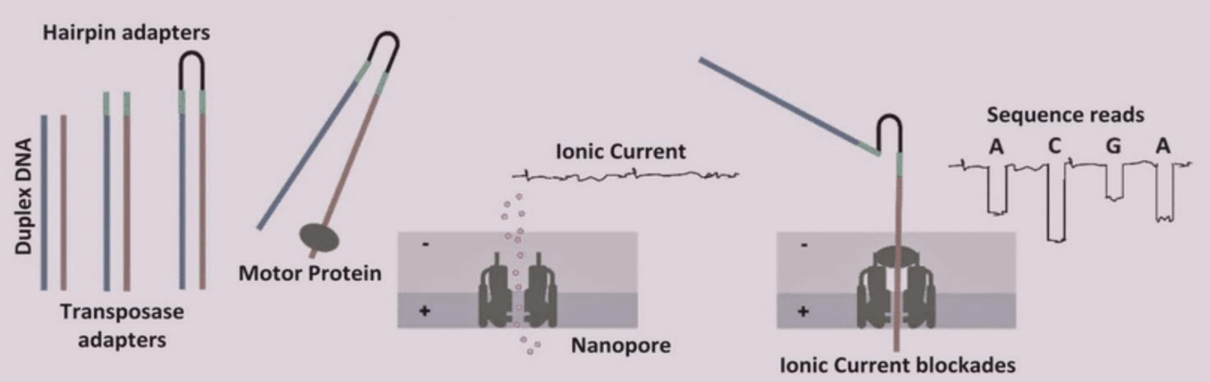


Figure 2.3: Nanopore Sequencing. The basic steps of the Nanopore sequencing workflow from the fragmentation of the DNA to the identification of the nucleotide sequence.
Figure adapted from (51).

Pacific Biosciences, commonly known as PacBio, stands as another long-read sequencing technique that was introduced for commercial use in 2010 by the biotechnology company bearing the same name. PacBio has both advantages and disadvantages similar to those of ONT (55). However, this method is not further described here, as it was not employed in any of the projects detailed in this dissertation.

2.2.1.1 Multiplexing

For cost-effective sequencing and efficient utilization of sequencing throughput, a procedure called multiplexing is normally applied to each of the above describes sequencing methods.

For this purpose, a barcode, a sample-specific identifiable fragment of DNA, is attached to the DNA molecules of each prepared sample before the sequencing process to later identify which read belongs to which sample. This is comparably cost-efficient for Illumina sequencing, with e.g. 386 available barcodes for a standard NexteraXT library and potentially more barcodes depending on the used workflow and sequencing kit, but less efficient for Nanopore sequencing with its currently limited number of 24 to 96 barcodes.

The standardized output format generated by sequencing methods is the FASTQ format. A FASTQ file comprises four lines for each read. The first line starts with the "@" symbol, followed by a sequence identifier and optional descriptions. The second line presents the raw sequence. The third line features the "+" symbol, optionally followed by the sequence identifier from the first line. Lastly, the fourth line provides quality values corresponding to the sequence in the second line, with one quality character per sequence letter. The structure of a single read entry might resemble the following:

```
@Sequence_Identifier
CAACTGTATAATATGGTCAAAATATATGAGATG
+
111>AFFFFFBF3FGGGFGGGGFDGBGF311FG
...
```

Regardless of the sequencing method employed, but depending on how this sequencing data should be processed, different methods of genome inference can be utilized in deducing the genomic content within the genome or even reconstructing the original genome structure. Common algorithms of genome inference employed for such purposes include assemblers, mappers, and variant-callers.

2.2.2 Assembly

Assembly is the process of combining sequence fragments, known as reads, into larger sequences called contigs. These reads are derived from quasi-random positions along a sequenced genome. When sufficient coverage is achieved, these reads overlap, allowing the merging of multiple reads into larger fragments, potentially reconstructing the entire genome. The ease of the assembly process is influenced by factors such as the length of the reads and their error rate. There are two primary assembly strategies: *de novo* assembly (Figure 2.4 A) and reference-guided assemblies (Figure 2.4 B) (56,57). *De novo* assembly relies solely on the overlaps of different reads, making the process more challenging but eliminating reference-based errors. In contrast, reference-guided assemblies align reads not only to each other but also to a reference genome closely related to the sequenced organism. While potentially easier, this method may introduce biases if the chosen reference does not align well with certain areas of the organism's genome.

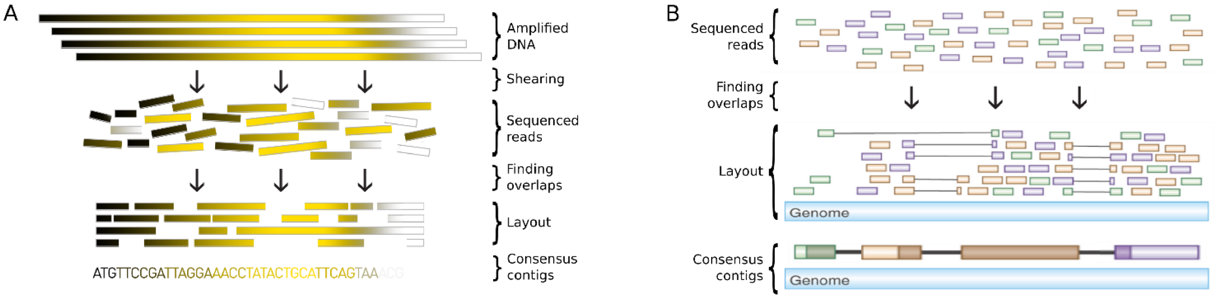


Figure 2.4: Assembly strategies. Simplified workflows of a de-novo and a reference guided assembly method. **A.** Sequenced reads are overlapped with each other and merged into one or multiple consensus contigs. **B**. Sequenced reads are mapped onto a reference genome and merged into one or multiple consensus contigs.
Figure adapted from (58,59).

2.2.3 De Bruijn graphs and overlap graphs

A De Bruijn graph is a directed graph that represents the relative order of sequence sub-fragments of length k (k-mers) as found in an input-set of sequencing reads. These graphs find applications in diverse tasks, including the assembly of sequencing data into genomes (60) or the assignment of reads to a specific dataset (17).

Briefly, De Bruijn graphs are constructed by segmenting sequencing data or a genome into overlapping sequence fragments of a specific length, termed k-mers. These k-mers are then represented as nodes. An edge between the nodes representing k-mers x and y is added if and only if, in the input sequence, x starts at position j and y at j+1. When two otherwise identical sequences exhibit a polymorphism, this introduces a so-called bubble into the graph structure (Figure 2.5). Adding new sequences to the graph involves identifying identical k-mers within the existing graph and introducing new branches where the k-mers do not align. Assembling a genome from such a graph entails navigating the longest possible route through the graph along branches with the most support (61,62). Such navigation is termed an eulerian walk, which is defined as a walk through a connected (so called "eulerian") graph that visits each edge exactly once (63). However, due to the fact that sequencing normally is imperfect, not every constructed de Bruijn graph is also an eulerian graph, e.g. when the graph is split into multiple parts du to sequencing errors or coverage issues. Thus, for the purpose of assembly the eulerian walk was changed into the de Bruijn superwalk (64), which contains each read as a subwalk, but is not dependent on fully connected graphs anymore.

An overlap graph operates similarly, with the distinction that sequencing data is not divided into k-mers. Instead, each read is represented in its entirety as one node, and the edges depict overlaps between these nodes (61,62). In comparison to the de Bruijn graph, this method is better at resolving repeats, since it does not split reads into shorter k-mers, which also removes read coherency from the graph. On the other hand, an overlap graph takes longer to construct and needs more space.

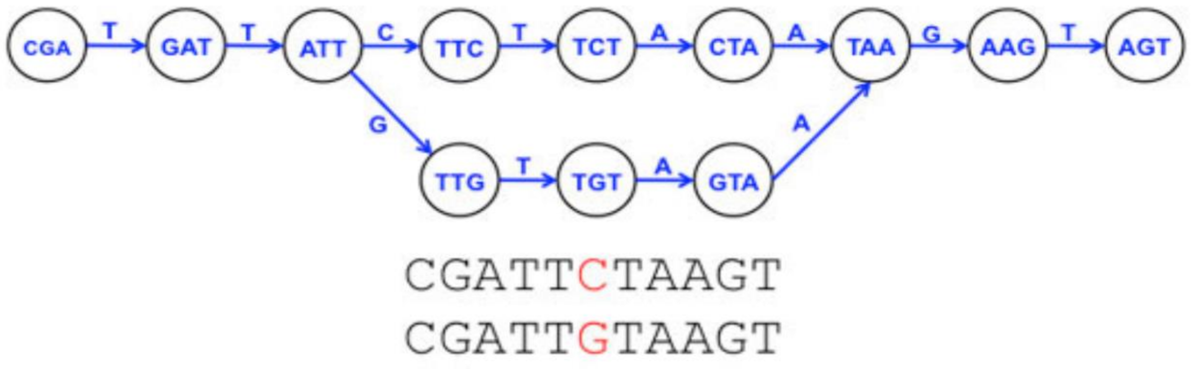


Figure 2.5: De Bruijn graph. A short depiction of a De Bruijn graph. Reads are displayed through an overlapping succession of k-mers. Here for "k" the length 3 is chosen. Disagreeing positions are presented by a forking in the succession.
Figure adapted from (65).

The standard output format typically generated by these sequencing data processing methods is the FASTA format. A FASTA file consists of two lines for each sequence, whether it represents an entire genome or a fraction of a genome, known as a contig. The format comprises the following: First, the ">"

symbol, followed by a description of the sequence; second, the raw sequence, either as nucleotides or amino acids. For enhanced readability, long sequences originally include line breaks after a fixed number of positions, dividing the raw sequence into more than one line. The structure of two sequence entries might appear as follows:

```
>Sequence_Identifier 1
CAACTGTATAATATGGTCAAAATATATGAGATG
>Sequence_Identifier 2
TTTTGAGTTAGTTTTAAGCGCATTAGTAGCGGGCGCTA
...
```

2.2.4 Mapping

Mapping is the process of identifying the best sequence alignment, subject to a scoring function, between a query sequence (a read) and a larger reference (a genome)(66). This enables the assignment of the query to a specific location within the reference. Typically, the reference is obtained beforehand by assembling high-quality and high-coverage sequencing data into a comprehensive genome or by downloading a similarly processed reference genome from a database. Mapping processes use the so called "Seed and Extend" method (67). "Seeding" is the determination of a preferably exact match between a substring of the read and the genome. Different algorithms either use fixed length seeds or maximal exact matching seeds. "Extension" is the process of prolonging the mapped seeds on 5' and 3' side (and ideally linking multiple seeds) by finding matches between the genome and a read that contains the corresponding seed. In this process, mismatches and InDels are potentially accepted and assessed depending on the used scoring function. Different techniques that are applied are global alignment algorithms, local alignment algorithms and BLAST-like seed extension (67).

Given that mapping tools must be able to handle a large number of nucleotide strings with varying lengths, efficient techniques for finding seeds have become crucial. Two common mechanisms employed for this purpose are the Burrows-Wheeler transformation and the Minimizer. The Burrows-Wheeler transformation rearranges the bases of a read in a reversible manner, enhancing text compression efficiency and the effective search for exact matches (68). Conversely, the Minimizer transforms the reads and the reference into subsets of short k-mers and searches for overlaps between them, while preserving information about sequence contiguity, resulting in an efficient technique for finding subsequent exact matches (69).

Mapping applications include scenarios such as mapping single reads onto reference genomes or core genes for e.g. variant calling (Figure 2.6). Various algorithms exist for diverse purposes, each handling nucleotide differences in its own way.

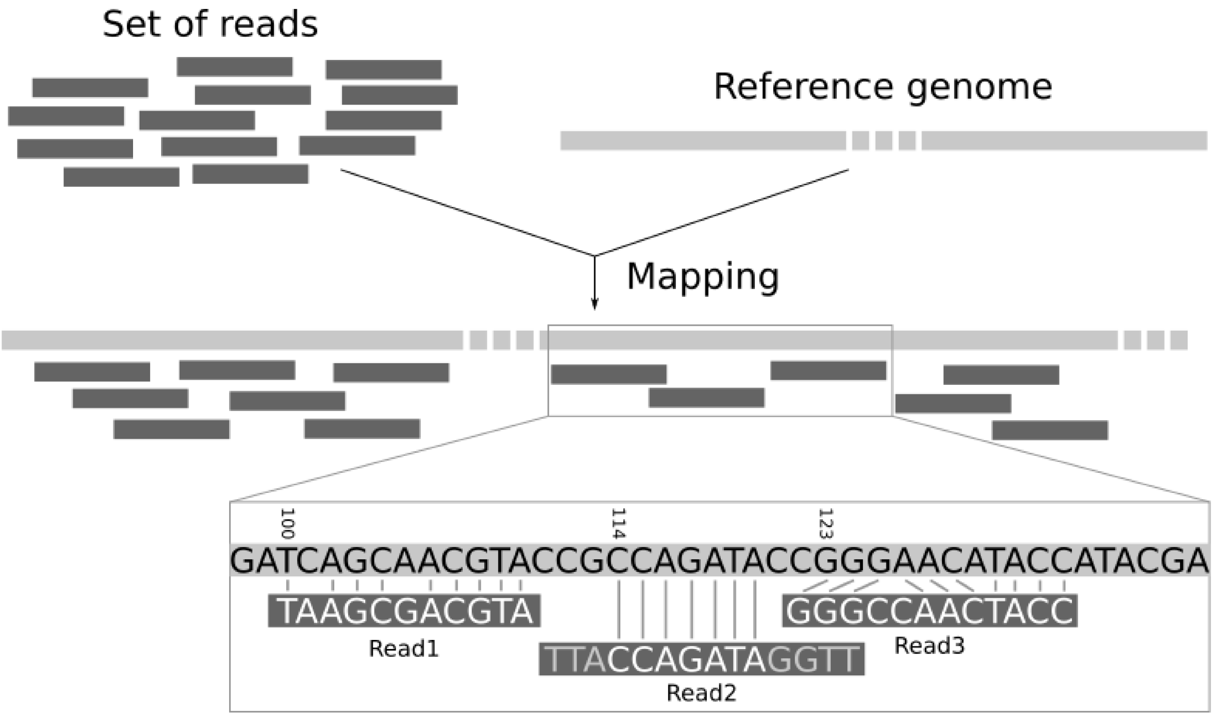


Figure 2.6: Mapping. A simple presentation of the mapping of reads to a reference genome. A perfect match is not always required for a read to be mapped.
Figure adapted from (70).

In general, the output formats generated by such mappings are the "Binary Alignment Map" (BAM) format and the "Sequence Alignment Map" (SAM) format (62). A BAM file comprises a header section that includes the names of the used files and/or samples, along with overall mapping process details such as the length of the reference or the used method. Additionally, the format involves an alignment section containing information about the corresponding mapped read, including its name, sequence, read quality, position within the reference, and more. While the BAM file is compressed and not directly human-readable, it can be reformatted into a SAM file, which contains the same information in an uncompressed and human-readable format.

2.2.5 Variant calling

The process of identifying differing positions, known as genetic variations or variants, from mappings of sequencing data to a reference genome is termed variant-calling. In this context, nucleotide discrepancies between sequences are assessed based on the ratio of reads carrying these variations, the base quality and potentially sequence context. Differences between reads and the utilized reference genome in regions with low coverage for example (e.g., only one mapped read) could be sequencing errors. Differences that are found in only one read in a region with adequate coverage, as well.

For variant-calling, two primary families of algorithms have been developed: likelihood-based callers and machine-learning-based callers.

Likelihood-based variant callers (71) estimate the likelihood of observed sequencing data under different statistical models. These models often rely on specific distributions, such as the binomial or Poisson distribution, and estimate parameters like sequencing error rates. One of their advantages is accuracy, particularly when the assumptions of the statistical models align well with the data. However, they may encounter challenges when handling complex patterns or complex base contexts.

Machine learning-based variant callers (72) employ algorithms that learn patterns from annotated training data. Consequently, these methods require a set of training data with known variant status to train the model. The flexibility of machine learning models enables them to adapt to a broader range of data patterns, which is advantageous in complex scenarios. While the accuracy of these methods heavily depends on the quality and representativeness of the training dataset, algorithms that use machine learning, like DeepVariant (73), currently belong to the state-of-the-art techniques of variant-calling.

The file format generated by variant-calling methods is known as the "variant call format" (VCF). A VCF file comprises different sections: A meta-information part, each line beginning with "##", including general information about the utilized variant caller, versions, parameters, and abbreviations used within the file; a header line, beginning with “#”, containing the column titles for the following data; and a variant section, listing information about the chromosome and the exact position where the variant was found (one variant per line). It also details how the reference and alternative alleles appear, the quality of the variant call, and various values relevant for potential filtering or statistical analysis. The structure of two variant entries might resemble the following:

```
##fileformat=VCFv1.1
##fileDate=20240704
##reference=file/path/to/ref/file.fasta
#CHROM  POS     REF     ALT     QUAL    FILTER  INFO    FORMAT  SAMPLE
Cont_1  550     A       G       30.23   PASS    F       GT:AF   1/1:0.9718
Cont_6  729     C       T       27.65   PASS    P       GT:AF   1/1:0.9597
…
```

2.3 Selected analyses of clinical interest regarding bacterial genomes

The bioinformatic analysis of bacterial genomes encompasses various aspects, including the already introduced sequencing methods and processing of sequencing data into assemblies. In this dissertation, two additional aspects will be addressed: the exploration of acquisition mechanisms of antibiotic resistances and the investigation of potential outbreaks in a clinical context.

2.3.1 Acquisition of antibiotic resistances

Resistance genes play a pivotal role in the survivability of bacteria, particularly within healthcare settings (74). Understanding different mechanisms of antibiotic resistances and their acquisition is thus crucial for human health. Illustrative examples include the vancomycin resistance of *Enterococcus faecium* (75) and the methicillin resistance of *Staphylococcus aureus* (76), both notorious and challenging hospital pathogens (77). Such resistance genes can be located in the chromosome or plasmids, and their functional mechanisms exhibit considerable variability (78).

The acquisition and dissemination of resistances are often facilitated by mobile genetic elements (MGEs), which transfer genes via horizontal gene transfer (HGT) to other organisms (refer to background 2.1). An example relevant in the context of this thesis is the *pmrCAB* operon of *Acinetobacter baumannii*. This operon comprises three genes, *pmrC*, *pmrA*, and *pmrB*, which, in colistin-resistant *A. baumannii*, underwent specific mutations. These modifications result in a higher colistin resistance. In brief, specific mutations in *pmrA* and pmrB lead to a constitutive activation of the *pmrAB* system, followed by an upregulation of the entire *pmrCAB* operon and other operons, ultimately enhancing colistin resistance (79,80).

Normally, the pmrCAB genotype is linked to the international clone group of the corresponding isolate (81). However, during the examination of an isolate belonging to international clone group 7(81), the identification of certain single nucleotide polymorphisms (SNPs) within the *pmrB* gene strongly suggested a relatedness to a reference genome of international clone group 4. This observation led to the hypothesis of horizontal gene transfer, and to the questions, if recombination boundaries could be mapped exactly and if certain MGEs are involved in this, initiating publication number two.

2.3.2 Detection and investigation of outbreaks

An outbreak investigation involves identifying links between pathogens, potential transmission pathways, and assessing the relatedness of pathogens among hosts. In the modern context, where pathogens may become more pathogenic due to factors like accumulated resistances, a rapid and accurate analysis of transmission chains is of paramount importance, also for infection prevention and control measures (82,83). Examples are the implementation of SARS-CoV-2 sequencing to investigate health-care associated cases (84), which helped to detect cryptic transmission events and identify opportunities to further reduce health-care associated infections, or the analysis of methicillin resistant *Staphylococcus aureus* on a special care baby unit (85), which uncovered transmission links within the baby unit, between mothers on a postnatal ward and to a staff member during periods without known infection, thus making it possible to prevent further infections through these pathways. Independent of the exact use-case, the faster an outbreak analysis is completed (while retaining high accuracy), the faster infection prevention measures can be implemented, resulting in reduced spread, and reducing patient harm. Important requirements are thus a high quality of the processed sequencing data, leading to more accurate genetic distances; as well as a rapid sequencing and analysis pipeline, leading to faster results.

With the advent of sequencing, outbreak investigations can be successfully conducted using genomic data, by sequencing isolates of interest and comparing their genomes, either based on raw sequencing data or assemblies. This involves quantification and interpretation of the similarity between isolates; one approach for quantifying the similarity is a phylogenetic tree, and interpretation of the tree may be based on a threshold of distances expected for closely related isolates. Phylogenetic trees or minimum-spanning-trees with clusters of closely related isolates can then be linked to information about potential transmission pathways, like patients on the same ward (86). Multiple methods have been developed for measuring the similarity of isolates in an outbreak.

Three established methods for this purpose are multi-locus-sequence-typing (MLST), core genome multi-locus-sequence-typing (cgMLST) and whole genome single-nucleotide-polymorphism (SNP) genotyping.

In an MLST scheme (87) the housekeeping genes (typically around 7) of a species are compared to a database, and a sequence type is assigned to each analysed isolate, based on the genotype of these genes. Here, each gene is treated as a unit and no distinctions are made if a gene has one ore multiple SNPs. The advantages of this method are its speed and cost-efficiency, since only a small amount of specific DNA is needed. However, this method has limitations in accuracy, comparing only a small subset of potentially multiple thousands of genes, and requires a curated MLST database.

A cgMLST scheme (88) on the other hand utilizes the entire core genome (usually few thousand genes) while treating each gene as a single unit, regardless of the number of variants found within. This approach is potentially less time- and resource-efficient than MLST, but offers a higher accuracy with good sensitivity, as core genes are expected to be present in all isolates of a species. Traditionally cgMLST analysis pipelines require a curated database, however, it is possible to construct cgMLST-like pipelines without a database, by skipping the assignment of specific genotypes, and instead comparing the pattern of detected variants.

Lastly, a whole-genome SNP genotyping scheme (89) completely abandons the distinction between genic and intergenic sequences when computing pairwise sample distances, focusing solely on the presence or absence of single-nucleotide polymorphisms and their differences between compared isolates. While this potentially adds a higher discriminatory power, the comparison of SNP positions itself, as well as the analysis of the accessory genome or non-coding regions (90), without necessarily increasing time- or resource-consumption in comparison to the cgMLST (both need whole genome sequencing data), it does not consider gene boundaries, possibly leading to decreased comparability, due to the fact, that not all isolates, per definition, contain all accessory genes or even intergenic regions (91).

This underscores the preference for an analysis centered on core genes to enhance comparability. Therefore, for publication number three a cgMLST-like approach was deemed the appropriate system for outbreak investigations.

Of note, genes, particularly those in the core genome, are subject to a consistent evolutionary pressure, so called purifying selection (92), unlike non-coding regions, where the likelihood of mutations emerging is higher.

Chapter 3: Publications

3.1 Publication 1: “Ultraplexing: increasing the efficiency of long-read sequencing for hybrid assembly with k-mer-based multiplexing”

Manuscript published in: Genome Biology

Impact factor: 13.583 (2020)

DOI: https://doi.org/10.1186/s13059-020-01974-9 (17)

Authorship: Shared first author

Contribution to the publication according to CRediT classification:

- Conceptualization (supporting)
- Data curation (equal)
- Formal analysis (lead)
- Investigation (lead)
- Methodology (supporting)
- Project administration (supporting)
- Software (equal)
- Validation (lead)
- Visualization (lead)
- Writing - original draft preparation (equal)
- Writing - review and editing (equal)

Dilthey *et al. Genome Biology* (2020) 21:68
https://doi.org/10.1186/s13059-020-01974-9

Genome Biology

SOFTWARE Open Access

Ultraplexing: increasing the efficiency of long-read sequencing for hybrid assembly with *k*-mer-based multiplexing

Alexander T. Dilthey[1,2*†], Sebastian A. Meyer[1†] and Achim J. Kaasch[1,3*]

Abstract

Hybrid genome assembly has emerged as an important technique in bacterial genomics, but cost and labor requirements limit large-scale application. We present Ultraplexing, a method to improve per-sample sequencing cost and hands-on time of Nanopore sequencing for hybrid assembly by at least 50% compared to molecular barcoding while maintaining high assembly quality. Ultraplexing requires the availability of Illumina data and uses inter-sample genetic variability to assign reads to isolates, which obviates the need for molecular barcoding. Thus, Ultraplexing can enable significant sequencing and labor cost reductions in large-scale bacterial genome projects.

Keywords: Bacterial genomics, Genome assembly, Assembly graph, Multiplexing, *k*-mer, Hybrid assembly, Barcoding

Background

Accurate characterization of large numbers of microbial genomes is becoming increasingly important in microbiology. For example, bacterial genome-wide association studies (bGWAS) rely on the sequencing of large numbers of samples to correlate genetic variants to phenotypes such as antibiotic resistance or virulence [1–3]. Further examples are phylogenetic analyses and quality assurance in industrial microbiology [4–7].

A variety of sequencing technologies with different technological trade-offs have emerged for the sequencing of microbial genomes. Short-read sequencing technologies (such as Illumina [8] have low error rates (< 0.1%) but provide only limited resolution of complex and repetitive genomic regions. Examples are the genes encoding *S. aureus* protein A (*spa*) and fibronectin binding-protein (*fnbpA*), which play key roles in the pathogenesis of *S. aureus* [9] and which cannot be reliably assembled from short-read data [10]. Long-read sequencing technologies (Pacific Biosciences [11], Oxford Nanopore [12]) generate sequencing reads of tens or even hundreds of kilobases in length, enabling the correct structural resolution of complex regions; their higher error rates (5–15%), however, can negatively impact consensus and small-variant genotyping accuracy [13–15].

Combining short- and long-read data has therefore emerged as a standard approach for the resolution of bacterial genomes [16]. Long-read sequence information can be used to deconvolute short-read-based assembly graphs (hybrid de novo assembly [17–20]). Alternatively, de novo assemblies from long reads [21] can be polished with short-read data to improve consensus accuracy [22]. By either approach, the coverage requirements to arrive at a high-quality assembly of a microbial genome are typically modest (50–100× for each data type [23, 24]).

Molecular barcoding approaches enable the cost-effective sequencing of multiple samples in one run ("multiplexing"). Molecular barcoding involves the labeling of each DNA sample with a unique barcode sequence, pooling and joint sequencing of the samples, and determining the source sample for each sequencing read, based on its

* Correspondence: alexander.dilthey@med.uni-duesseldorf.de; achim.kaasch@med.ovgu.de
†Alexander Dilthey and Sebastian A. Meyer contributed equally to this work.
[1]Institute of Medical Microbiology and Hospital Hygiene, University Hospital, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany
Full list of author information is available at the end of the article

barcode sequences. Highly efficient, automated implementations of molecular barcoding exist for the Illumina platform, enabling the sequencing of hundreds of microbial isolates to sufficient coverage with a single flow cell. Molecular barcoding approaches for long-read platforms, however, are less effective. A maximum of 24 samples can currently be multiplexed on an Oxford Nanopore MinION flow cell using the manufacturer's kits for "native" (PCR-free) barcoding. In addition, the preparation of multiplex libraries requires significant hands-on time (> 12 h compared to 3 h for a non-multiplexed library) and comes with significant losses of input material, and presumably, the pipetting steps reduce attainable read lengths by shearing. These factors make barcoded long-read sequencing costly and labor-intensive, and the availability of a more scalable approach to multiplexed long-read sequencing would be highly desirable.

Here, we present Ultraplexing, a new method that allows the pooling of multiple samples in long-read sequencing without relying on molecular barcodes. Ultraplexing uses inter-sample genetic variability, as measured by Illumina sequencing, to assign long reads to individual isolates (Fig. 1). Specifically, each isolate genome is represented by its de Bruijn graph, constructed from sample-specific short-read data, and each long read is assigned to the sample de Bruijn graph it is most compatible with (or randomly in cases of a draw). A similar approach enables haplotype-aware assembly in eukaryotic genomes [25].

The intuition behind Ultraplexing is that there will typically be a high-quality alignment between a read and the assembly graph of the source genome it emanates from. Importantly, the assignment of reads completely contained in genomic regions shared among multiple samples (e.g., due to mobile genetic elements or inter-sample genetic homology) may remain ambiguous. This, however, will typically have no or only a small effect on the accuracy of the hybrid assembly process, for the affected reads will spell equally valid assembly graph traversals in all compatible samples.

Ultraplexing requires the availability of Illumina data. It is applicable to studies that either incorporate the generation of these from the beginning, or it can serve as a cost-effective method to generate additional long-read data for samples that have already been short-read sequenced. In the following, we demonstrate that Ultraplexing can match or even outperform classical molecular long-read barcoding approaches in terms of assembly quality while enabling significant reductions in cost and hands-on time.

Results

We used simulated and real Nanopore and Illumina sequencing data to evaluate the performance of Ultraplexing in the context of bacterial hybrid de novo assembly. In all experiments, we relied on Unicycler as an established method for hybrid assembly [17]. We primarily focused on the quality of the generated assemblies, i.e., structural accuracy (number of contigs, reference recall, assembly precision) and consensus accuracy (single nucleotide polymorphisms; SNPs), measured against the utilized reference genomes (in simulations) or barcoding-based assemblies (for real data). To distinguish between Ultraplexing-mediated effects and intrinsic assembly complexity for the selected isolates, we reported assembly accuracy for random (in all experiments) and perfect (in simulations) assignment of long reads. Additionally, we assessed the proportion of correctly assigned reads. Of note, all simulation experiments were based on conservative assumptions (e.g., 5 Gb throughput per long-read flow cell; see the "Methods" section for further details), and no mis-assemblies were identified through visual inspection in any of the Ultraplexing-based sets.

Simulation experiment I: Multi-species Ultraplexing

In a first step, we evaluated Ultraplexing on a sample of 10 different clinically important bacterial species (Additional file 1), covering a wide range of genome sizes (2.0–6.3 Mb), GC contents (32–60%), and between-species mash [26] distances (0.02–0.20; Additional file 2). The Ultraplexing algorithm assigned all but 2 of 477,890 simulated long reads to the correct bacterial isolate (close to 100% classification accuracy, Additional file 12: Figure S1). Ultraplexing-based assemblies were highly concordant (Additional file 12: Figure S1 and Additional file 3) with the underlying reference genomes, achieving near-perfect structural agreement (average reference recall and assembly precision > 99.999%) and low divergence (average number of SNPs against the reference genome, 57). Furthermore, assembly accuracy metrics for Ultraplexing and perfect read assignment were virtually identical (for example, an average of 57 SNPs for Ultraplexing compared to 56 SNPs for perfect assignment; Additional file 12: Figure S2). To assess how the performance of multi-species Ultraplexing was affected when combining more than one strain per species, we repeated the experiment for 5 clinically important species, each represented by 2 strains (Additional file 2) with mash distance < 0.01 (Additional file 2) [23]. Ultraplexing-based assemblies were virtually identical to assemblies based on perfect read assignment (for example, identical SNP count observed for 6/10 genomes) and of generally very high quality (Additional file 12: Figure S3 and Additional file 4), except for two *E. coli* genomes; in these, large repeat structures (Additional file 1) led two assembly fragmentation (> 100 contigs) for both Ultraplexing and perfect read assignment.

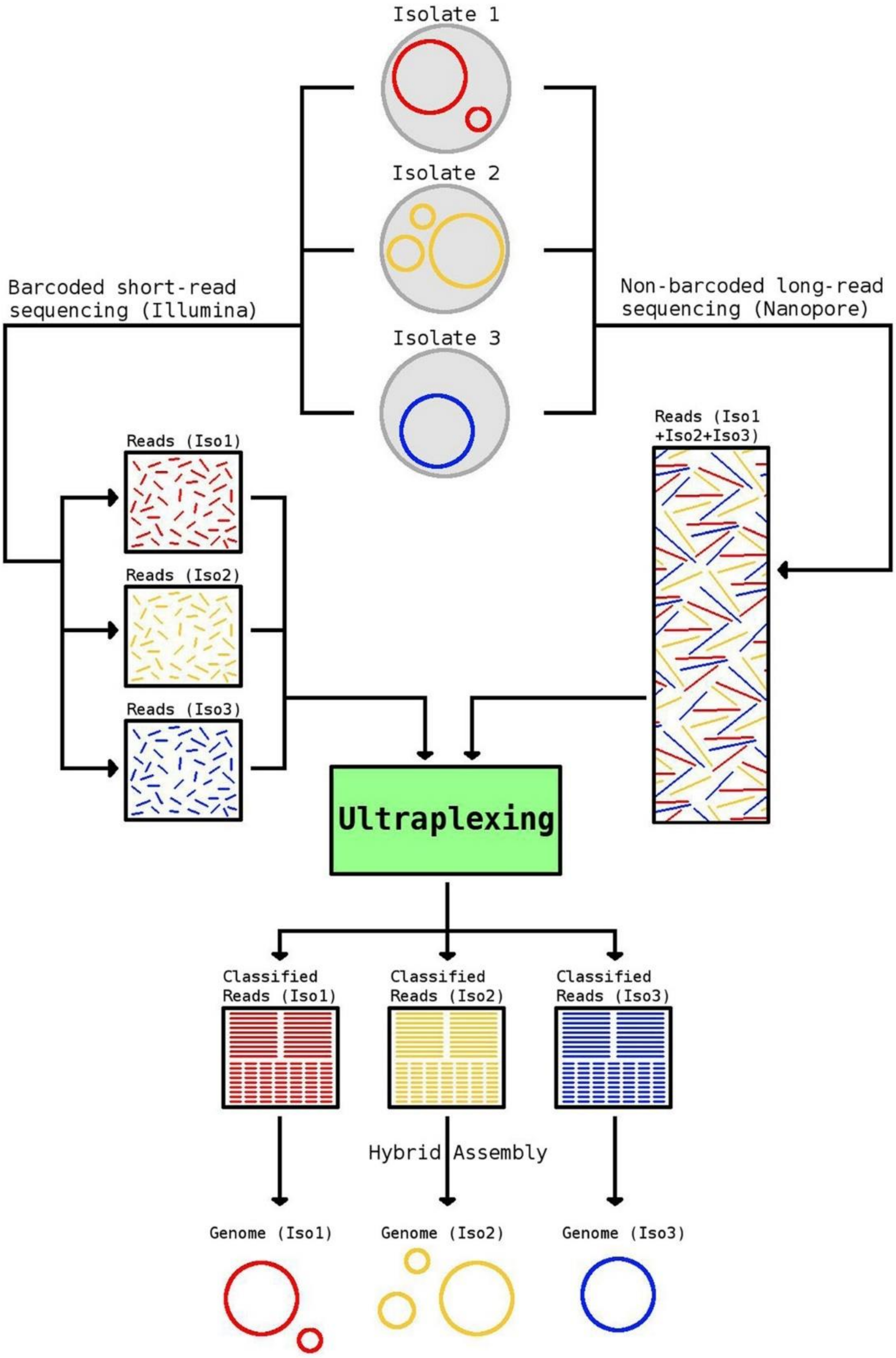


Fig. 1 Overview of the Ultraplexing approach. Long reads are generated in simple pooled sequencing runs. The Ultraplexing algorithm determines the most likely source genome for each long read by carrying out a comparison between the read and the de Bruijn graphs of the sequenced sample genomes, inferred from short-read data. Hybrid assembly of sample-specific long and short reads enables the recovery of complete bacterial genomes

Simulation experiment II: Single-species Ultraplexing with 10–50 isolates

To assess Ultraplexing performance on closely related isolates and with increasing sample numbers, we randomly selected sets of 10, 20, 30, 40, and 50 genomes from 181 publicly available complete assemblies of the human pathogen *Staphylococcus aureus* (Additional file 1). Of note, as simulated long-read flow cell capacity was held constant, sets with more genomes contained less long-read data per isolate. Across experiments, the proportion of correctly assigned reads decreased as sample numbers increased and varied between 35 and 95% (Fig. 2a). To test whether reduced read assignment accuracies were due to inter-sample sequence homologies, we computed the metric *Δedit distance* for random samples of mis-assigned reads and found an average *Δedit distance* of 0.3%, with more than 50% of mis-assigned reads exhibiting a *Δedit distance* of 0 (Fig. 2b). At the read alignment level, the genomes that the mis-assigned reads were assigned to are thus indistinguishable or very similar to the true source genomes. Consistent with this, the generated Ultraplexing-based assemblies were highly concordant with the utilized reference genomes (average reference recall ≥ 99.96% and assembly precision ≥ 99.99% across sets; average number of SNPs 46; Fig. 2c–f). Furthermore, assembly accuracy metrics for Ultraplexing and perfect read assignment were comparable even with increasing number of bacterial isolates; for example, the average number of SNPs per genome in the run with 50 bacterial isolates was 59 for Ultraplexing (QV 47) and 32 for perfect read assignment (QV 49). Complete results for this experiment are presented in Additional file 5 and visualized in Fig. 2. Finally, to evaluate to which extent assembly accuracy was influenced by genome complexity [23, 27], we repeated the experiment for 30 *S. aureus* isolates of class I complexity and for 30 *S. aureus* isolates of class III genome complexity (Additional file 1). Individual outliers in the set of class III genomes notwithstanding (Additional file 12: Figure S4), overall assembly quality remained high even for class III genomes (average reference recall, 99.98% for class compared to 99.86% for class III; average assembly precision, 100.00% for class I and III; average number of SNPs, 34 for class I and 77 for class III; Additional file 4). What is more, the quality of Ultraplexing-based assemblies remained comparable to that of assemblies based on perfect read assignment for class III genomes (for example, 77 SNPs on average for Ultraplexing, corresponding to QV 46, compared to 52 SNPs on average for perfect read assignment, corresponding to QV 47).

Simulation experiment III: Impact of plasmids

In addition to the chromosomal genome, many bacterial cells harbor plasmids. Plasmids are extrachromosomal circular strings of DNA that are generally much smaller than the chromosomal DNA. Plasmids can vary in copy number within each cell, and they often exhibit complex and repetitive sequence structures. Since plasmid sequences could reduce the performance of the Ultraplexing algorithm, we repeated the previous simulation experiments with sets of 10–50 *S. aureus* genomes that all harbored plasmids (Additional file 1; Additional file 12: Figure S5). We found that the accuracy of chromosomal genome assemblies was not affected by the presence of plasmids. Additionally, the plasmid recovery rate was comparable to assemblies based on reads assigned to their true source; complete recovery was achieved in 135 of 150 total isolate genomes with Ultraplexing, and in 137 with perfect read assignments. Identified reasons for incompletely recovered plasmids included high sequence homology to other plasmids or the genomic DNA (Additional file 6). Complete results for this experiment are presented in Additional file 7 and visualized in Additional file 12: Figure S6 (chromosomal genome) and Additional file 12: Figure S7 (plasmids). Finally, we further explored the impact of repeats between the chromosomal and plasmid genomes on a set of 10 complex (class III) *Pseudomonas* isolates, 9 of which harbor chromosome-plasmid repeats ranging from 669 bp to 69 kb in size (Additional file 1; Additional file 12: Figure S8). Assembly accuracy remained high at slightly reduced levels (reference recall > 97% and assembly precision > 99% for all 10 genomes), and Ultraplexing- and truth-based assemblies are almost identical in terms of accuracy metrics (identical reference recall for 10/10 isolates and identical assembly precision for 9/10 at very similar SNP levels; Additional file 4).

Real-data experiment I: Nanopore-based Ultraplexing of 10 *S. aureus* clinical samples

To assess the performance of Ultraplexing on real data, we randomly selected ten bacterial isolates of the species *Staphylococcus aureus* from our collection of clinical isolates. To generate a reference genome for each isolate, we sequenced each sample on an Illumina system, performed barcoded Oxford Nanopore sequencing with the 12-sample barcoding kit (~ 214× coverage per isolate; mean read length 8.3 kb), and carried out hybrid de novo assembly. The generated reference genomes consist of 1–3 circular contigs per isolate, representing the chromosomal genome (~ 2.8 Mb in length) and plasmids (2.3–34.9 kb in length, all circular; BLAST [28] classification results are shown in Additional file 8).

To test Ultraplexing on these isolates, we demultiplexed the barcoded Nanopore sequencing data with the Ultraplexing algorithm and carried out hybrid de novo assembly. The Ultraplexing-based assemblies showed a high degree of concordance (Fig. 3) with the generated reference genomes in terms of contig number, assembly length, genome structure (average reference recall and assembly precision > 99.9%), and consensus accuracy (4 SNPs per isolate on average and 6 of 10 isolates with no detected SNPs). In contrast, assemblies based on random read assignment yielded lower-quality assemblies across all considered metrics (for example, 136 SNPs per genome; Fig. 3d). Complete results for all genomes are presented in Additional file 9 and visualized in Fig. 3. Summary statistics of the Illumina and Nanopore sequencing runs can be found in Additional file 10.

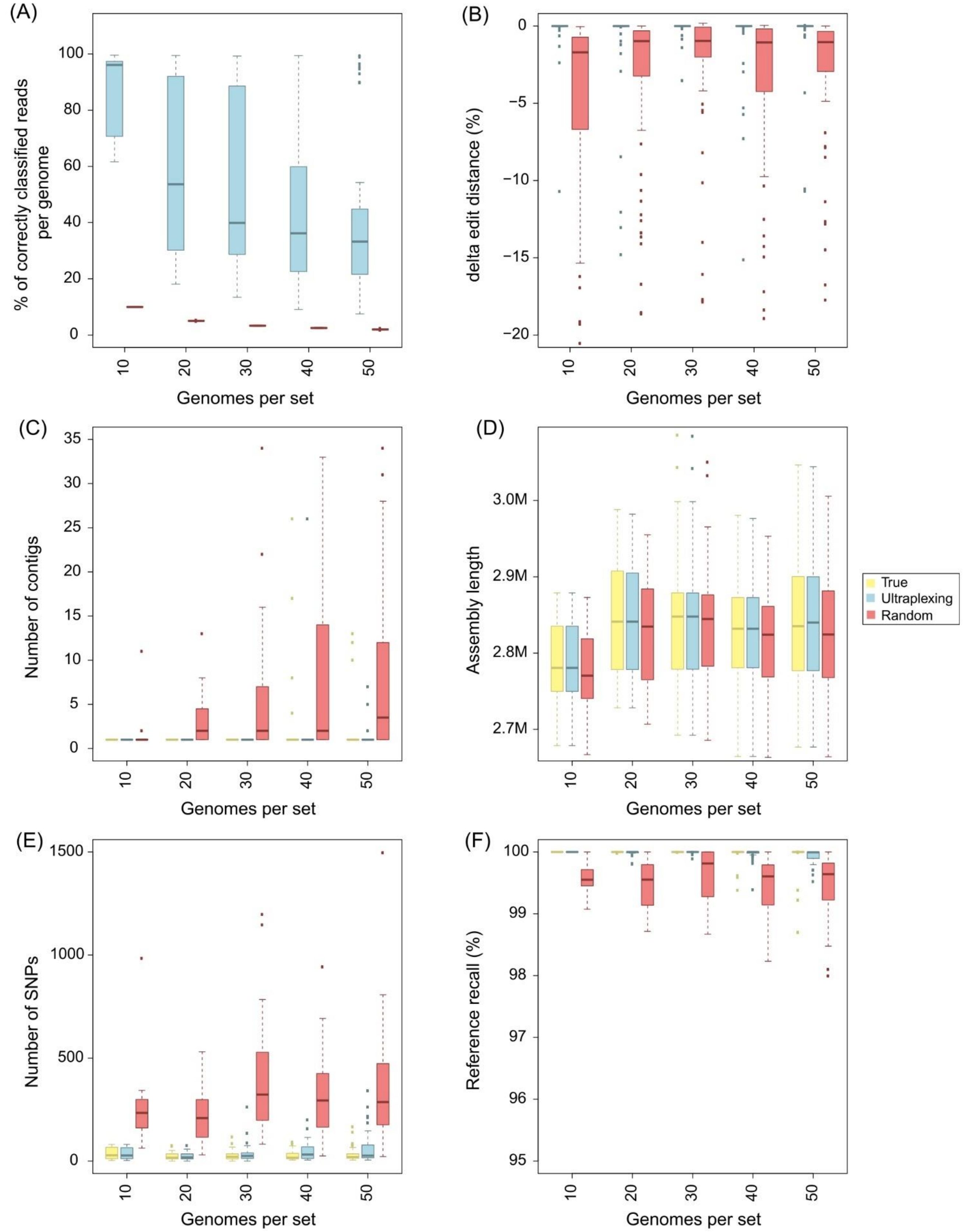


Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 Simulated Ultraplexing runs with 10–50 *S. aureus* genomes, in comparison to perfect (True) and random (Random) assignment of long reads. **a** The proportion of correctly assigned long reads. **b** The Δedit distance for random samples of falsely classified long reads. **c** The distribution of contigs per assembly. **d** The distribution of assembly lengths. **e** The distribution of SNPs per assembly. **f** The distribution of reference recall. SNPs and reference recall were calculated relative to the utilized reference genomes, and all metrics within the same set of genomes are based on the same simulated short-read data

Read-data experiment II: Nanopore-based Ultraplexing of 48 clinical isolates

To assess the feasibility of applying Ultraplexing to a larger number of samples, we repeated the previous experiment with 48 samples. As in the previous experiment, barcoded Nanopore (~ 446× coverage per isolate; average read length 10.4 kb) and Illumina (~ 44× coverage per isolate; 2 × 250 bp reads with MiSeq v2 chemistry) sequencing was carried out to generate reference genomes for the 48 samples.

For Ultraplexing, long-read sequencing data (~ 87× coverage per isolate; average read length 11.7 kb) was generated in a single MinION run by pooling DNA from the 48 isolates. Reads were demultiplexed with the Ultraplexing algorithm, and hybrid de novo assembly was carried out.

The generated assemblies exhibited a plausible profile in terms of assembly length, and for 29/48 assemblies, the Ultraplexing-based assembly had the same number of contigs as the generated reference genomes (Fig. 4). Further investigation showed a high degree of concordance between the Ultraplexing-based assemblies and the reference genomes both in terms of genome structure (average reference recall and assembly precision > 99.8%) and the number of SNPs per genome (126 on average, equivalent to QV 43). Complete results for the comparison of the 48 Ultraplexing-based assemblies against the reference genomes are presented in Additional file 9 and visualized in Fig. 4. Read length and coverage statistics for all sequencing runs can be found in Additional file 10; the read length distribution of all generated Nanopore sequencing runs is visualized in Additional file 12: Figure S9.

Discussion

We have presented Ultraplexing, a method that resolves pooled long-read sequencing data in the context of hybrid de novo assembly without the use of barcoding. Ultraplexing leverages inter-sample genetic variation to assign pooled long reads to individual isolates and benefits from the fact that Illumina sequencing enables the reliable characterization of the *k*-mer spectra of individual genomes.

Using simulated sequencing data, we demonstrated that Ultraplexing enables the generation of highly accurate hybrid assemblies and reliably detects plasmids, even in datasets that contain multiple isolates of the same bacterial species, complex plasmid-chromosome repeat structures, or genomes of high complexity. We have also validated the method on two real Nanopore sequencing datasets and shown that Ultraplexing-based assemblies are virtually identical to barcoding-based assemblies when comparing multiplexed runs with the same number of isolates; remaining errors in the assemblies based on both Ultraplexing and perfect read assignment may represent residual errors introduced by the hybrid assembly approach. When using Ultraplexing to increase the number of samples over the current maximum of PCR-free molecular barcoding approaches on the Nanopore platform, Ultraplexing-based assemblies generally maintain high accuracy.

A key advantage of Ultraplexing in comparison to molecular barcoding is decreased cost and hands-on time. The number of samples sequenced per flow cell can at least be doubled, and barcoding reagents are not necessary. Hands-on time was reduced eightfold in our 48-sample experiment (~ 5 h per flow cell with 10 barcoded samples compared to 3 h for one Ultraplexing run with 48 samples). Taking into account potential differences in sample handling operator performance, we conservatively estimate that the hands-on time benefit conferred by Ultraplexing is at least 50%.

On the other hand, Ultraplexing has a number of limitations. First, Ultraplexing can consume significant computational resources (70 CPU hours and 175 Gb of memory for the demultiplexing step in the experiment with 48 samples). Improvements in hands-on time do therefore not necessarily translate into decreased time-to-result. Second, Ultraplexing relies on Illumina data for read assignment and hybrid assembly; systematic biases in Illumina sequencing, as observed for certain bacterial genomes with high or low GC content [29], may affect the accuracy of Ultraplexing. Third, the application of Ultraplexing requires high molecular weight DNA, the extraction of which may be challenging for certain bacterial species. Fourth, while we have shown that Ultraplexing is generally robust against the presence of complex repeat structures, assembly accuracy was slightly reduced for class III genomes. For these reasons, the method is best suited to applications in which large numbers of genomes need to be resolved to very high, but not perfect, accuracy, and in which turnaround times on the order of 3–5 days are acceptable. Examples of this include bacterial genome-wide association studies and retrospective outbreak sequencing. For other applications, such as the generation of a small number of

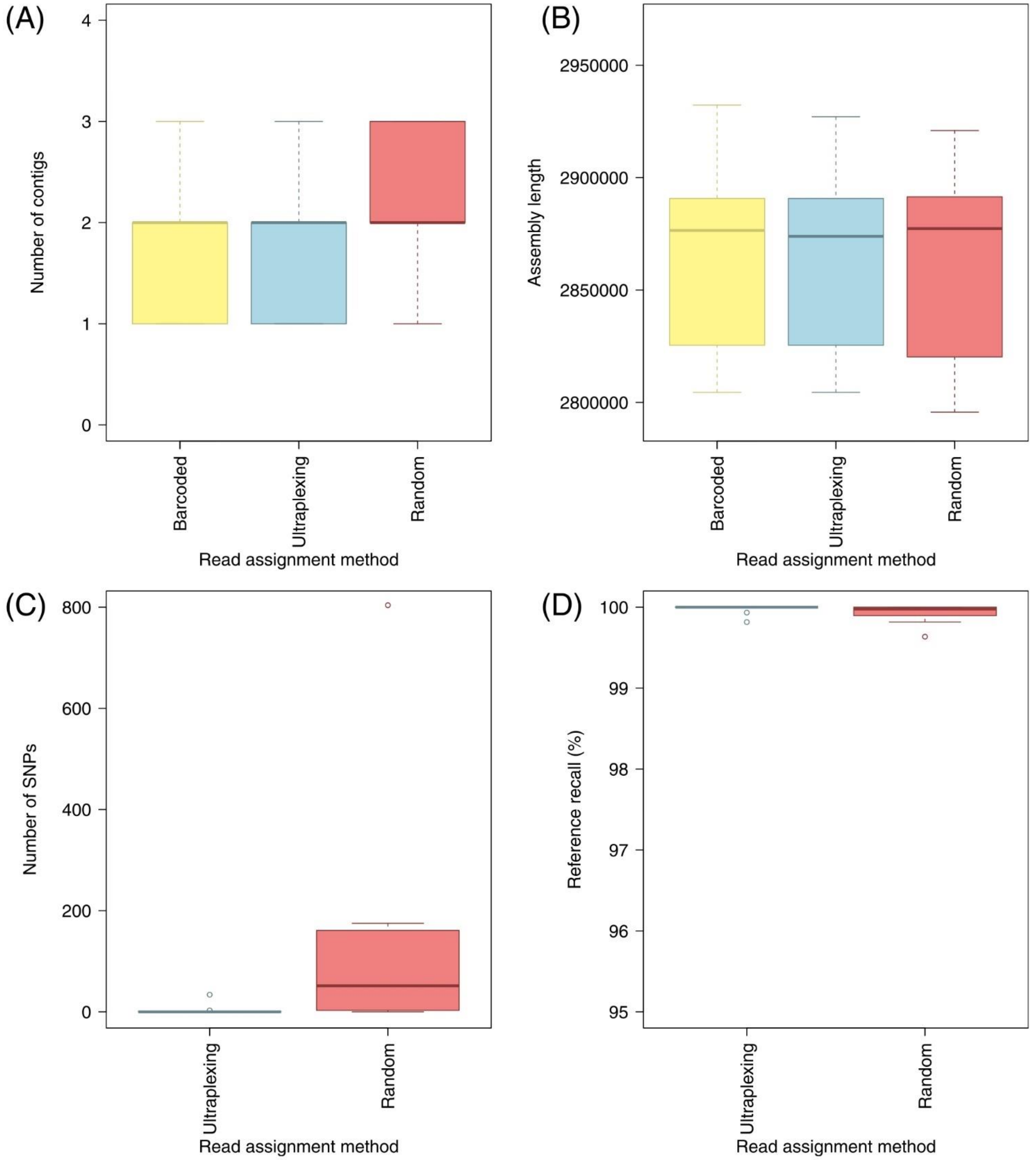


Fig. 3 Ultraplexing and classical molecular barcoding on a set of ten *S. aureus* isolates. For different read assignment methods applied to the same set of Nanopore reads, the distribution of contigs per assembly (**a**), the distribution of assembly lengths (**b**), the distribution of SNPs per assembly (**c**), and the distribution of reference recall (**d**) are shown. SNPs and reference recall were calculated relative to assemblies based on molecular barcoding, and the same Illumina sequencing data were used throughout. Barcoded, reads assigned according to molecular barcodes; Ultraplexing, reads assigned by the Ultraplexing algorithm; Random, reads assigned randomly

reference-grade assemblies or time-critical diagnostic applications, conventional barcoding approaches may remain preferable.

Although our primary focus was on assembly accuracy, we also evaluated the accuracy of individual read assignments in the simulation experiments. One important

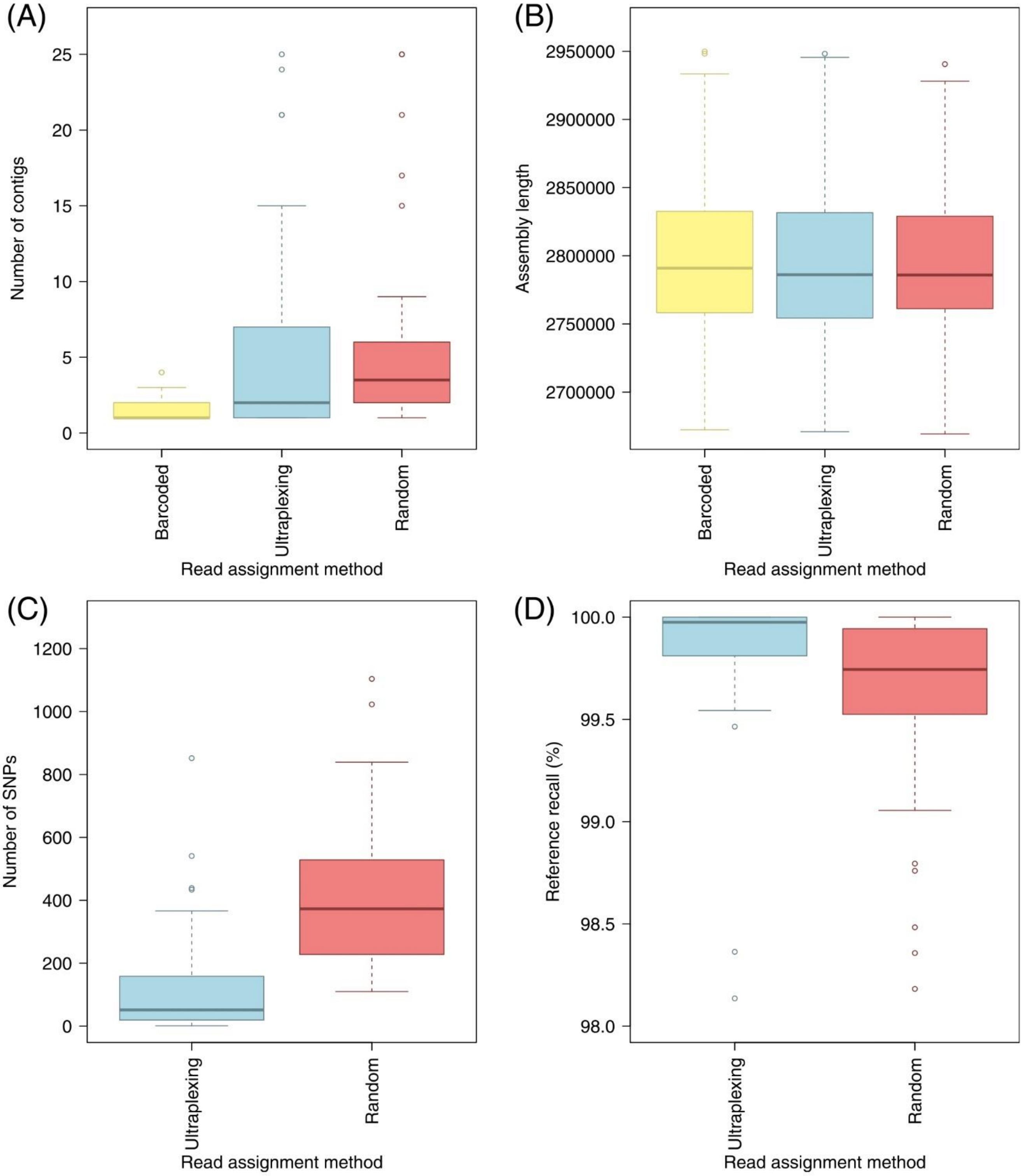


Fig. 4 Ultraplexing and classical molecular barcoding on a set of 48 *S. aureus* isolates. **a** The distribution of contigs per assembly. **b** The distribution of assembly lengths. **c** The distribution of SNPs per assembly. **d** The distribution of reference recall. SNPs and reference recall are calculated relative to assemblies based on molecular barcoding, and the same Illumina sequencing data were used throughout. Barcoded, molecularly barcoded Nanopore data, 5 flow cells with ≤ 10 samples each; Ultraplexing, reads assigned by the Ultraplexing algorithm, 1 flow cell with 48 samples; Random, reads from the Ultraplexing run, assigned randomly

factor driving read assignment accuracy was the extent of genetic variability between the pooled samples. Consistent with this, Ultraplexing achieved near-perfect read assignment in the first multi-species experiment but reduced assignment accuracy when species were represented by more than one strain. We hypothesized that mis-assignments driven by inter-sample sequence homology would have a negligible effect on assembly accuracy. Consistent with this, assembly accuracy was relatively insensitive to increasing numbers of mis-assigned reads in the single-species experiment, and we could confirm that inter-sample sequence homology accounts for the majority of mis-assigned reads using edit distance metrics. Furthermore, assembly accuracy was significantly reduced for random read assignment, reflecting higher proportions of falsely assigned reads in the absence of underlying sequence homologies. In addition, Ultraplexing may be less well suited for applications that depend on accurate assignments of individual reads, such as read-based methylation calling.

Our study has a number of limitations. First, we have only validated Ultraplexing on a single long-read technology, Oxford Nanopore. However, based on prior work demonstrating successful *k*-mer-based classification of eukaryotic PacBio reads [30, 31], we expect that Ultraplexing could also be applied to PacBio data, though the shorter subread distribution of the technology may negatively impact accuracy [32]. Second, although Ultraplexing was validated on a number of clinically important bacterial species covering a wide array of genome sizes and genome complexities, we cannot exclude the possibility that performance may degrade for genome or repeat configurations not included in the test set. Third, we have not rigorously tested the technical limits of Ultraplexing, including the maximum number of isolates and the necessary properties of the short-read sequencing data. Given that flow cell output has been increasing steadily, extraction of high molecular weight DNA for long-read sequencing may plausibly become the most significant limiting factor. Fourth, in terms of bioinformatics methods development, Ultraplexing relies on simple *k*-mer statistics instead of proper graph alignment [33–35], and we have not explored methods for the optimization of intra-batch genetic diversity in large sequencing projects. These points could be addressed in future work.

Conclusion

Ultraplexing is a new method for multiplexed long-read sequencing in the context of hybrid de novo assembly. Ultraplexing-based assemblies are highly accurate in terms of genome structure and consensus accuracy and exhibit quality characteristics comparable to assemblies based on molecular barcoding. Through increasing the number of samples per flow cell and simplified library preparation, Ultraplexing enables significant reductions of long-read sequencing costs and hands-on time. Thus, Ultraplexing enables the cost-effective complete resolution of large numbers of bacterial genomes.

Methods

The Ultraplexing read assignment algorithm

Let *n* denote the number of sequenced bacterial samples. We assume the availability of high-coverage Illumina sequencing data for each of the *n* individual isolates and that a pool of high molecular weight DNA, representing a mixture of the genomes of the *n* isolates, has been sequenced with a long-read sequencing technology like Oxford Nanopore or Pacific Biosciences. For each sample, a de Bruijn graph ($k = 19$) is constructed from the sample-specific Illumina short-read data and the graph is cleaned (removal of low-coverage supernodes) with Cortex [16]. Each long read from the pooled run is assigned to the sample for which the number of read *k*-mers present in the cleaned sample de Bruijn graph is maximal (or randomly in cases of a draw). We note that our approach can be understood as a heuristic approach to read-to-graph alignment. After the long-read assignment process is complete (i.e., after each long read has been assigned to one of the *n* isolates), the Cortex graph is discarded for the subsequent assembly steps. Of note, the choice of a *k* is a trade-off between the number of isolate-specific *k*-mers at a given *k* and the expected *k*-mer survival rate in the long-read data, calculated as $(1 - e)$^k, where *e* is the long-read sequencing error rate. $k = 19$ was chosen based on published work [25] on *k*-mer-based binning of long reads and based on preliminary simulation experiments.

Hybrid assembly and assembly evaluation criteria

Unicycler (version 0.4.4) [17] was used for all hybrid assembly experiments in this publication. Unicycler receives, for each sample, (I) the sample-specific Illumina reads and (II) the long reads assigned to the sample. Long reads are assigned according to the Ultraplexing long-read assignment algorithm, the molecular barcodes, or the underlying ground truth, depending on the evaluation scenario.

The performance of Ultraplexing was assessed (I) by assessing the proportion of reads assigned to the correct sample (in simulations), (II) by comparing the generated Ultraplexing-based hybrid de novo assemblies to reference genomes (downloaded from RefSeq for simulations and based on barcoding-based hybrid assembly for real data, see below), and (III) by comparing the accuracy of Ultraplexing-based assemblies to that of assemblies based on random (all experiments) or perfect (in simulations) assignment of long reads.

To assess the accuracy of an assembly, we compared the assembly to the corresponding reference genome. As baseline characteristics, we considered the total number of contigs and the combined assembly length. Furthermore, nucmer v3.1 [36] was used to generate an alignment between the assembly and the reference genome, globally filtering identified diagonals with "delta-filter -1." We used the filtered diagonals to compute three quality metrics: "SNPs," measuring consensus accuracy; "reference recall," the fraction of the reference covered by the assembly; and "assembly precision," the fraction of the assembly covered by the reference. When reported, QV scores are calculated as $round(-10 \times \log10(\frac{average\ \#\ SNPs\ per\ genome}{average\ reference\ genome\ size}))$ (Phred scale). Of note, assembly precision was close to 100% in all experiments, and we do not separately report on this metric.

For the simulation experiment with plasmids, we separately evaluated the sets of chromosomal and plasmid contigs for each assembly. We relied on RefSeq annotations for determining the status (chromosomal or plasmid) of each contig in the reference and assigned the status of each assembly contig according to the status of its highest-scoring nucmer hit in the reference.

Read assignment accuracy and edit distance

In simulated datasets, we calculated the proportion of correctly assigned long reads. A read was counted as correctly assigned if, and only if, it was assigned to the genome it was simulated from. For mis-assigned reads, we additionally defined a metric referred to as "Δedit distance," using edlib (version 1.2.6) [37]. Let d_1 be the ends-free edit distance between a read and the genome it was simulated from, and let d_2 be the edit distance between a read and the genome it was assigned to. Δedit distance is defined as d_1–d_2, divided by the length of the read. A negative value indicates a better alignment to the source genome than to the predicted genome. To assess the distributional properties of Δedit distance, the metric was calculated for random samples of 100 mis-assigned reads per method.

Simulation experiments

For the multi-species simulation experiments, chromosomal sequences of 10 clinically important species were downloaded from RefSeq [38]. For the single-species experiments without plasmids, chromosomal sequences of 181 complete *S. aureus* genomes were downloaded from RefSeq [38]. For the single-species simulation experiment with plasmids, 169 complete genomes were downloaded that contained between 2 and 11 annotated plasmids. The accessions of all downloaded genomes are listed in Additional file 1, and the selected genome subsets are listed in the corresponding results tables (Additional files 4 and 5).

For each genome, 300 Mb of short-read data was simulated with wgsim (version 0.3.1-r13) [39], using the parameters base error rate (-e 0.005), length of first read (-1 150), length of second read (-2 150), outer distance between the read ends (-d 278), standard deviation (-s 128), mutation rate (-r 0), and fraction of indels (-R 0). Long-read data were simulated with pbsim (version 1.0.3)[40], using the parameters prefix of the output (--prefix [prefix]), coverage (--depth 200), mean read length (--length-mean 8370), standard deviation of the read length (--length-sd 6389), maximum read length (--length-max 61011), minimum read length (--length-min 230), mean sequencing accuracy (--accuracy-mean 0.88), and model of quality code (--model_qc model_qc_clr). Mean read length was adjusted to match that of our first Nanopore sequencing run, and maximum read length was set to approximately 85% of that observed on the first run (Additional file 10). For all experiments, we assumed a constant long-read flow cell capacity of 5 Gb, and per-isolate coverage was adjusted accordingly (i.e., 5 Gb total output divided by the number of simulated isolates). Simulated long-read data were pooled and demultiplexed with the Ultraplexing algorithm. Hybrid de novo assembly was carried out, and the generated assemblies were benchmarked against the utilized reference genomes.

DNA extraction and long-read sequencing

DNA was extracted from overnight bacterial cultures in 3 ml LB broth. For short-read sequencing, the "DNeasy UltraClean Microbial" Kit was used according to the manufacturer's instruction. One nanogram of DNA per isolate was used for the library preparation with the TruePrep DNA Library Prep Kit. Short-read sequencing was conducted on a MiSeq instrument (Illumina) using 250 bp paired end sequencing using v2 chemistry. DNA extraction for long-read sequencing was performed with the MagAttract HMW DNA Kit (QIAGEN). Wide bore pipette tips were used to avoid shearing. Long-read sequencing was carried out on a MinION device with FLO-MIN106 flow cells and the SQK-LSK108 ligation sequencing kit (real-data experiment I) and SQK-LSK109 ligation sequencing kit (real-data experiment II). Of note, SQK-LSK109 involves reduced pipetting, possibly decreasing shearing. For barcoded long-read sequencing, samples were labeled with barcodes using the Oxford Nanopore ligation sequencing kit (EXP-NBD103 kit for 12 samples per run), and reads were demultiplexed with Albacore (version 2.1.3). For Ultraplexing, DNA from individual samples was pooled based on equal weight to yield a total of 700 ng of DNA, and demultiplexing was carried out with the Ultraplexing algorithm. Summary statistics of all sequencing runs are presented in Additional file 10.

Real-data validation experiments

For all experiments with real data, we used hybrid assembly with Unicycler [17] to generate high-quality reference genomes for all isolates, combining molecularly barcoded short- and long-read data.

Molecular long-read barcoding was carried out using the 12-sample barcoding kit (EXP-NBD103) for the first real-data experiment (1 flow cell) and for the second real-data experiment (5 flow cells with ≤ 10 samples per run). Barcoded Illumina sequencing runs were carried out for all samples in the real-data experiments. All sequencing runs are summarized in Additional file 10. Read mappability was determined with BWA MEM (version 0.7.17-r1188) (with standard settings and read mapping mode -x ont2d) [41].

Plasmid identification

To check if smaller contigs in barcoded assemblies of the real-data experiments represented plasmids, we used the online version of BLAST [28]. All non-chromosomal contigs (assumed to be all contigs but the longest in each assembly) were blasted against the nucleotide (nt) database, restricted to sequences that correspond to bacteria (taxid: 2), and if the best hit was characterized as plasmid and had a high identity (≥ 90%) and a low *e* value (0 or close to 0), we assumed that the contig represented a correctly assembled plasmid (Additional file 8). Three plasmids that generated hits to human BAC constructs were removed from the corresponding assemblies.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-01974-9.

Additional file 1. Sample summary. Names, accessions and summary statistics of all utilized reference genomes.

Additional file 2. Mash distances. Relatedness of genomes within each experiment.

Additional file 3. Main Evaluation Simulation Experiment I. Read classification and assembly accuracy in a simulation experiment with 10 different human pathogens.

Additional file 4. Evaluation 3 Additional Simulation Experiments. Read classification and assembly accuracy for 3 additional simulation experiments (5 species x 2 strains, class I and class III *S. aureus*, 10 *Pseudomonas*).

Additional file 5. Main Evaluation Simulation Experiment II. Read classification and assembly accuracy in a simulation experiment with 10 – 50 *S. aureus* genomes.

Additional file 6. Incorrectly assembled plasmids (simulations). Incorrectly assembled or incompletely recovered plasmids in the simulated sets with 10 – 50 plasmid-containing *S. aureus* isolates.

Additional file 7. Main Evaluation Simulation Experiment III. Read classification and assembly accuracy in a simulation experiment with 10 – 50 plasmid-containing *S. aureus* genomes.

Additional file 8. Putative plasmids (real data). BLAST results for contigs putatively representing plasmids in two real-data experiments.

Additional file 9. Evaluation of real-data experiments. Assembly accuracy and properties of the utilized reference genomes in two real-data experiments.

Additional file 10. Sequencing data summary. Summary statistics of all generated read sets (Oxford Nanopore and Illumina).

Additional file 11. Detailed legends for the supplementary tables.

Additional file 12. Supplementary figures.

Additional file 13. Review history.

Acknowledgements

We thank Lisanna Hülse for technical assistance regarding DNA extraction, library preparation, and Nanopore sequencing. We thank Harald Seifert (University of Cologne) for providing the bacterial isolates. Computational support and infrastructure were provided by the "Centre for Information and Media Technology" (ZIM) at the University of Düsseldorf (Germany). Illumina-based sequencing was performed by the Biologisch-Medizinisches Forschungszentrum der Heinrich-Heine-Universität Düsseldorf (BMFZ).

Peer review information

Andrew Cosgrove was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 13.

Authors' contributions

AD and AJK contributed to the study concept and design, data management, data analysis, data interpretation, and manuscript writing. SM contributed to the data management, data analysis and data interpretation, and manuscript writing. All authors have read and approved the final draft submitted.

Authors' information

Twitter handles: @AlexDilthey (Alexander Dilthey), @Bioinformeyer (Sebastian A. Meyer), @AchimKaasch (Achim J. Kaasch).

Funding

This work was supported by the Jürgen Manchot Foundation and the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

Availability of data and materials

The datasets generated and analyzed during the current study, as well as the generated reference assemblies, are available under the BioProject accession number PRJNA528186: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA528186 [42].
Assemblies from Ultraplexing-based and random assignment of reads and the source code of the Ultraplexer are available on OSF: https://doi.org/10.17605/OSF.IO/4M9VH [43].
The source code of the Ultraplexing algorithm is also available on GitHub: https://github.com/SebastianMeyer1989/UltraPlexer [44].
The Ultraplexing algorithm is made available under the MIT license and implemented in C++, Perl, and R. Sequence-to-graph alignment depends on the Cortex (cortex_var) package version 1.0.5.21 [16].

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Institute of Medical Microbiology and Hospital Hygiene, University Hospital, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany. [2]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, MD 20892, USA. [3]Institute of Medical Microbiology and Hospital Hygiene, University Hospital, Otto-von-Guericke-University Magdeburg, Magdeburg, Germany.

Received: 19 July 2019 Accepted: 24 February 2020
Published online: 14 March 2020

References

1. Falush D. Bacterial genomics: microbial GWAS coming of age. Nat Microbiol. 2016;1(5):16059.
2. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015;25:17–24.
3. Young BC, Earle SG, Soeng S, Sar P, Kumar V, Hor S, et al. Panton-Valentine leucocidin is the key determinant of Staphylococcus aureus pyomyositis in a bacterial GWAS. eLife. 2019;8:e42486.
4. Jagadeesan B, Gerner-Smidt P, Allard MW, Leuillet S, Winkler A, Xiao Y, et al. The use of next generation sequencing for improving food safety: translation into practice. Food Microbiol. 2019;79:96–115.
5. Cocolin L, Mataragas M, Bourdichon F, Doulgeraki A, Pilet M-F, Jagadeesan B, et al. Next generation microbiological risk assessment meta-omics: the next need for integration. Int J Food Microbiol. 2018;287:10–7.
6. Diaz-Sanchez S, Hanning I, Pendleton S, D'Souza D. Next-generation sequencing: the future of molecular genetics in poultry production and food safety. Poult Sci. 2013;92(2):562–72.
7. Taboada EN, Graham MR, Carriço JA, Van Domselaar G. Food safety in the age of next generation sequencing, bioinformatics, and open data access. Front Microbiol. 2017;8:909.
8. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 2011;39(13):e90.
9. Menzies BE. The role of fibronectin binding proteins in the pathogenesis of Staphylococcus aureus infections. Curr Opin Infect Dis. 2003;16(3):225–9.
10. Bartels MD, Petersen A, Worning P, Nielsen JB, Larner-Svensson H, Johansen HK, et al. Comparing whole-genome sequencing with sanger sequencing for spa typing of methicillin-resistant Staphylococcus aureus. J Clin Microbiol. 2014;52(12):4305–8.
11. Rhoads A, Au KF. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics. 2015;13(5):278–89.
12. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, et al. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif. 2015;3:1–8.
13. Krishnakumar R, Sinha A, Bird SW, Jayamohan H, Edwards HS, Schoeniger JS, et al. Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias. Sci Rep. 2018;8:3159.
14. Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, et al. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. Bioinformatics. 2014;30(19):2709–16.
15. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36(4):338–45.
16. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. 2012; 44(2):226–32.
17. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017; 13(6):e1005595.
18. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
19. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics. 2016; 32(7):1009–15.
20. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821–9.
21. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.
22. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):e112963.
23. Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, Mcvey SD, et al. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol. 2013;14(9):R101.
24. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb Genomics. 2017; 3(10):e000132.
25. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol. 2018;36(12):1174–82.
26. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17:132.
27. Schmid M, Frei D, Patrignani A, Schlapbach R, Frey JE, Remus-Emsermann MNP, et al. Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. Nucleic Acids Res. 2018;46(17):8953–65.
28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
29. Goldstein S, Beka L, Graf J, Klassen JL. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. BMC Genomics. 2019;20:23.
30. De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND, Swann J, et al. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microb Genomics. 2019;5(9):e000294.
31. Ip CLC, Loose M, Tyson JR, de Cesare M, Brown BL, Jain M, et al. MinION Analysis and Reference Consortium: phase 1 data release and analysis. F1000Research. 2015;4:1075.
32. Hestand MS, Van Houdt J, Cristofoli F, Vermeesch JR. Polymerase specific error rates and profiles identified by single molecule sequencing. Mutat Res. 2016;784–785:39–45.
33. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018;36(9):875–9.
34. Rautiainen M, Mäkinen V, Marschall T. Bit-parallel sequence-to-graph alignment. Bioinforma Oxf Engl. 2019;35(19):3599–607.
35. Jain C, Dilthey A, Misra S, Zhang H, Aluru S. Accelerating sequence alignment to graphs. bioRxiv. 2019;27:651638.
36. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.
37. Šošić M, Šikić M. Edlib: a C/C ++ library for fast, exact sequence alignment using edit distance. Bioinformatics. 2017;33(9):1394–5.
38. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35(suppl_1):D61–5.
39. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16): 2078–9.
40. Ono Y, Asai K, Hamada M. PBSIM: PacBio reads simulator—toward accurate genome assembly. Bioinformatics. 2013;29(1):119–21.
41. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio. 2013;arXiv:1303.3997.
42. Dilthey A, Meyer SA, Kaasch AJ. Ultraplexing validation: BioProject; 2019. Available from: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA528186. [cited 2020 Feb 19].
43. Dilthey A, Meyer SA, Kaasch AJ. Ultraplexing validation: OSF; 2020. Available from: https://osf.io/4m9vh/, https://doi.org/10.17605/OSF.IO/4M9VH. [cited 2020 Feb 19].
44. Dilthey A, Meyer SA, Kaasch AJ. UltraPlexer: GitHub; 2019. Available from: https://github.com/SebastianMeyer1989/UltraPlexer. [cited 2020 Feb 19].

Publisher's Note

3.2 Supplementary material to publication 1

Due to their size, the supplementary tables could not be inserted here. They can be accessed from the original publication source (https://doi.org/10.1186/s13059-020-01974-9), or via the enclosed CD instead. The supplementary figures mentioned in the publication are listed on the following pages.

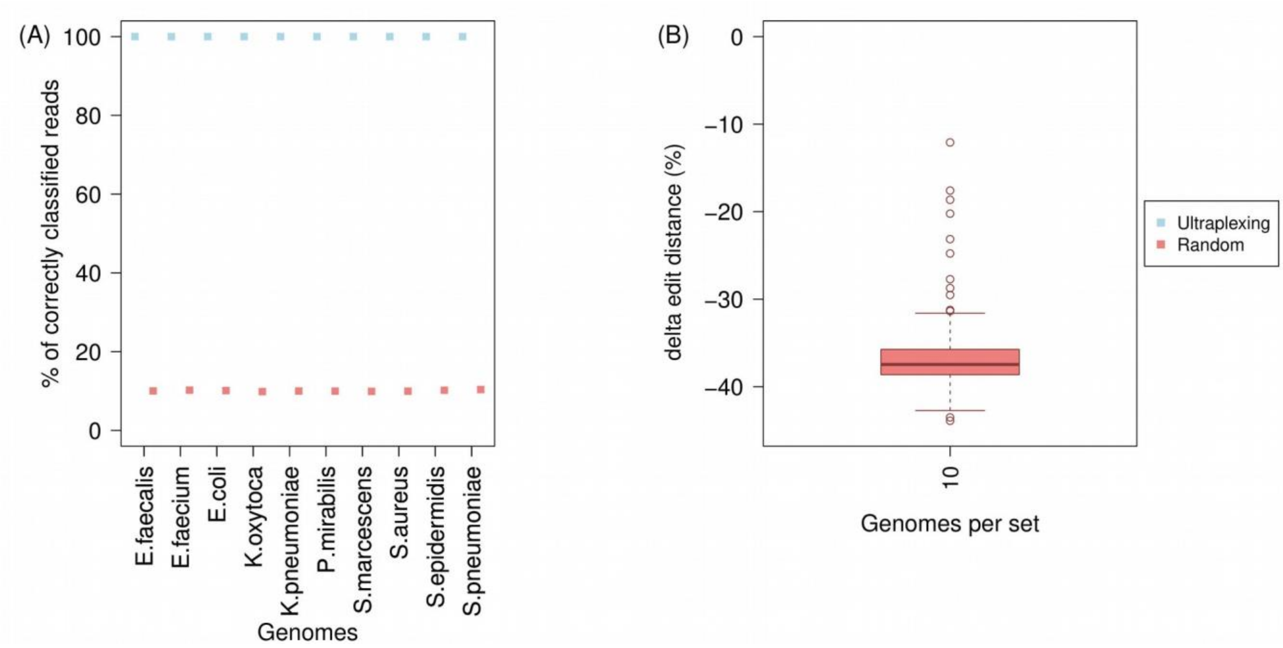


Figure S1: Read classification in a simulation experiment with ten different human pathogens. The figure shows the percentage of correctly classified simulated long reads (A) and Δedit distance for falsely classified reads (B). Reads were assigned according to the Ultraplexing algorithm (Ultraplexing) and randomly (Random).

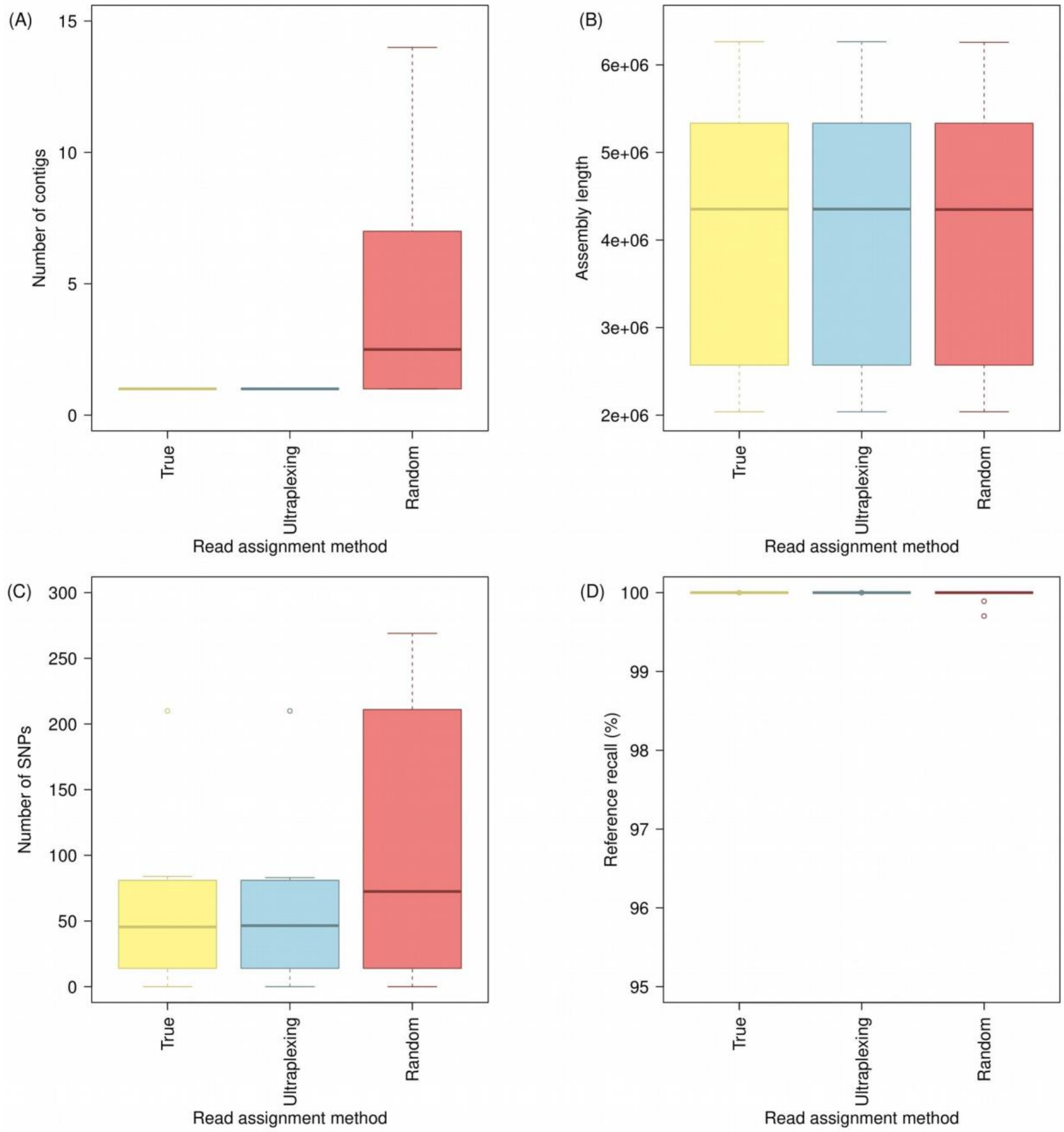


Figure S2: Assembly accuracy in a simulation experiment with ten different human pathogens. The figure shows the distribution of contigs per assembly (A); the distribution of assembly lengths (B); the distribution of SNPs per assembly (C); and thedistribution of reference recall (D). Long reads were assigned to their true origin (True); by the Ultraplexing algorithm (Ultraplexing); and randomly (Random).

Independent of long-read assignment method, the same simulated short-read data areused for all hybrid assemblies of the same species. SNPs and reference recall were calculated relative to the utilized reference genomes.

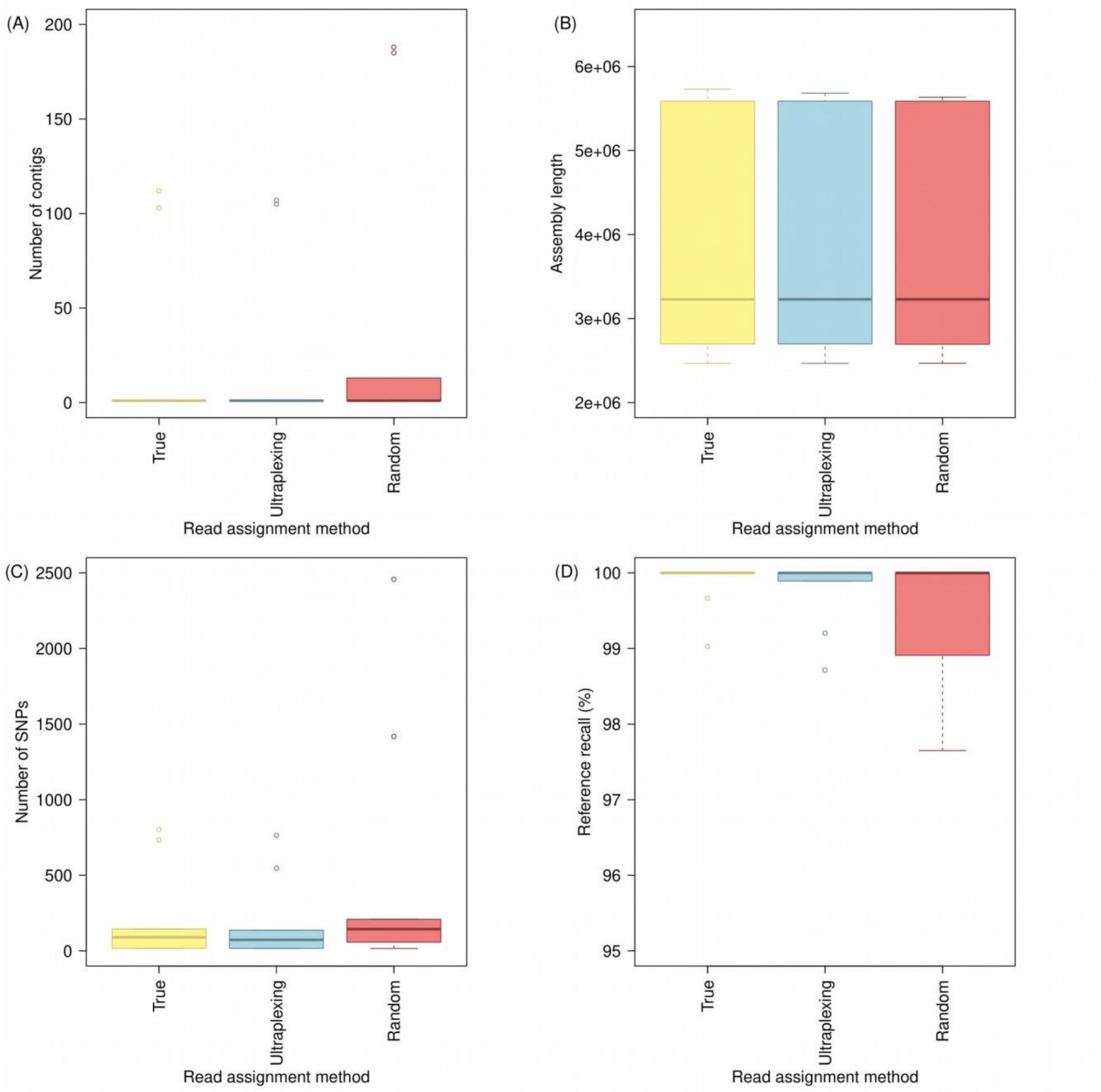


Figure S3: Assembly accuracy in a simulation experiment with five different human pathogens, each represented by two closely related strains. The figure shows the distribution of contigs per assembly (A); the distribution of assembly lengths (B); the distribution of SNPs per assembly (C); and the distribution of reference recall (D). Long reads were assigned to their true origin (True); by the Ultraplexing algorithm (Ultraplexing); and randomly (Random). Independent of long-read assignment method, the same simulated short-read data are used for all hybrid assemblies of the same species. SNPs and reference recall were calculated relative to the utilized reference genomes.

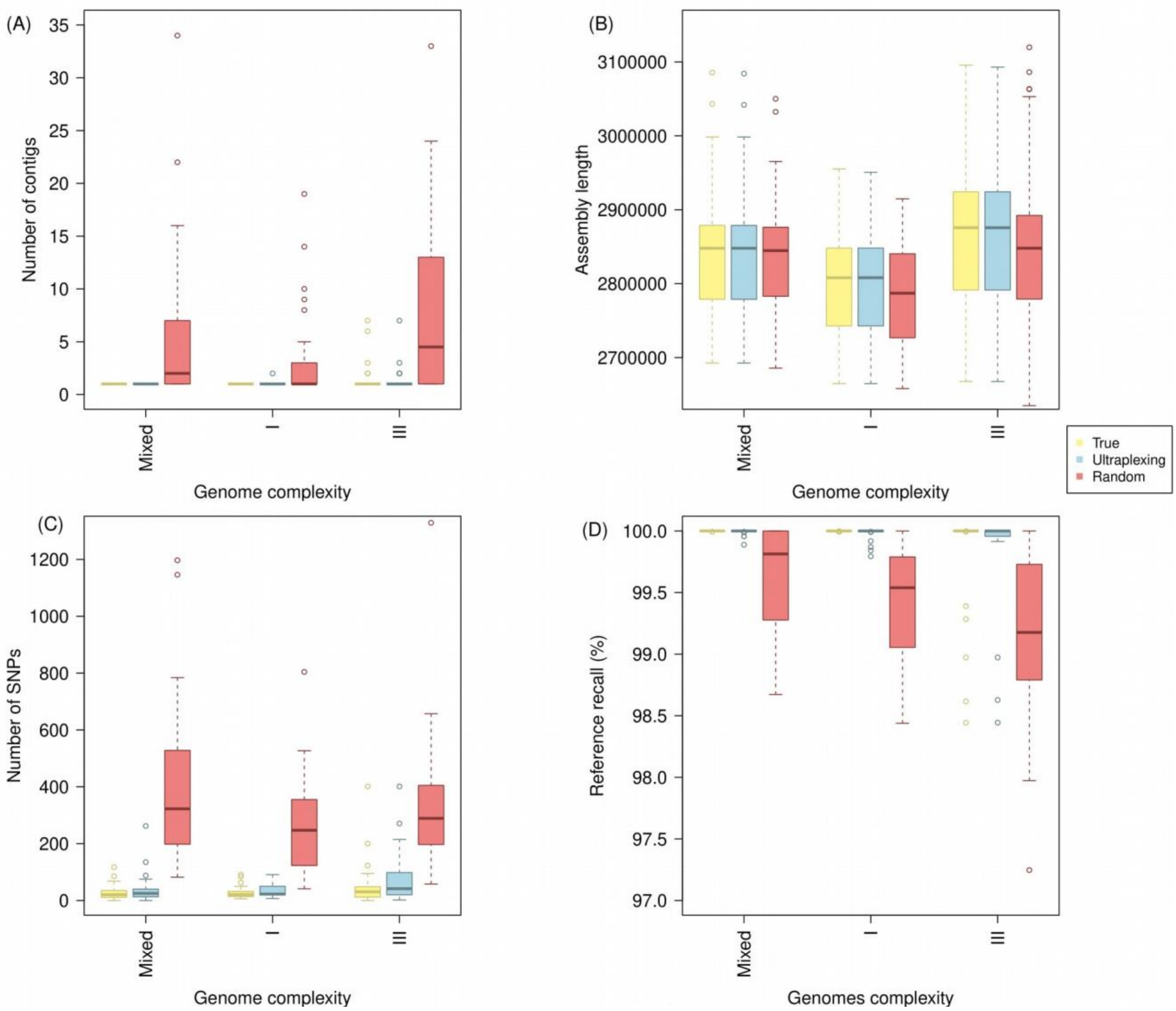


Figure S4: Assembly accuracy in three simulation experiments with 30 *S. aureus* genomes of different genome complexity each, based on 30 genomes with mixed complexity randomly drawn from the set used for the main part of Simulation experiment II (Mixed); 30 class I complexity (I) genomes; and 30 class III complexity (III) genomes. The figure shows the distribution of contigs per assembly (A); the distribution of assembly lengths (B); the distribution of SNPs per assembly (C); and the distribution of reference recall (D).

Long reads were assigned to their true origin (True); by the Ultraplexing algorithm (Ultraplexing); and randomly (Random). Independent of long-read assignment method, the same simulated short-read data areused for all hybrid assemblies of the same isolate. SNPs and reference recall were calculated relative to the utilized reference genomes.

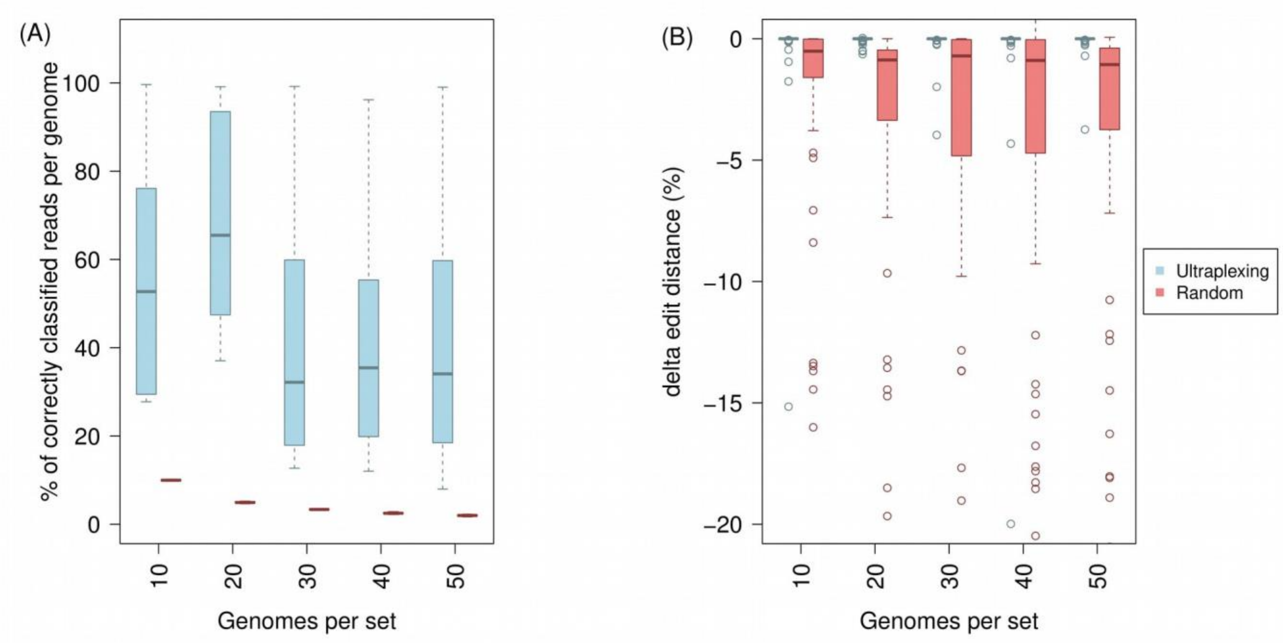


Figure S5: Read classification in five simulation experiments with 10 – 50 different plasmid-containing *S. aureus* genomes. The figure shows the distribution of the percentage of correctly classified simulated long reads (A) and the distribution of Δeditdistance for falsely classified reads (B). Reads were assigned by the Ultraplexing algorithm (Ultraplexing) and randomly (Random).

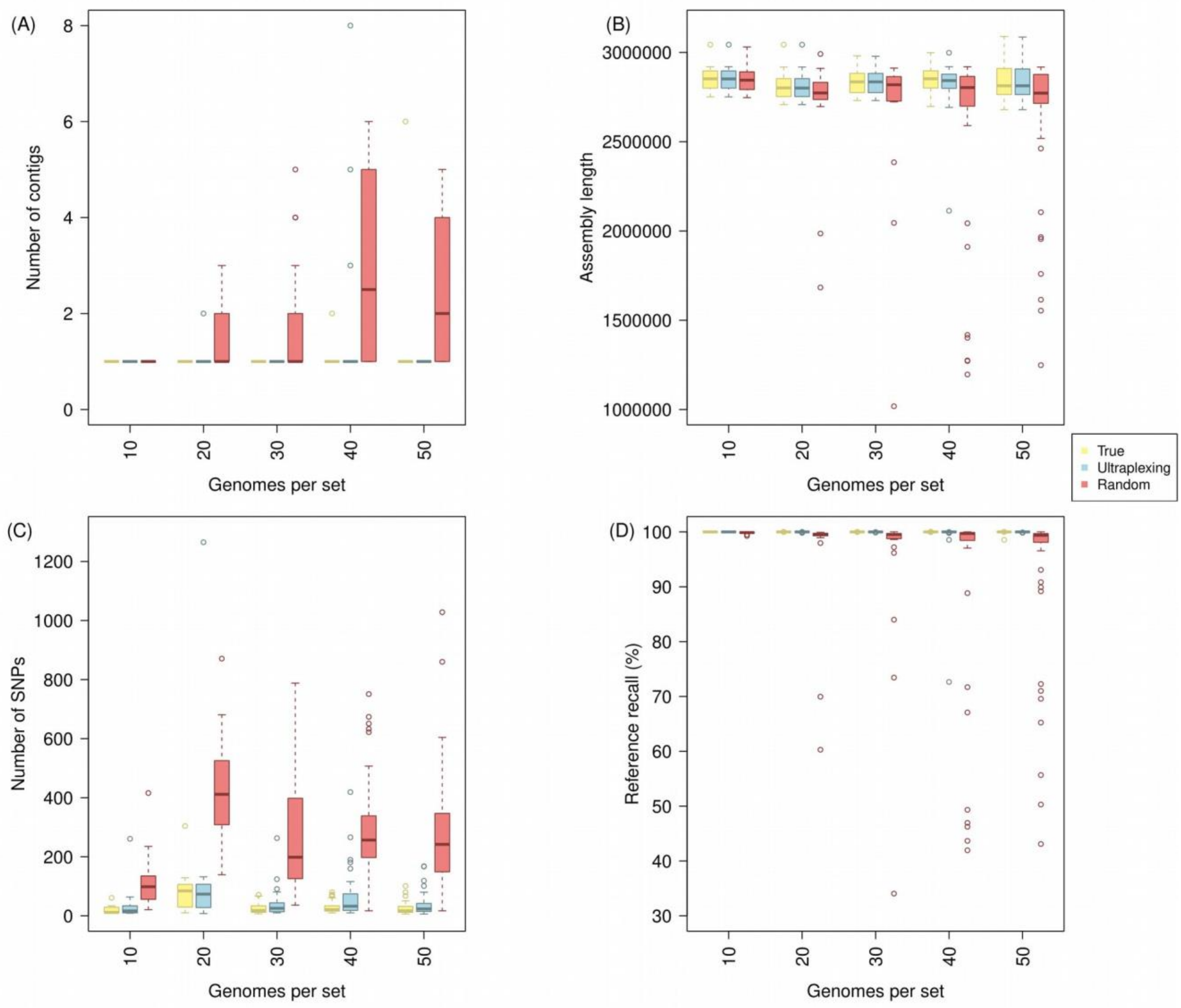


Figure S6: Chromosomal assembly accuracy in five simulation experiments with 10 – 50 different plasmid-containing *S. aureus* genomes. Reference and assembly contigs were classified as 'chromosomal' or 'plasmid' and evaluated separately (see Methods); shown here are results for the 'chromosomal' compartment. The figure shows the distribution of contigs per assembly (A); the distribution of assembly lengths (B); the distribution of SNPs per assembly (C); and the distribution of referencerecall (D).
Long reads were assigned to their true origin (True); by the Ultraplexing algorithm (Ultraplexing); and randomly (Random). Independent of long-read assignment method, the same simulated short-read data are used for all hybrid assemblies of the same isolate. SNPs and reference recall were calculated relative to the utilized reference genomes.

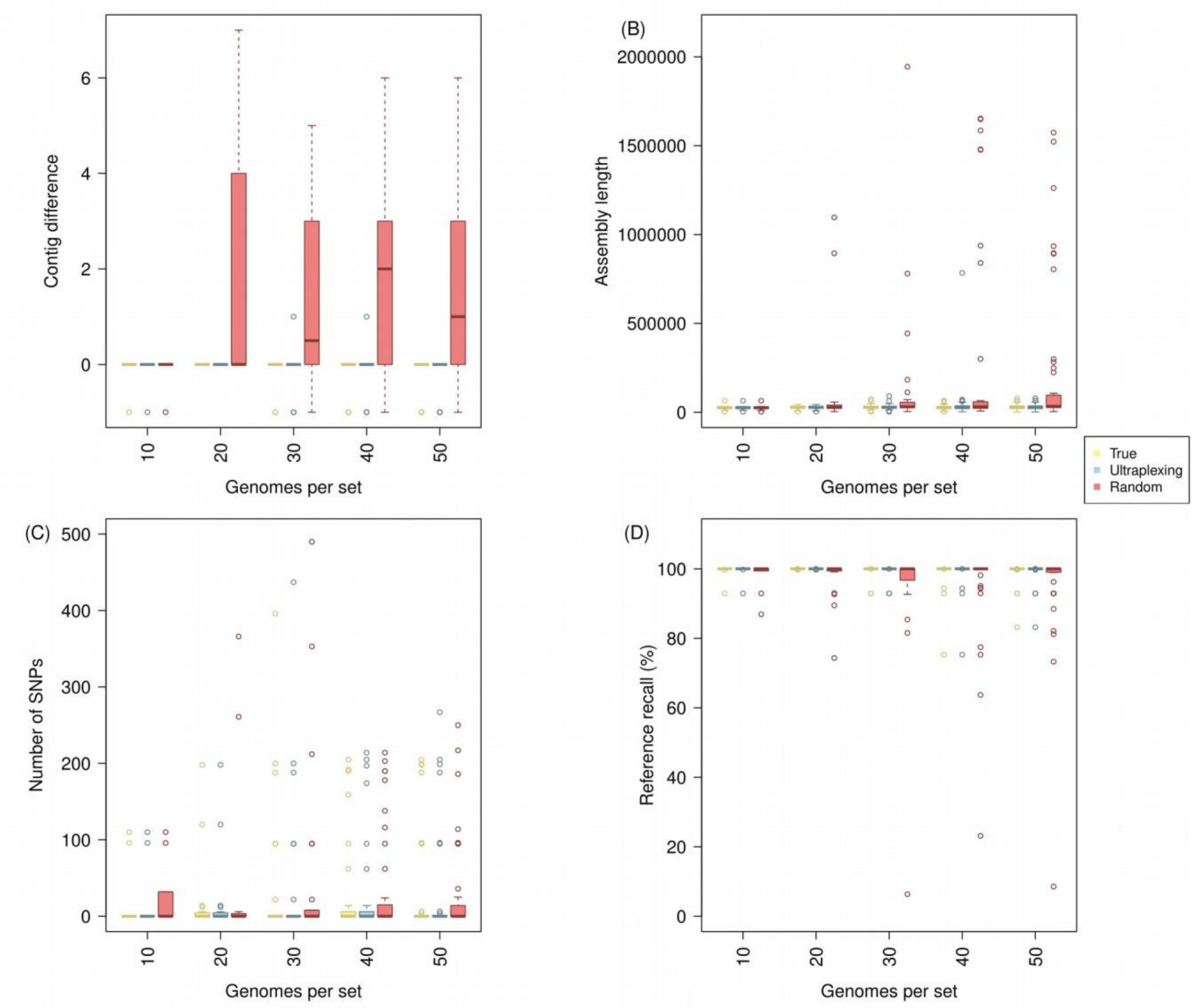


Figure S7: Plasmid assembly accuracy in five simulation experiments with 10 – 50 different plasmid-containing *S. aureus* genomes. Reference and assembly contigs were classified as 'chromosomal' or 'plasmid' and evaluated separately (see Methods); shown here are results for the 'plasmid' compartment. The figure shows thedistribution of contigs per assembly (A); the distribution of assembly lengths (B); the distribution of SNPs per assembly (C); and the distribution of reference recall (D).
Longreads were assigned to their true origin (True); by the Ultraplexing algorithm (Ultraplexing); and randomly (Random). Independent of long-read assignment method, the same simulated short-read data are used for all hybrid assemblies of the same isolate. SNPs and reference recall were calculated relative to the utilized reference genomes.

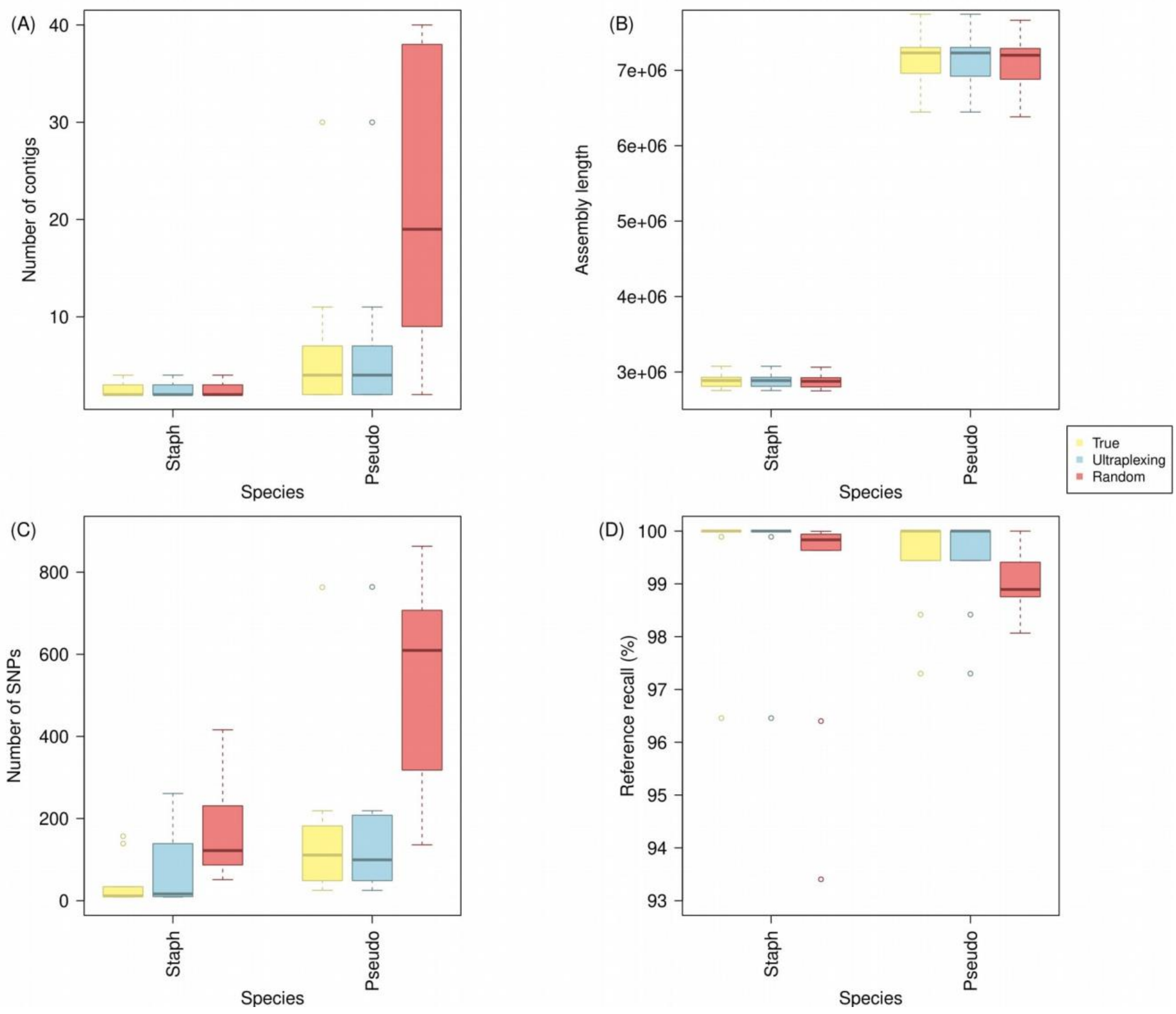


Figure S8: Assembly accuracy in two simulation experiments with 10 plasmid-containing genomes each, based on 10 *S. aureus* genomes randomly drawn from the set used for the main part of Simulation experiment II (Staph) and 10 Pseudomonas genomes with high repeat richness (Pseudo). The figure shows the distribution of contigs per assembly (A); the distribution of assembly lengths (B); the distribution of SNPs per assembly (C); and the distribution of reference recall (D).
Long reads were assigned to their true origin (True); by the Ultraplexing algorithm (Ultraplexing); and randomly (Random). Independent of long-read assignment method, the same simulated short-read data are used for all hybrid assemblies of the same isolate. SNPsand reference recall were calculated relative to the utilized reference genomes. Metrics for the *S. aureus* isolates were calculated for the chromosomal genome as described in the Methods section, metrics for the Pseudomonas isolates for the complete genome, not distinguishing between chromosomal and plasmid contigs.

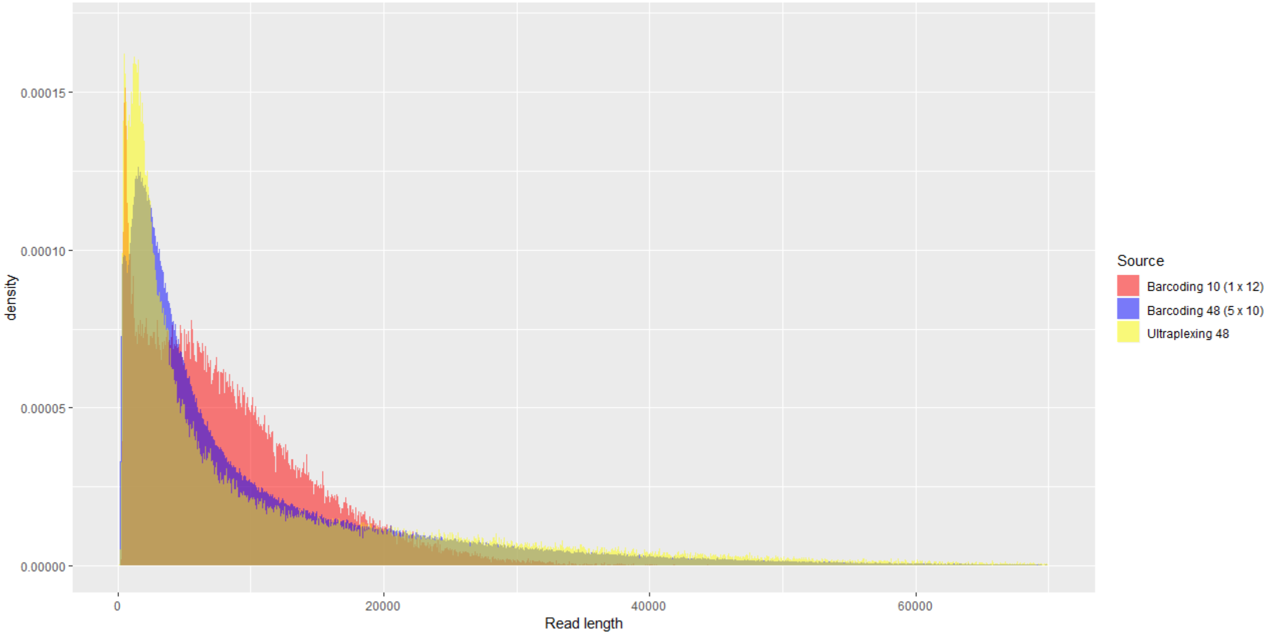


Figure S9: Read length distributions of the generated Oxford Nanopore datasets.

3.3 Publication 2: “pmrCAB Recombination Events among Colistin-Susceptible and -Resistant *Acinetobacter baumannii* Clinical Isolates Belonging to International Clone 7”

Manuscript published in: ASM Journals, mSphere

Impact factor: 5.029 (2021)

DOI: https://doi.org/10.1128/msphere.00746-21 (18)

Authorship: Shared first author

Contribution to the publication according to CRediT classification:

- Conceptualization (supporting)
- Data curation (equal)
- Formal analysis (lead)
- Investigation (equal)
- Methodology (equal)
- Project administration (supporting)
- Software (lead)
- Validation (equal)
- Visualization (equal)
- Writing - original draft preparation (supporting)
- Writing - review and editing (equal)

AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

RESEARCH ARTICLE

pmrCAB Recombination Events among Colistin-Susceptible and -Resistant *Acinetobacter baumannii* Clinical Isolates Belonging to International Clone 7

Carolina Silva Nodari,[a,b*] Sebastian Alexander Fuchs,[c] Kyriaki Xanthopoulou,[b,d] Rodrigo Cayô,[a,e] Harald Seifert,[b,d] Ana Cristina Gales,[a] Alexander Dilthey,[c] Paul G. Higgins[b,d]

[a]Universidade Federal de São Paulo-UNIFESP, Laboratório Alerta, Division of Infectious Diseases, Department of Internal Medicine, Escola Paulista de Medicina (EPM), São Paulo, Brazil

[b]Institute for Medical Microbiology, Immunology and Hygiene, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

[c]Institute of Medical Microbiology and Hospital Hygiene, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[d]German Center for Infection Research (DZIF), partner site Bonn-Cologne, Cologne, Germany

[e]Universidade Federal de São Paulo (UNIFESP), Laboratório de Bacteriologia e Imunologia (LIB), Setor de Biologia Molecular, Microbiologia e Imunologia, Departamento de Ciências Biológicas (DCB), Instituto de Ciências Ambientais, Químicas e Farmacêuticas (ICAQF), Diadema, Brazil

Carolina Silva Nodari and Sebastian Alexander Fuchs contributed equally to the study. The order of first authors was determined based on the initial date of enrollment in the study.

ABSTRACT *Acinetobacter baumannii* is a successful nosocomial pathogen due to its genomic plasticity. Homologous recombination allows genetic exchange and allelic variation among different clonal lineages and is one of the mechanisms associated with horizontal gene transfer (HGT) of resistance determinants. The main mechanism of colistin resistance in *A. baumannii* is mediated through mutations in the *pmrCAB* operon. Here, we describe two *A. baumannii* clinical isolates belonging to International Clone 7 (IC7) that have undergone recombination in the *pmrCAB* operon and evaluate the contribution of mobile genetic elements (MGE) to this phenomenon. Isolates 67569 and 72554 were colistin susceptible and resistant, respectively, and were submitted for short- and long-read genome sequencing using Illumina MiSeq and MinION platforms. Hybrid assemblies were built with Unicycler, and the assembled genomes were compared to reference genomes using NUCmer, Cortex, and SplitsTree. Genomes were annotated using Prokka, and MGEs were identified with ISfinder and repeat match. Both isolates presented a 21.5-kb recombining region encompassing *pmrCAB*. In isolate 67659, this region originated from IC5, while in isolate 72554 multiple recombination events might have happened, with the 5-kb recombining region encompassing *pmrCAB* associated with an isolate representing IC4. We could not identify MGEs involved in the mobilization of *pmrCAB* in these isolates. In summary, *A. baumannii* belonging to IC7 can present additional sequence divergence due to homologous recombination across clonal lineages. Such variation does not seem to be driven by antibiotic pressure but could contribute to HGT mediating colistin resistance.

IMPORTANCE Colistin resistance rates among *Acinetobacter baumannii* clinical isolates have increased over the last 20 years. Despite reports of the spread of plasmid-mediated colistin resistance among *Enterobacterales*, the presence of *mcr*-type genes in *Acinetobacter* spp. remains rare, and reduced colistin susceptibility is mainly associated with the acquisition of nonsynonymous mutations in *pmrCAB*. We have recently demonstrated that distinct *pmrCAB* sequences are associated with different *A. baumannii* International Clones (IC). In this study, we identified the presence of homologous recombination as an additional cause of genetic variation in this operon, which, to the best of our knowledge, was not mediated by mobile genetic elements. Even though this phenomenon was observed in both colistin-susceptible and -resistant

Editor Patricia A. Bradford, Antimicrobial Development Specialists, LLC

Address correspondence to Paul G. Higgins, paul.higgins@uni-koeln.de.

*Present address: Carolina Silva Nodari, Enteric Bacterial Pathogens Unit, Institut Pasteur, Paris, France.

The authors declare a conflict of interest.

Received 3 September 2021
Accepted 9 November 2021
Published 1 December 2021

isolates, it has the potential to contribute to the spread of resistance-conferring alleles, leading to reduced susceptibility to this last-resort antimicrobial agent.

KEYWORDS polymyxins, colistin resistance, mobile genetic elements, insertion sequences, Gram-negative bacilli

Acinetobacter baumannii is an opportunistic pathogen causing a variety of difficult-to-treat infections owing to their high incidence of antimicrobial resistance. One of the reasons for this is its high genomic plasticity and its ability to acquire resistance determinants (1, 2). The *A. baumannii* population can be grouped into nine international clonal lineages (3), which differ from each other in at least 1,800 alleles, as shown by core genome multilocus sequence typing (cgMLST) (4). Furthermore, each lineage has distinct alleles associated with them, such as the intrinsic $bla_{\text{OXA-51}}$-like (5).

Homologous recombination allows foreign DNA to be integrated into the chromosome, and in *A. baumannii* it has already been associated with the acquisition of resistance determinants to aminoglycosides (6, 7). Additionally, other studies have shown that homologous recombination contributes to the allelic variation of intrinsic resistance determinants, such as the outer membrane protein CarO (8) and the chromosome-encoded *Acinetobacter*-derived cephalosporinase (ADC) (9).

Mutations in the *pmrCAB* operon are the main mechanism causing reduced susceptibility to colistin among *A. baumannii* strains (10). We have recently demonstrated the allelic variation of *pmrCAB* between distinct International Clones (ICs) and that colistin-susceptible isolates belonging to the same clonal lineage should be used as reference strains when investigating point mutations potentially associated with colistin resistance (11, 12). Interestingly, some of the IC2 isolates described in the study by Gerson and colleagues (11) presented *pmrCAB* sequences that are associated with IC4, suggesting homologous recombination between these clonal lineages. Kim and Ko (13) have also suggested that *pmrCAB* genetic variation between distinct species belonging to the *A. baumannii*-*A. calcoaceticus* complex was due to recombination.

Here, we describe two *A. baumannii* clinical isolates belonging to IC7 with distinct colistin susceptibility profiles and presenting recombined *pmrCAB* operons and evaluate the contribution of mobile genetic elements (MGE) to this phenomenon.

(This work was presented in part at the 12th International Symposium on the Biology of *Acinetobacter* in Frankfurt, Germany, 2019)

RESULTS AND DISCUSSION

Some divergence was observed when the PmrCAB protein sequences of the IC7 isolates 67659 and 72554 were aligned against MC1 (IC7 reference genome). The colistin-susceptible isolate 67659 showed one amino acid substitution in both PmrA and PmrB as well as five in PmrC. In contrast, isolate 72554 presented 4, 18, and 71 amino acid substitutions in PmrA, PmrB, and PmrC, respectively (Fig. 1A to C). The k-mer sharing analysis of *pmrCAB* and its flanking regions demonstrated that sequence similarities were increased when isolates 67659 and 72554 were compared to those belonging to IC5 and IC4, respectively (Fig. 2). Furthermore, no amino acid substitutions were observed in PmrC or PmrA when isolates 67659 and 72554 were compared against isolate 67098 (IC5) and isolate 71813 (IC4), respectively. Higher sequence similarity was also observed in PmrB, with only a single substitution ($\text{Arg}_{389}\text{Gln}$) identified when isolates 71813 and 72554 were compared, as well as two substitutions ($\text{Pro}_{187}\text{Thr}$ and $\text{Asn}_{256}\text{Ile}$) in the comparison between isolates 67098 and 67659 (Fig. 1A to C). The representativeness of the included reference genomes was also explored in an additional set of isolates as well as in a larger genomic region (see Fig. S1 to S5 in the supplemental material).

The presence of regions with such high polymorphism rates suggests that horizontal transfer through recombination, rather than the accumulation of multiple point mutations over time, is involved in the variability of these specific DNA fragments. This is particularly important and more frequent in naturally transformable species, such as

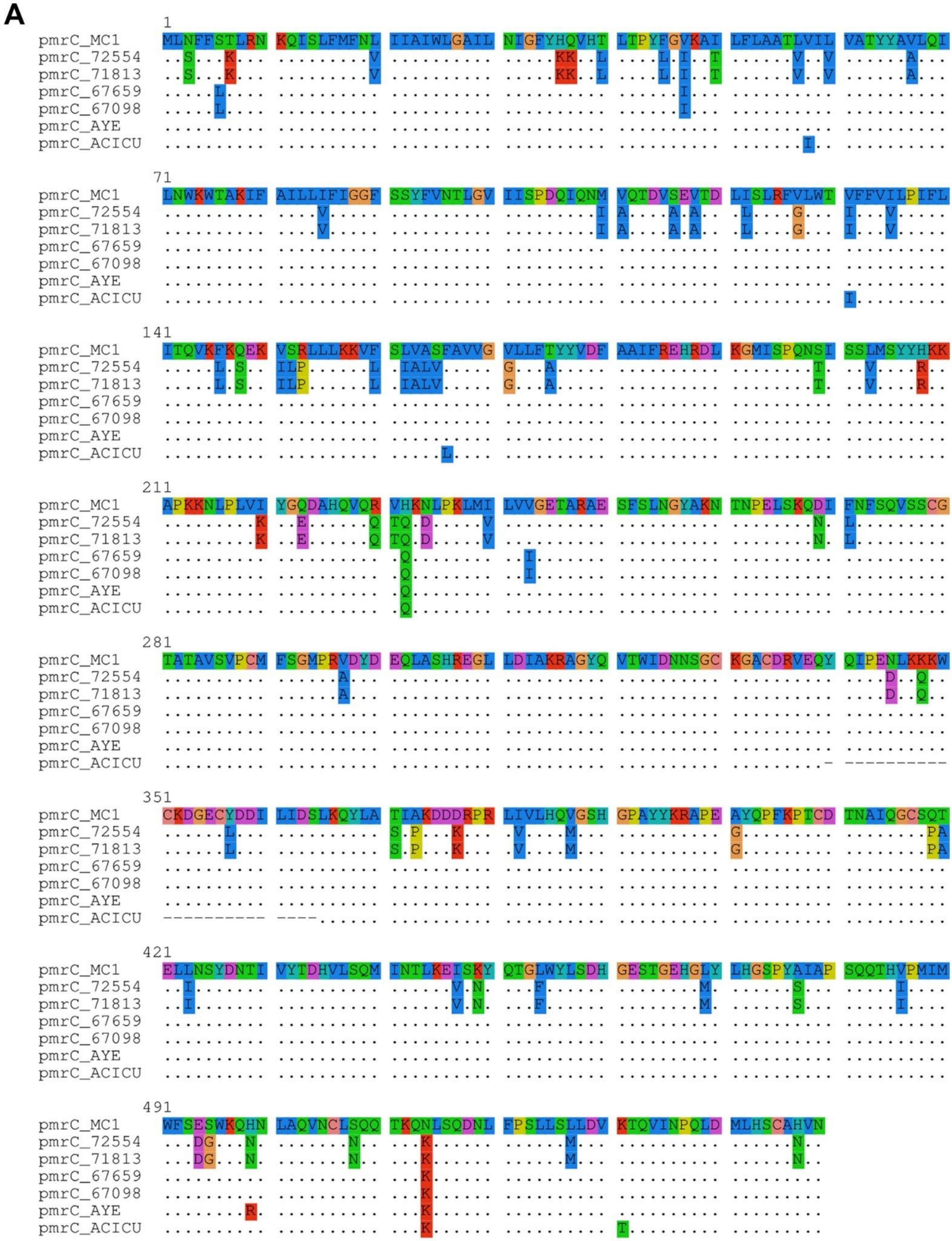

A
1
pmrC_MC1 MLNFFSTLRN KQISLFMFNL IIAIWLGAIL NIGFYHQVHT LTPYFGVKAI LFLAATLVIL VATYYAVLQI
pmrC_72554 ..S...K... ....V ....KK..L ....L.I..T ......V..V ......A...
pmrC_71813 ..S...K... ....V ....KK..L ....L.I..T ......V..V ......A...
pmrC_67659 .....L.... ....I... ....
pmrC_67098 .....L.... ....I... ....
pmrC_AYE ....
pmrC_ACICU ....I.. ....
71
pmrC_MC1 LNWKWTAKIF AILLIFIGGF SSYFVNTLGV IISPDQIQNM VQTDVSEVTD LISLRFVLWT VFFVILPIFL
pmrC_72554 ....V..... ....I A....A.A.. .L....G... I...V.....
pmrC_71813 ....V..... ....I A....A.A.. .L....G... I...V.....
pmrC_67659 ....
pmrC_67098 ....
pmrC_AYE ....
pmrC_ACICU .... I.........
141
pmrC_MC1 ITQVKFKQEK VSRLLLKKVF SLVASFAVVG VLLFTYYVDF AAIFREHRDL KGMISPQNSI SSLMSYYHKK
pmrC_72554 .....L.S.. ILP......L .IALV..... G...A..... ....T. ..V....R..
pmrC_71813 .....L.S.. ILP......L .IALV..... G...A..... ....T. ..V....R..
pmrC_67659 ....
pmrC_67098 ....
pmrC_AYE ....
pmrC_ACICU ....L.... ....
211
pmrC_MC1 APKKNLPLVI YGQDAHQVQR VHKNLPKLMI LVVGETARAE SFSLNGYAKN TNPELSKQDI FNFSQVSSCG
pmrC_72554 ....K ..E......Q TQ.D.....V ....N. L.........
pmrC_71813 ....K ..E......Q TQ.D.....V ....N. L.........
pmrC_67659 .... .Q........ ..I....... ....
pmrC_67098 .... .Q........ ..I....... ....
pmrC_AYE .... .Q........ ....
pmrC_ACICU .... .Q........ ....
281
pmrC_MC1 TATAVSVPCM FSGMPRVDYD EQLASHREGL LDIAKRAGYQ VTWIDNNSGC KGACDRVEQY QIPENLKKKW
pmrC_72554 ....A... ....D..Q..
pmrC_71813 ....A... ....D..Q..
pmrC_67659 ....
pmrC_67098 ....
pmrC_AYE ....
pmrC_ACICU ....- ----------
351
pmrC_MC1 CKDGECYDDI LIDSLKQYLA TIAKDDDRPR LIVLHQVGSH GPAYYKRAPE AYQPFKPTCD TNAIQGCSQT
pmrC_72554 ......L... .... S.P...K... .V....M... .... G......... ....PA
pmrC_71813 ......L... .... S.P...K... .V....M... .... G......... ....PA
pmrC_67659 ....
pmrC_67098 ....
pmrC_AYE ....
pmrC_ACICU ---------- ----...... ....
421
pmrC_MC1 ELLNSYDNTI VYTDHVLSQM INTLKEISKY QTGLWYLSDH GESTGEHGLY LHGSPYAIAP SQQTHVPMIM
pmrC_72554 ..I....... ....V.N. ...F...... ....M. ......S... .....I....
pmrC_71813 ..I....... ....V.N. ...F...... ....M. ......S... .....I....
pmrC_67659 ....
pmrC_67098 ....
pmrC_AYE ....
pmrC_ACICU ....
491
pmrC_MC1 WFSESWKQHN LAQVNCLSQQ TKQNLSQDNL FPSLLSLLDV KTQVINPQLD MLHSCAHVN
pmrC_72554 ...DG...N. ....N.. ...K...... ......M... ....N..
pmrC_71813 ...DG...N. ....N.. ...K...... ......M... ....N..
pmrC_67659 ....K...... ....
pmrC_67098 ....K...... ....
pmrC_AYE ....R. ....K...... ....
pmrC_ACICU ....K...... .... T......... ....

FIG 1 (A to D) Protein sequence alignment of PmrC (A), PmrA (B), and PmrB (C) and SplitsTree-based neighbor-net of a 23.6-kb genomic region encompassing *pmrCAB* (D) between isolates MC1 (IC7), 72554 (IC7), 71813 (IC4), 67659 (IC7), 67098 (IC5), AYE (IC1), and ACICU (IC2). Sequences belonging to isolate MC1 were used as references for sequence alignment. Amino acid differences are highlighted in colors (panels A to C).

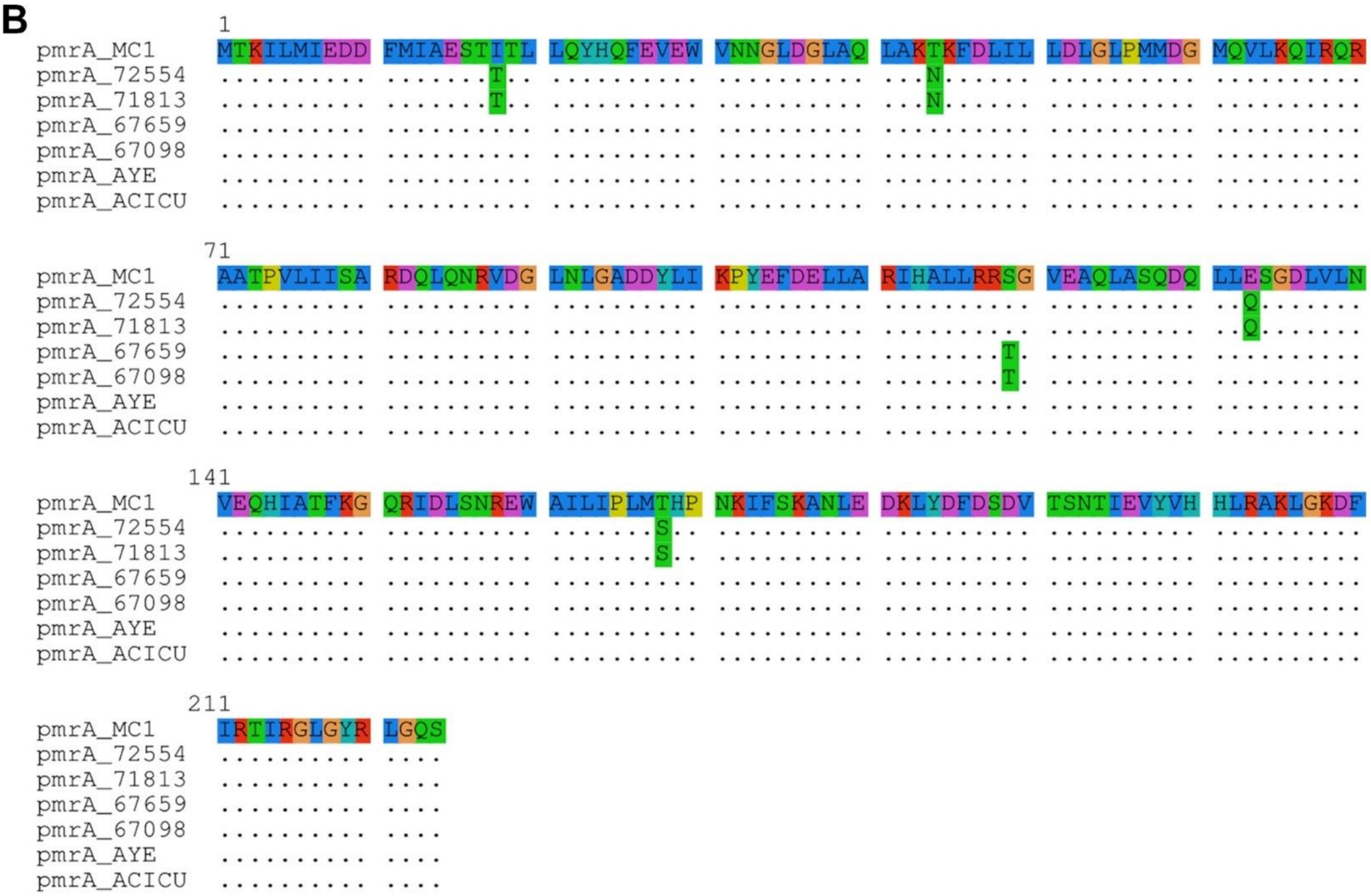
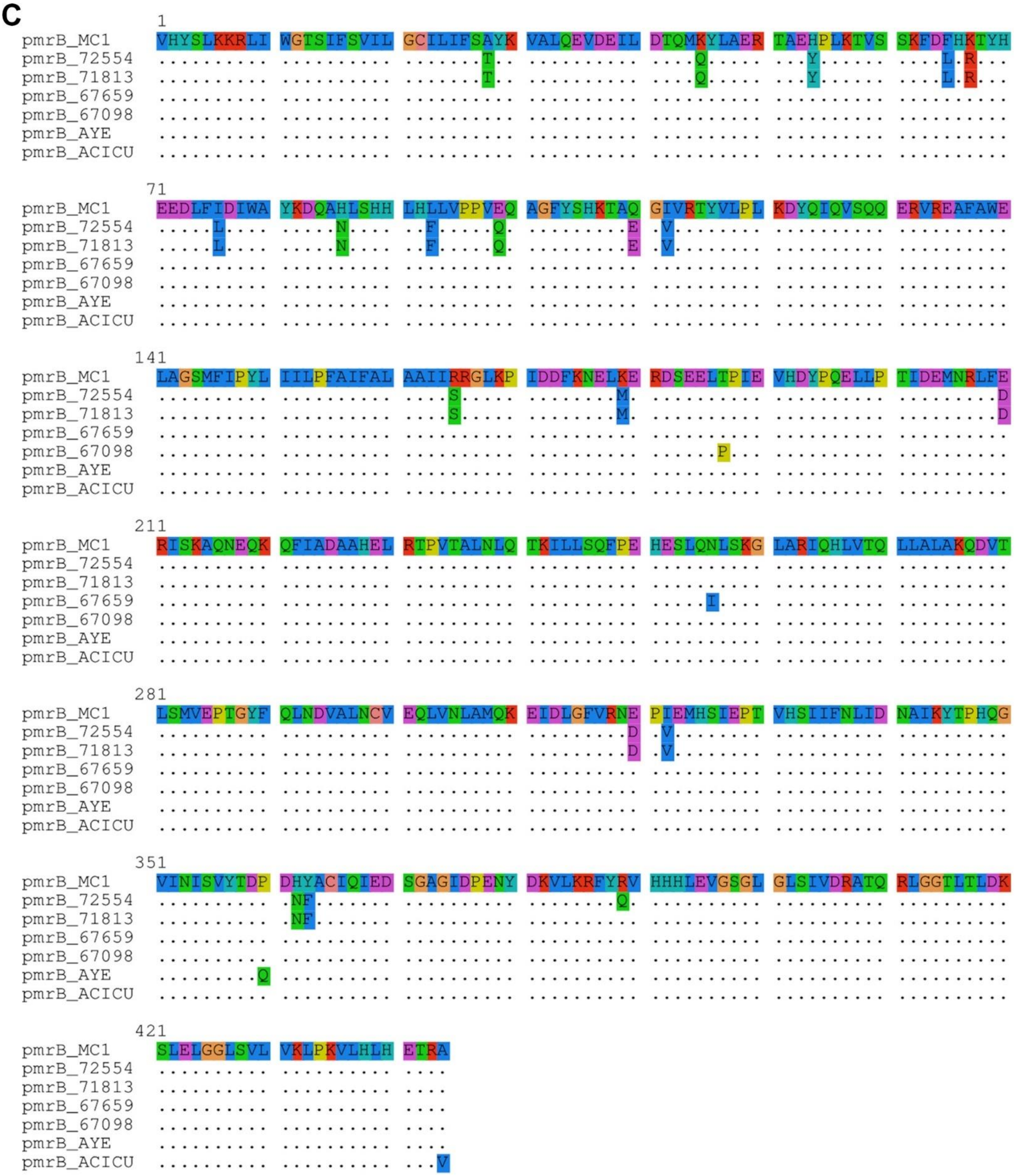

B
1
pmrA_MC1 MTKILMIEDD FMIAESTITL LQYHQFEVEW VNNGLDGLAQ LAKTKFDLIL LDLGLPMMDG MQVLKQIRQR
pmrA_72554 ....T.. ....N...... ....
pmrA_71813 ....T.. ....N...... ....
pmrA_67659 ....
pmrA_67098 ....
pmrA_AYE ....
pmrA_ACICU ....
71
pmrA_MC1 AATPVLIISA RDQLQNRVDG LNLGADDYLI KPYEFDELLA RIHALLRRSG VEAQLASQDQ LLESGDLVLN
pmrA_72554 ....Q.......
pmrA_71813 ....Q.......
pmrA_67659 ....T. ....
pmrA_67098 ....T. ....
pmrA_AYE ....
pmrA_ACICU ....
141
pmrA_MC1 VEQHIATFKG QRIDLSNREW AILIPLMTHP NKIFSKANLE DKLYDFDSDV TSNTIEVYVH HLRAKLGKDF
pmrA_72554 ....S.. ....
pmrA_71813 ....S.. ....
pmrA_67659 ....
pmrA_67098 ....
pmrA_AYE ....
pmrA_ACICU ....
211
pmrA_MC1 IRTIRGLGYR LGQS
pmrA_72554 ....
pmrA_71813 ....
pmrA_67659 ....
pmrA_67098 ....
pmrA_AYE ....
pmrA_ACICU ....

FIG 1 (Continued)

A. baumannii (1, 2). Based on the large number of nonsynonymous mutations observed in *pmrCAB*, with PmrC protein sequences presenting up to 13% divergence from what is expected for their lineage, we can infer that this operon has been transferred across clonal lineages through homologous recombination. The likely presence of recombination around the *pmrCAB* operon was confirmed by a SplitsTree analysis, also including reference genomes for IC1 and IC2 (Fig. 1D; phi test for recombination, $P = 0.0$). Considering that IC4 and IC5, together with IC7, are the most frequent lineages observed in South America (3) and were already described in the same hospital (12, 14), it comes as no surprise that horizontal gene transfer occurred among those lineages.

Using a k-mer-based analysis, it was noticed that the length of the region presenting high sequence divergence surrounding *pmrCAB* was similar between the two evaluated isolates and extended to at least 8 kb up- and downstream of *pmrCAB* (Fig. 2A and B, top). However, when using the same approach to compare those isolates to the reference genomes belonging to IC4 and IC5, which presumably acted as donors of the recombining regions, some differences were observed. While k-mer sharing proportion between isolates 67659 and 67098 was close to 1 through the whole extension of the recombining region (Fig. 2A, bottom), the similarities between isolates 72554 and 71813 were restricted to only 700 bp upstream of *pmrC* as well as 1,000 bp downstream of *pmrB* (Fig. 2B, bottom). This finding suggests that additional recombination events have taken place and that the *pmrCAB* allele belonging to IC4 went through some other intermediary host before making it into 72554, consistent with SplitsTree results. Boinett and colleagues (15) have previously suggested that a 700-kb genomic region that included *pmrCAB* had undergone homologous recombination in laboratory-induced colistin-resistant isolates. Those isolates, however, belonged to IC2, suggesting that recombining regions vary depending on their genetic background. This



```
              1
pmrB_MC1      VHYSLKKRLI WGTSIFSVIL GCILIFSAYK VALQEVDEIL DTQMKYLAER TAEHPLKTVS SKFDFHKTYH
pmrB_72554    .......... .......... .......T.. .......... ....Q..... ...Y...... ....L.R...
pmrB_71813    .......... .......... .......T.. .......... ....Q..... ...Y...... ....L.R...
pmrB_67659    .......... .......... .......... .......... .......... .......... ..........
pmrB_67098    .......... .......... .......... .......... .......... .......... ..........
pmrB_AYE      .......... .......... .......... .......... .......... .......... ..........
pmrB_ACICU    .......... .......... .......... .......... .......... .......... ..........

             71
pmrB_MC1      EEDLFIDIWA YKDQAHLSHH LHLLVPPVEQ AGFYSHKTAQ GIVRTYVLPL KDYQIQVSQQ ERVREAFAWE
pmrB_72554    .....L.... .....N.... ..F.....Q. .........E .V........ .......... ..........
pmrB_71813    .....L.... .....N.... ..F.....Q. .........E .V........ .......... ..........
pmrB_67659    .......... .......... .......... .......... .......... .......... ..........
pmrB_67098    .......... .......... .......... .......... .......... .......... ..........
pmrB_AYE      .......... .......... .......... .......... .......... .......... ..........
pmrB_ACICU    .......... .......... .......... .......... .......... .......... ..........

            141
pmrB_MC1      LAGSMFIPYL IILPFAIFAL AAIIRRGLKP IDDFKNELKE RDSEELTPIE VHDYPQELLP TIDEMNRLFE
pmrB_72554    .......... .......... ....S..... ........M. .......... .......... .........D
pmrB_71813    .......... .......... ....S..... ........M. .......... .......... .........D
pmrB_67659    .......... .......... .......... .......... .......... .......... ..........
pmrB_67098    .......... .......... .......... .......... ......P... .......... ..........
pmrB_AYE      .......... .......... .......... .......... .......... .......... ..........
pmrB_ACICU    .......... .......... .......... .......... .......... .......... ..........

            211
pmrB_MC1      RISKAQNEQK QFIADAAHEL RTPVTALNLQ TKILLSQFPE HESLQNLSKG LARIQHLVTQ LLALAKQDVT
pmrB_72554    .......... .......... .......... .......... .......... .......... ..........
pmrB_71813    .......... .......... .......... .......... .......... .......... ..........
pmrB_67659    .......... .......... .......... .......... .....I.... .......... ..........
pmrB_67098    .......... .......... .......... .......... .......... .......... ..........
pmrB_AYE      .......... .......... .......... .......... .......... .......... ..........
pmrB_ACICU    .......... .......... .......... .......... .......... .......... ..........

            281
pmrB_MC1      LSMVEPTGYF QLNDVALNCV EQLVNLAMQK EIDLGFVRNE PIEMHSIEPT VHSIIFNLID NAIKYTPHQG
pmrB_72554    .......... .......... .......... .........D .V........ .......... ..........
pmrB_71813    .......... .......... .......... .........D .V........ .......... ..........
pmrB_67659    .......... .......... .......... .......... .......... .......... ..........
pmrB_67098    .......... .......... .......... .......... .......... .......... ..........
pmrB_AYE      .......... .......... .......... .......... .......... .......... ..........
pmrB_ACICU    .......... .......... .......... .......... .......... .......... ..........

            351
pmrB_MC1      VINISVYTDP DHYACIQIED SGAGIDPENY DKVLKRFYRV HHHLEVGSGL GLSIVDRATQ RLGGTLTLDK
pmrB_72554    .......... .NF....... .......... ........Q. .......... .......... ..........
pmrB_71813    .......... .NF....... .......... .......... .......... .......... ..........
pmrB_67659    .......... .......... .......... .......... .......... .......... ..........
pmrB_67098    .......... .......... .......... .......... .......... .......... ..........
pmrB_AYE      .........Q .......... .......... .......... .......... .......... ..........
pmrB_ACICU    .......... .......... .......... .......... .......... .......... ..........

            421
pmrB_MC1      SLELGGLSVL VKLPKVLHLH ETRA
pmrB_72554    .......... .......... ....
pmrB_71813    .......... .......... ....
pmrB_67659    .......... .......... ....
pmrB_67098    .......... .......... ....
pmrB_AYE      .......... .......... ....
pmrB_ACICU    .......... .......... ...V
```

FIG 1 (Continued)

observation would be in agreement to the phenomenon described by Kim and Ko (13), where the authors reported that recombination could happen within *pmrC*, generating mosaic alleles. Such variation, however, was not observed in either of the two isolates evaluated in this study.

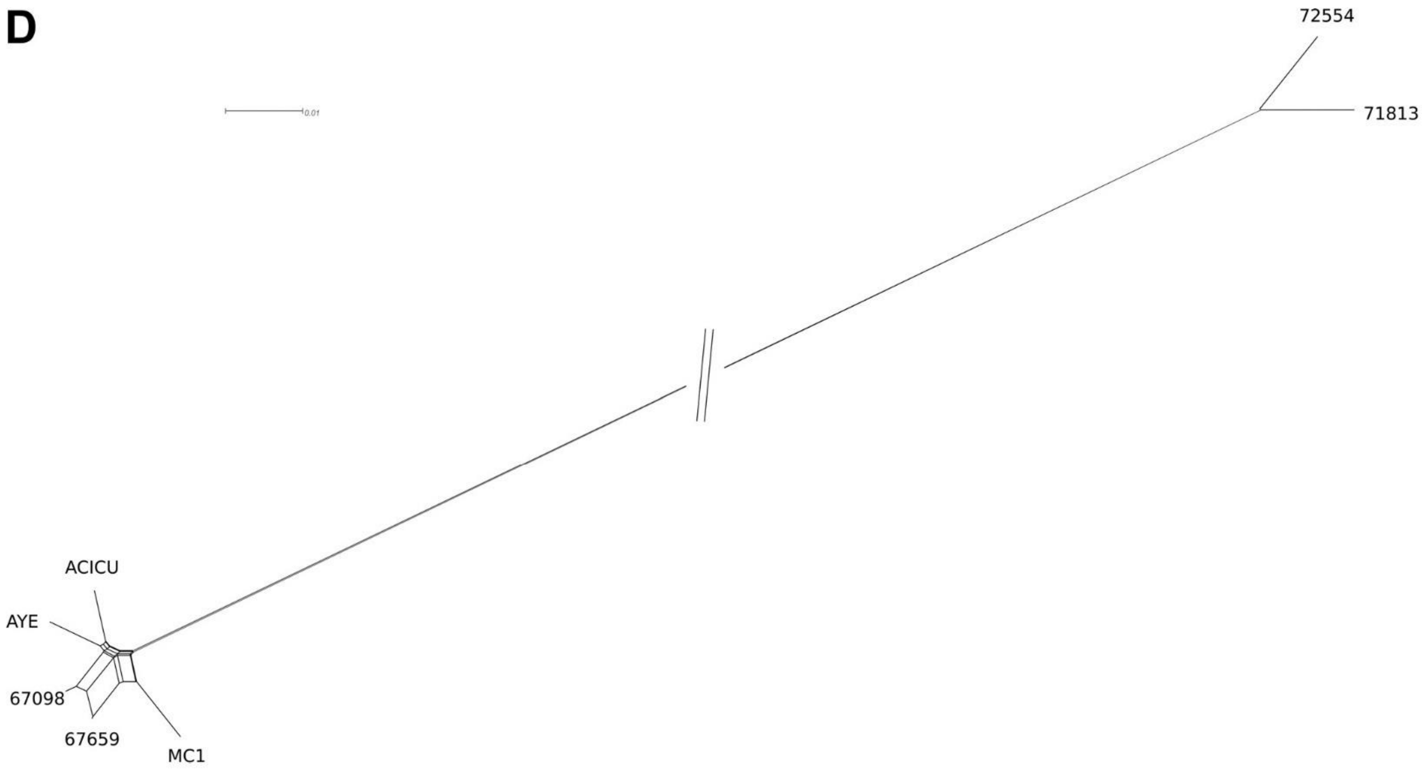


FIG 1 (Continued)

MGEs are often involved in horizontal gene transfer and, in *A. baumannii*, are frequently related to insertion sequences (ISs) and/or composite transposons (7, 16). Despite multiple copies of distinct IS elements being identified in the genomes of isolates 67659 and 72554 (data not shown), none of them was observed within or flanking the recombining region encompassing *pmrCAB*. In fact, the nearest IS detected was a copy of IS*Aba125* that was ~14 kb upstream of *pmrC* in both isolates, while in the other direction the closest IS element identified (a copy of IS*17*) was located >120 kb downstream of *pmrB*, suggesting that recombination was not mediated by DNA mobilization either through an IS or a composite transposon. Phage-related structures were also observed through the genome of both isolates. However, similar to the IS elements, none of them was found flanking the recombining regions, and the closest intact phage was observed >300 kb downstream of *pmrB*.

Considering that IS elements are self-transposable structures (17), we investigated the presence of inverted repeats flanking the recombining region, since they indicate that MGEs were lost postrecombination. A large number of repeats was observed within and flanking the recombining region in both isolates, with an average of 44 repeats per 1,000 bp. However, sequence analysis revealed that none of them were part of or constituted an insertion site for known IS elements. Moreover, they were also found at the same position in isolates 67098 and 71813, suggesting that they were translocated from IC5 and IC4 to IC7 during recombination, respectively, rather than being responsible for the DNA mobilization. Therefore, the mechanisms involved in the mobilization of *pmrCAB* into IC7 isolates remain to be elucidated.

Allelic variation in the *pmrCAB* operon is associated with natural polymorphisms within each *A. baumannii* IC. In our study, we demonstrated that IC7 isolates can present additional sequence divergence as a consequence of homologous recombination of regions with variable lengths across distinct clonal lineages. Interestingly, the recombination appears not to be driven by antibiotic pressure, since it was observed in both colistin-susceptible and -resistant isolates, and a variety of clonal lineages can act as donors of the recombining region. Additionally, we observed that MGEs were not required for the transfer of *pmrCAB* in our isolates, since neither IS elements nor

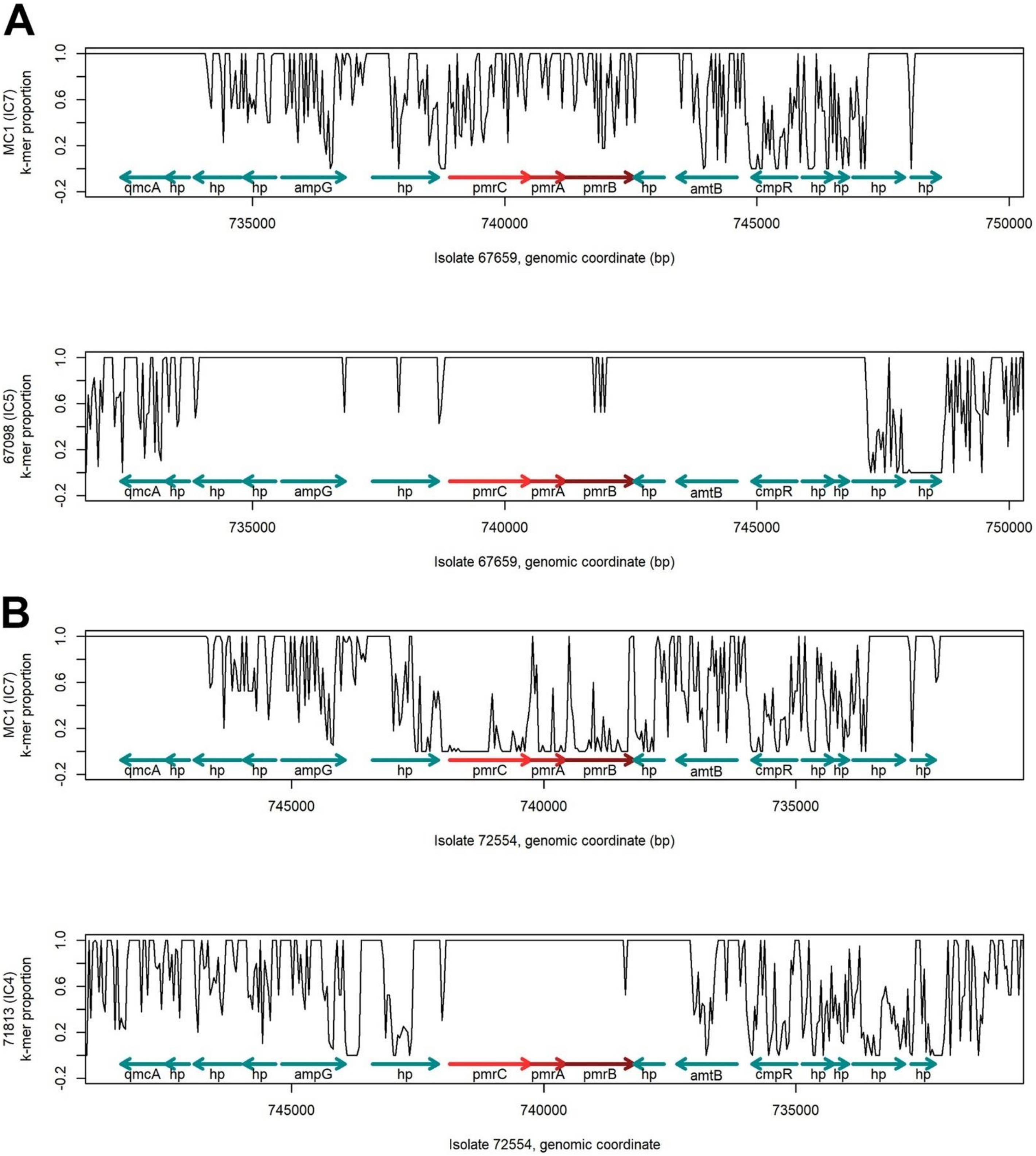


FIG 2 (A and B) Spatial k-mer sharing plots of a 23.6-kb genomic region encompassing *pmrCAB* and flanking genes of isolate 67659 against isolates MC1 (IC7, top) and 67098 (IC5, bottom) (A) and 72554 against MC1 (IC7, top) and 71813 (IC4, bottom) (B). The plots show spatial variations in the proportion of k-mers present in the genomes described on the *x* axis also present in the genome of the different references described on the *y* axis, calculated in sliding windows of 40 bases along the genome of the first isolate and for *k* = 19. Plots are based on k-mer counts computed with Cortex and a custom R visualization script. *pmrCAB* coding regions are highlighted in red, and flanking genes are indicated in green.

other MGEs were detected flanking the recombining region. Further studies are required to determine the mechanisms driving the mobilization of *pmrCAB* and to evaluate the presence of this phenomenon in other ICs as well as its frequency in the *A. baumannii* population.

MATERIALS AND METHODS

Bacterial isolates. *A. baumannii* clinical isolates 67659 and 72554 were recovered from the same tertiary hospital in the city of São Paulo, Brazil, 2 years apart (2015 and 2017, respectively). Their antimicrobial susceptibility profile was previously determined (14), and they were found to be colistin susceptible (MIC, 1 mg/liter) and resistant (MIC, >128 mg/liter), respectively. Their genomes were previously sequenced using the Illumina MiSeq platform, and cgMLST analysis revealed that the isolates had 28 allele differences and were grouped under IC7 (14). Additionally, previously described colistin-susceptible isolates belonging to IC4 (71813), IC5 (67098), and IC7 (MC1) were included as reference genomes for each IC (14, 18).

Long-read WGS using MinION platform. Genomic DNA of isolates 67659 and 72554 was extracted using the Genomic-Tips 100/G kit and genomic DNA buffers kit (Qiagen, Hilden, Germany). Libraries were prepared using the ligation sequencing kit (SQK-LSK109), combined with a native barcoding kit (EXP-NBD104) and the rapid barcoding kit (SQK-RBK004) (Oxford Nanopore Technologies, Oxford, United Kingdom), and were loaded onto an R9.4 flow cell (Oxford Nanopore Technologies). Genomes were assembled with a hybrid approach using Unicycler version 0.4.4 (19) with default parameters.

Genome alignment and identification of the recombining region including *pmrCAB*. The exact position of the *pmrCAB* operon was identified by aligning the *pmrCAB* sequence from *A. baumannii* ATCC 19606 (GenBank accession number NZ_CP045110.1) against the hybrid assemblies using the NUCmer tool of the MUMmer package, version 4.0.0beta2 (20), with default parameters. K-mer sharing plots were used for the robust identification of sequence homologies and recombination boundaries between lineages by visualizing spatial variation in the proportion of k-mers from one isolate (X) also present in another isolate (Y), calculated in sliding windows of 40 bases along the genome of X. In contrast to other alignment approaches, k-mer sharing plots do not require full assembly of genome Y but can be created based on short-read-derived k-mer counts. For a given region in isolate X, k-mer sharing values close to 1 indicate the likely presence of a homologous region in Y, whereas lower values indicate reduced similarity or the absence of the corresponding region from Y. The k-mer sharing plots were used to determine sequence homology patterns between different isolates around the *pmrCAB* operon and were created with a custom R script executed in RStudio (version 1.3.1093) (21). k-mer presence or absence was determined with Cortex (version 1.0.5.21; options "–mem_height 25," "–mem_width 100," and "–kmer_size 19") (22), employing a minimum k-mer coverage threshold of 10 for the analysis of short-read data. A neighbor-net analysis of the *pmrCAB* region was carried out with SplitsTree (23) with default settings, based on a MUSCLE (24) multiple-sequence alignment of identified *pmrCAB* sequences plus 10 kb of adjacent sequence from either side of *pmrCAB*. The phi test implemented in SplitsTree (null hypothesis: no recombination) was used to test for recombination.

Characterization of the mobile structures involved in *pmrCAB* recombination. To fully annotate the hybrid assemblies and to search for MGEs, Prokka version 1.14.5 (25) was used with default parameters. Putative IS elements and phage-related structures were further identified with the blast tools of IS-finder (https://isfinder.biotoul.fr/) and Phaster (https://phaster.ca/), respectively, using default parameters. Inverted repeats (IR) were identified using the repeat-match tool of the MUMmer package version 4.0.0beta2 (20) with a minimum repeat length of 10 bases.

Data availability. Short and long raw reads generated for IC7 isolates 67659 and 72554, as well as the reference isolates 67098 and 71813, were submitted to the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra/) of the National Center for Biotechnology Information (NCBI) under BioProject number PRJNA632943. Genome data from isolate MC1 are available under GenBank accession number NZ_QXPV00000000.1. Additional isolates presented in the supplemental material had their short raw reads submitted to the European Nucleotide Archive (http://www.ebi.ac.uk/ena/) of EMBL European Bioinformatics Institute (EBI) under the study accession numbers PRJEB12082 and PRJEB27899.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 0.9 MB.

FIG S2, PDF file, 0.5 MB.

FIG S3, PDF file, 0.9 MB.

FIG S4, PDF file, 0.1 MB.

FIG S5, PDF file, 0.1 MB.

TABLE S1, PDF file, 0.02 MB.

ACKNOWLEDGMENTS

Computational support and infrastructure were provided by the Centre for Information and Media Technology (ZIM) at the University of Düsseldorf (Germany).

This work was supported by the Jürgen Manchot Foundation. C.S.N. received a grant from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES-PDSE; process number 88881.187573/2018-01). K.X. was supported by the German Center for Infection Research (DZIF) and the Koeln Fortune Program/Faculty of Medicine, University of Cologne. A.C.G. received a research grant from the National Council for Science and

Technological Development (CNPq-process number 312066/2019). P.G.H. was supported by the German Research Council (DFG)-FOR2251 (www.acinetobacter.de).

A.C.G. has recently received research funding and/or consultation fees from Eurofarma, InfectoPharm, MSD, Pfizer, and United Medical. Other authors have nothing to declare. This study was not financially supported by any diagnostic/pharmaceutical company.

A.D. and P.G.H. jointly supervised this work.

REFERENCES

1. Antunes LC, Visca P, Towner KJ. 2014. *Acinetobacter baumannii*: evolution of a global pathogen. Pathog Dis 71:292–301. https://doi.org/10.1111/2049-632X.12125.
2. Snitkin ES, Zelazny AM, Montero CI, Stock F, Mijares L, Murray PR, Segre JA, NISC Comparative Sequence Program. 2011. Genome-wide recombination drives diversification of epidemic strains of *Acinetobacter baumannii*. Proc Natl Acad Sci U S A 108:13758–13763. https://doi.org/10.1073/pnas.1104404108.
3. Müller C, Stefanik D, Wille J, Hackel M, Higgins PG, Seifert H. 2019. Molecular epidemiology of carbapenem-resistant Acinetobacter baumannii clinical isolates and identification of the novel international clone IC9: results from a worldwide surveillance study (2012–2016). 29th Eur Cong Clin Microbiol Infect Dis (ECCMID), Amsterdam, Netherlands.
4. Higgins PG, Prior K, Harmsen D, Seifert H. 2017. Development and evaluation of a core genome multilocus typing scheme for whole-genome sequence-based typing of *Acinetobacter baumannii*. PLoS One 12:e0179228. https://doi.org/10.1371/journal.pone.0179228.
5. Zander E, Nemec A, Seifert H, Higgins PG. 2012. Association between *β*-lactamase-encoding *bla*(OXA-51) variants and DiversiLab rep-PCR-based typing of *Acinetobacter baumannii* isolates. J Clin Microbiol 50:1900–1904. https://doi.org/10.1128/JCM.06462-11.
6. Furuya EY, Lowy FD. 2006. Antimicrobial-resistant bacteria in the community setting. Nat Rev Microbiol 4:36–45. https://doi.org/10.1038/nrmicro1325.
7. Zhang G, Leclercq SO, Tian J, Wang C, Yahara K, Ai G, Liu S, Feng J. 2017. A new subclass of intrinsic aminoglycoside nucleotidyltransferases, ANT (3″)-II, is horizontally transferred among *Acinetobacter* spp. by homologous recombination. PLoS Genet 13:e1006602. https://doi.org/10.1371/journal.pgen.1006602.
8. Mussi MA, Limansky AS, Relling V, Ravasi P, Arakaki A, Actis LA, Viale AM. 2011. Horizontal gene transfer and assortative recombination within the *Acinetobacter baumannii* clinical population provide genetic diversity at the single *carO* gene, encoding a major outer membrane protein channel. J Bacteriol 193:4736–4748. https://doi.org/10.1128/JB.01533-10.
9. Hamidian M, Hancock DP, Hall RM. 2013. Horizontal transfer of an IS*Aba125*-activated *ampC* gene between *Acinetobacter baumannii* strains leading to cephalosporin resistance. J Antimicrob Chemother 68:244–245. https://doi.org/10.1093/jac/dks345.
10. Poirel L, Jayol A, Nordmann P. 2017. Polymyxins: antibacterial activity, susceptibility testing, and resistance mechanisms encoded by plasmids or chromosomes. Clin Microbiol Rev 30:557–596. https://doi.org/10.1128/CMR.00064-16.
11. Gerson S, Lucaßen K, Wille J, Nodari CS, Stefanik D, Nowak J, Wille T, Betts JW, Roca I, Vila J, Cisneros JM, Seifert H, Higgins PG. 2020. Diversity of amino acid substitutions in PmrCAB associated with colistin resistance in clinical isolates of *Acinetobacter baumannii*. Int J Antimicrob Agents 55:105862. https://doi.org/10.1016/j.ijantimicag.2019.105862.
12. Nodari CS, Cayô R, Streling AP, Lei F, Wille J, Almeida MS, de Paula AI, Pignatari ACC, Seifert H, Higgins PG, Gales AC. 2020. Genomic analysis of carbapenem-resistant *Acinetobacter baumannii* isolates belonging to major endemic clones in South America. Front Microbiol 11:584603. https://doi.org/10.3389/fmicb.2020.584603.
13. Kim DH, Ko KS. 2015. A distinct alleles and genetic recombination of *pmrCAB* operon in species of *Acinetobacter baumannii* complex isolates. Diagn Microbiol Infect Dis 82:183–188. https://doi.org/10.1016/j.diagmicrobio.2015.03.021.
14. Nodari CS, Higgins PG, Silva RC, Streling A, Lei F, Wille J, Seifert H, Gales A. 2019. Emergence of a pan-drug resistance phenotype among Acinetobacter baumannii isolates belonging to major epidemic clones in Brazil. 29th Eur Cong Clin Microbiol Infect Dis (ECCMID), Amsterdam, Netherlands.
15. Boinett CJ, Cain AK, Hawkey J, Do Hoang NT, Khanh NNT, Thanh DP, Dordel J, Campbell JI, Lan NPH, Mayho M, Langridge GC, Hadfield J, Chau NVV, Thwaites GE, Parkhill J, Thomson NR, Holt KE, Baker S. 2019. Clinical and laboratory-induced colistin-resistance mechanisms in *Acinetobacter baumannii*. Microb Genom 5:e000246. https://doi.org/10.1099/mgen.0.000246.
16. Hamidian M, Hall RM. 2014. Tn*6168*, a transposon carrying an IS*Aba1*-activated *ampC* gene and conferring cephalosporin resistance in *Acinetobacter baumannii*. J Antimicrob Chemother 69:77–80. https://doi.org/10.1093/jac/dkt312.
17. Vandecraen J, Chandler M, Aertsen A, Van Houdt R. 2017. The impact of insertion sequences on bacterial genome plasticity and adaptability. Crit Rev Microbiol 43:709–730. https://doi.org/10.1080/1040841X.2017.1303661.
18. Cerezales M, Xanthopoulou K, Wille J, Krut O, Seifert H, Gallego L, Higgins PG. 2020. Mobile genetic elements harboring antibiotic resistance determinants in *Acinetobacter baumannii* isolates from Bolivia. Front Microbiol 11:919. https://doi.org/10.3389/fmicb.2020.00919.
19. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595.
20. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol 14:e1005944. https://doi.org/10.1371/journal.pcbi.1005944.
21. RStudio Team. 2020. RStudio: integrated development for R. RStudio, PBC, Boston, MA. http://www.rstudio.com/.
22. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet 44:226–232. https://doi.org/10.1038/ng.1028.
23. Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267. https://doi.org/10.1093/molbev/msj030.
24. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113. https://doi.org/10.1186/1471-2105-5-113.
25. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

3.4 Supplementary material to publication 2

All supplementary figures and tables mentioned in the publication are listed on the following pages. They can also be accessed from the original publication source (https://doi.org/10.1128/msphere.00746-21), or via the enclosed CD.

Fig S1 – 1/6

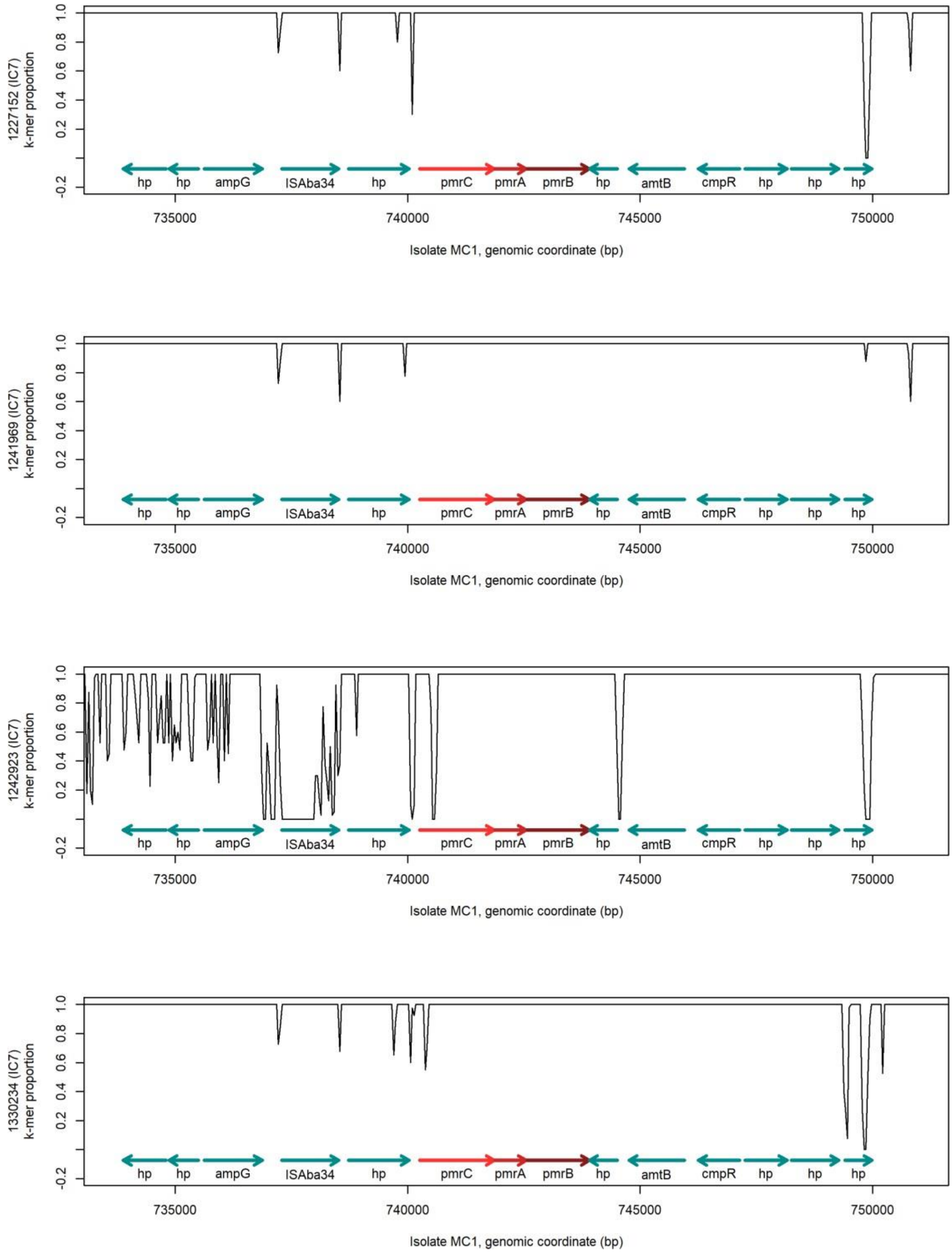
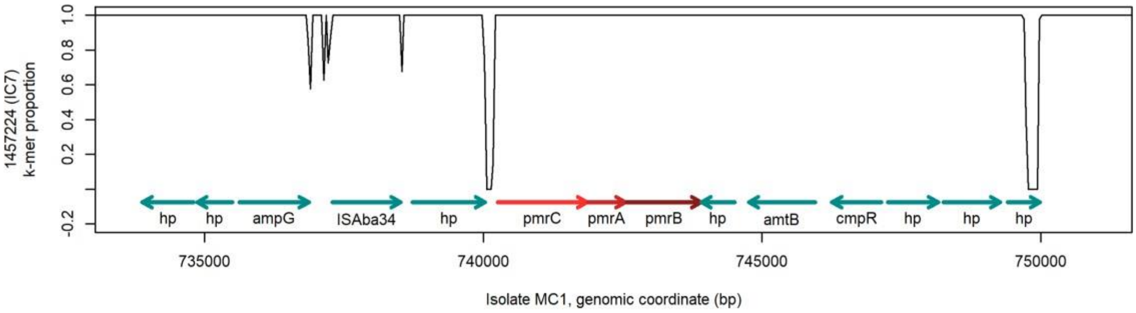
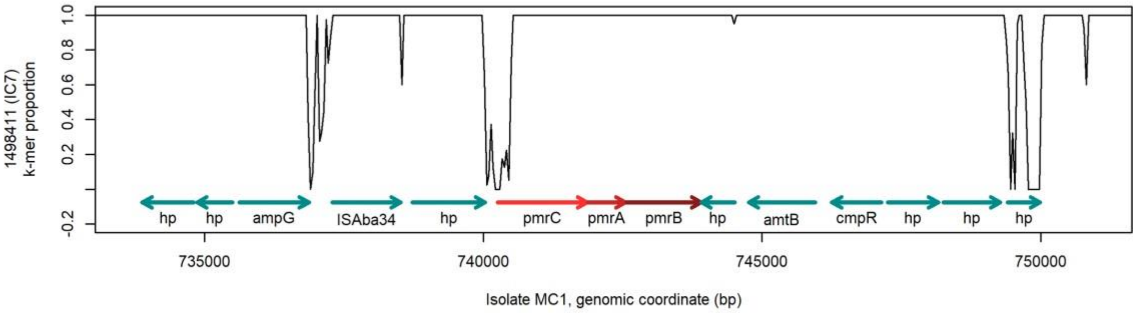
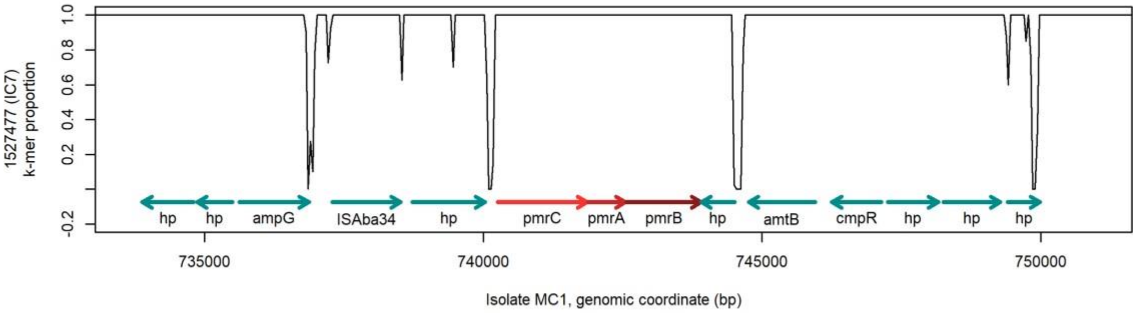


Fig S1 – 2/6


1457224 (IC7)
k-mer proportion
1.0
0.8
0.6
0.4
0.2
-0.2
hp
hp
ampG
ISAba34
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
735000
740000
745000
750000
Isolate MC1, genomic coordinate (bp)


1498411 (IC7)
k-mer proportion
1.0
0.8
0.6
0.4
0.2
-0.2
hp
hp
ampG
ISAba34
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
735000
740000
745000
750000
Isolate MC1, genomic coordinate (bp)


1527477 (IC7)
k-mer proportion
1.0
0.8
0.6
0.4
0.2
-0.2
hp
hp
ampG
ISAba34
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
735000
740000
745000
750000
Isolate MC1, genomic coordinate (bp)

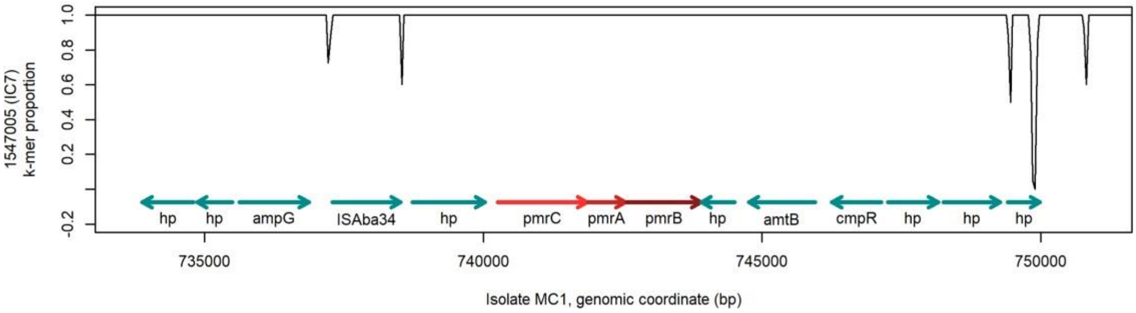

1547005 (IC7)
k-mer proportion
1.0
0.8
0.6
0.4
0.2
-0.2
hp
hp
ampG
ISAba34
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
735000
740000
745000
750000
Isolate MC1, genomic coordinate (bp)

Fig S1 – 3/6

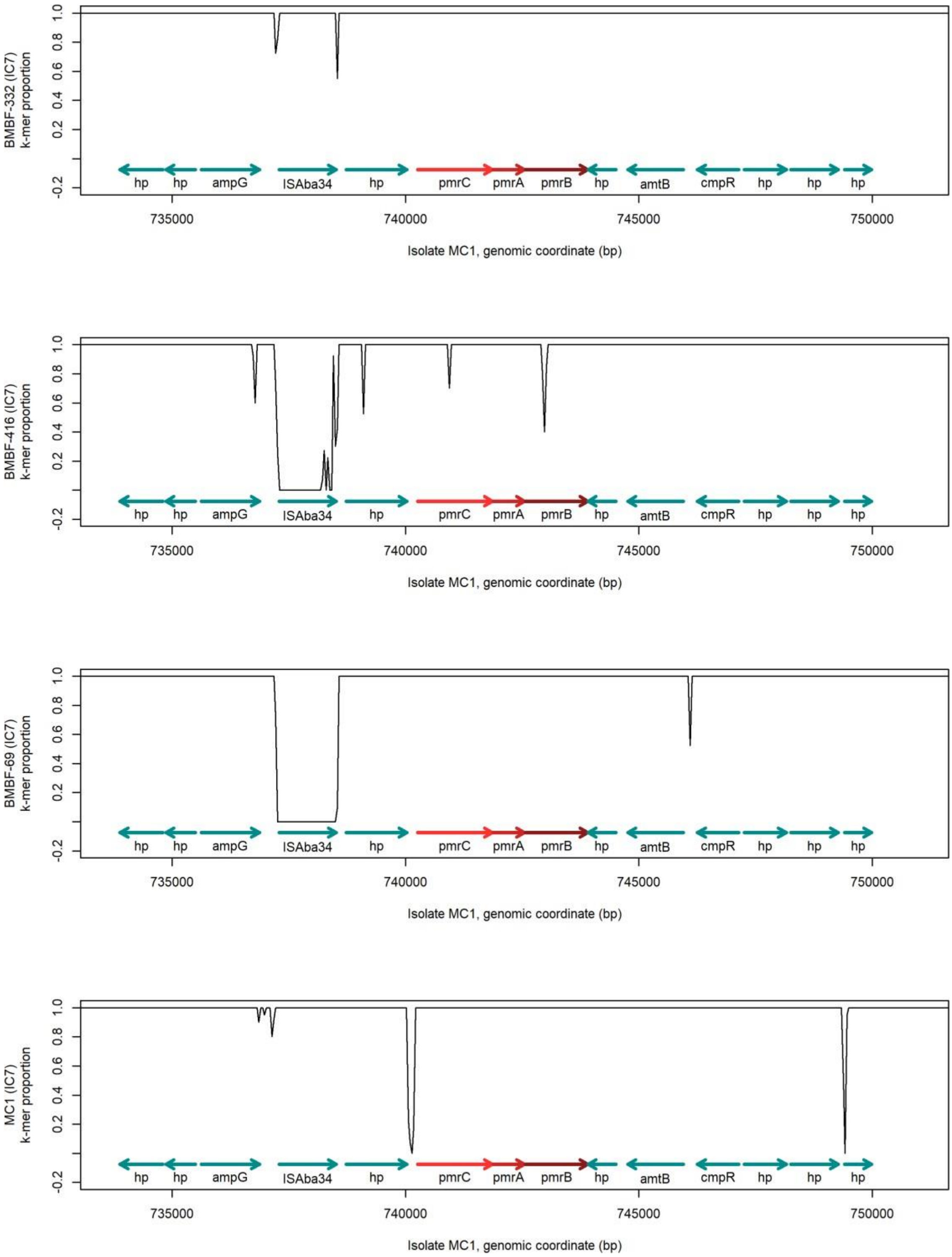

BMBF-332 (IC7)
k-mer proportion
1.0
0.8
0.6
0.4
0.2
-0.2
hp
hp
ampG
ISAba34
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
735000
740000
745000
750000
Isolate MC1, genomic coordinate (bp)
BMBF-416 (IC7)
k-mer proportion
1.0
0.8
0.6
0.4
0.2
-0.2
hp
hp
ampG
ISAba34
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
735000
740000
745000
750000
Isolate MC1, genomic coordinate (bp)
BMBF-69 (IC7)
k-mer proportion
1.0
0.8
0.6
0.4
0.2
-0.2
hp
hp
ampG
ISAba34
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
735000
740000
745000
750000
Isolate MC1, genomic coordinate (bp)
MC1 (IC7)
k-mer proportion
1.0
0.8
0.6
0.4
0.2
-0.2
hp
hp
ampG
ISAba34
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
735000
740000
745000
750000
Isolate MC1, genomic coordinate (bp)

Fig S1 – 4/6

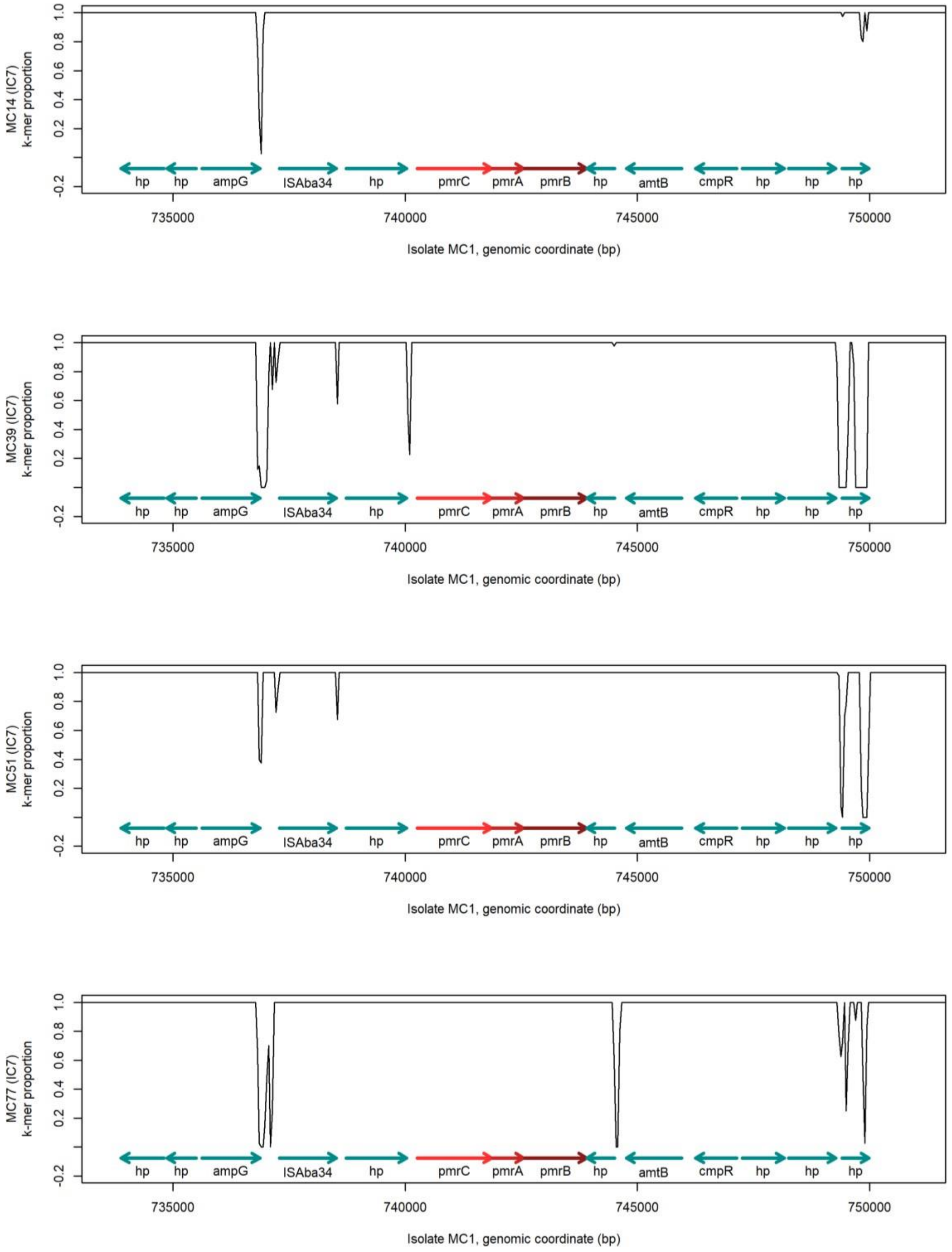

MC14 (IC7)
k-mer proportion
MC39 (IC7)
k-mer proportion
MC51 (IC7)
k-mer proportion
MC77 (IC7)
k-mer proportion
-0.2
0.2
0.4
0.6
0.8
1.0
hp
hp
ampG
ISAba34
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
735000
740000
745000
750000
Isolate MC1, genomic coordinate (bp)

Fig S1 – 5/6

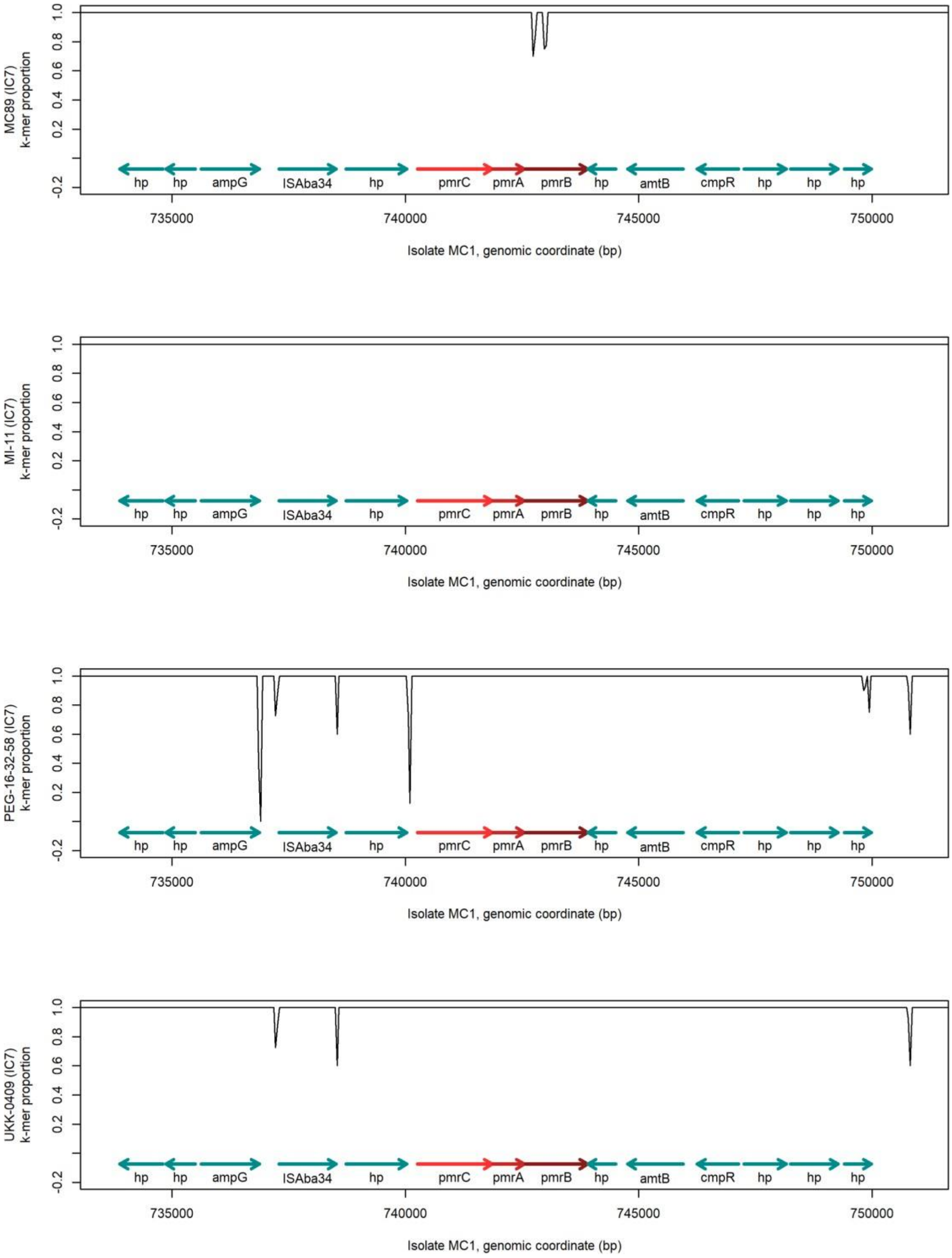
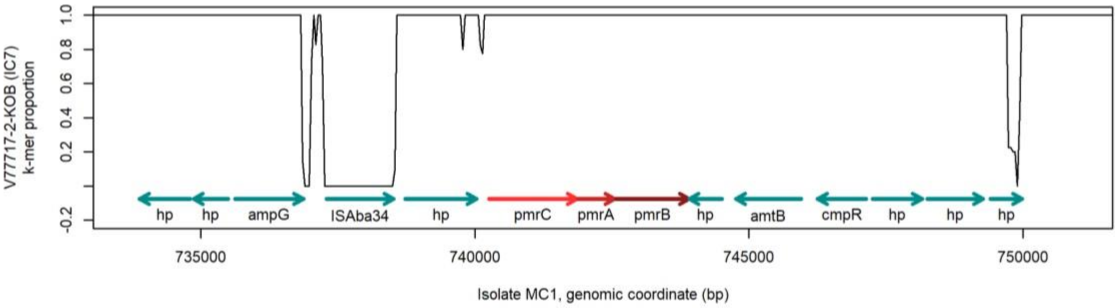

MC89 (IC7)
k-mer proportion
MI-11 (IC7)
k-mer proportion
PEG-16-32-58 (IC7)
k-mer proportion
UKK-0409 (IC7)
k-mer proportion
-0.2
0.2
0.4
0.6
0.8
1.0
hp
hp
ampG
ISAba34
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
735000
740000
745000
750000
Isolate MC1, genomic coordinate (bp)

Fig S1 – 6/6



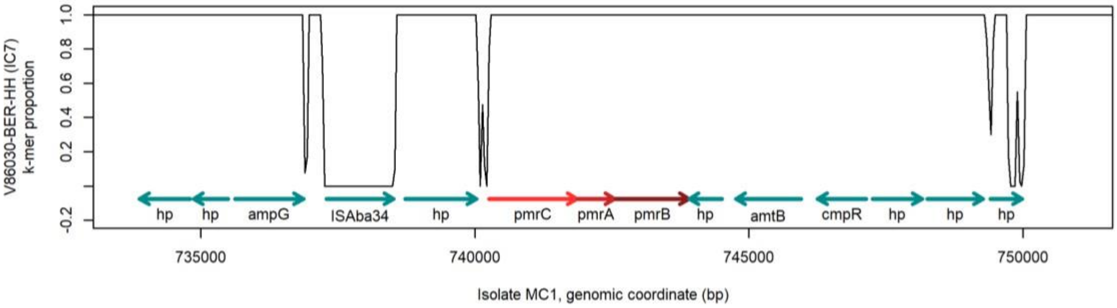


Figure S1: Spatial k-mer sharing plots of *pmrCAB* and flanking regions of isolates belonging to IC7 against MC1. The plots show spatial variations in the proportion of k-mers present in MC1 also observed in the genome of the different isolates described in the *y* axis, calculated in sliding windows of 40 bases along the genome of the MC1 and for $k = 19$. Plots are based on k-mer counts computed with Cortex and a custom R visualization script. *pmrCAB* coding regions are highlighted in red, and flanking genes are indicated in green.

Fig S2 – 1/3

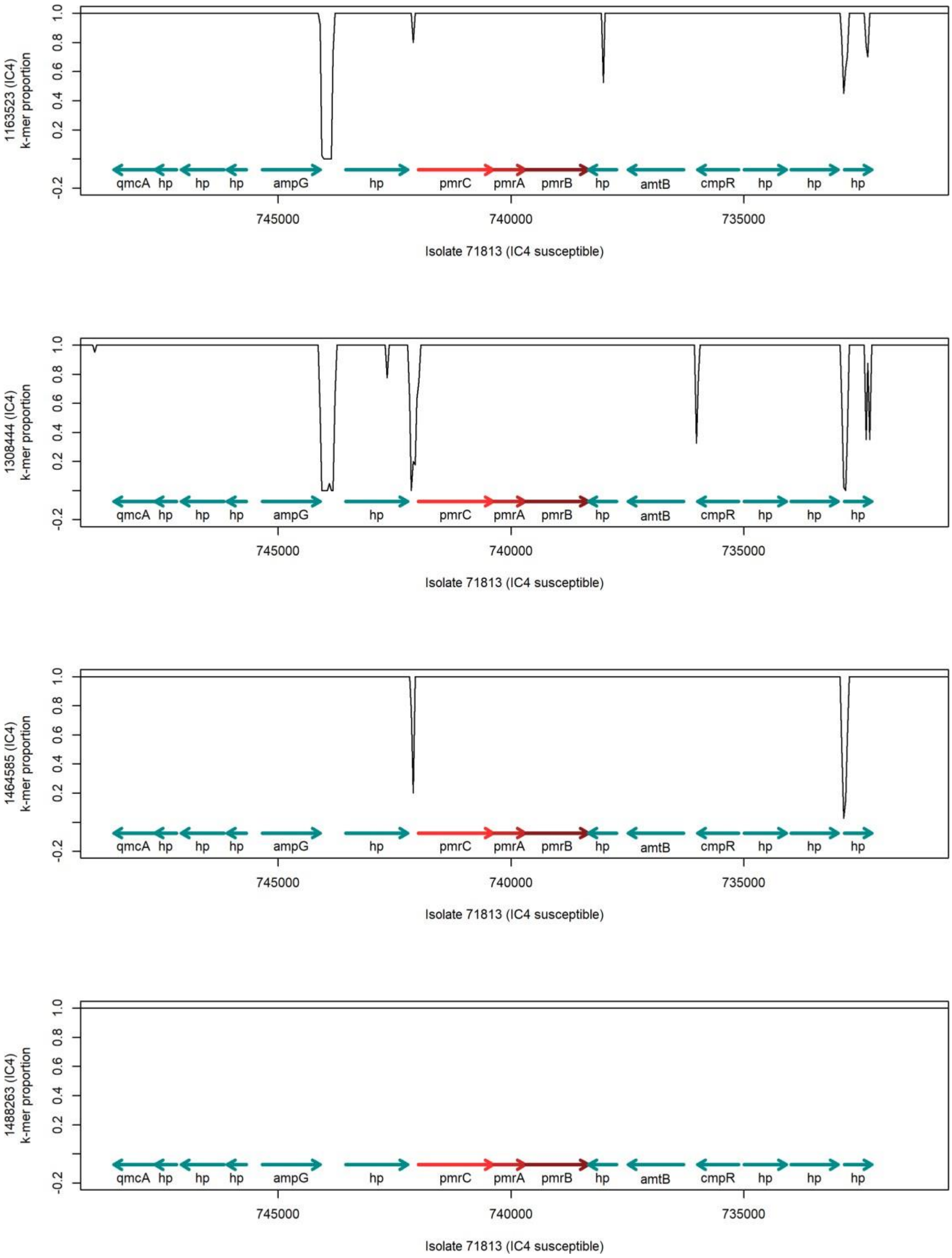

1163523 (IC4)
k-mer proportion
-0.2
0.2
0.4
0.6
0.8
1.0
qmcA
hp
hp
hp
ampG
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
745000
740000
735000
Isolate 71813 (IC4 susceptible)
1308444 (IC4)
k-mer proportion
-0.2
0.2
0.4
0.6
0.8
1.0
qmcA
hp
hp
hp
ampG
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
745000
740000
735000
Isolate 71813 (IC4 susceptible)
1464585 (IC4)
k-mer proportion
-0.2
0.2
0.4
0.6
0.8
1.0
qmcA
hp
hp
hp
ampG
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
745000
740000
735000
Isolate 71813 (IC4 susceptible)
1488263 (IC4)
k-mer proportion
-0.2
0.2
0.4
0.6
0.8
1.0
qmcA
hp
hp
hp
ampG
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
745000
740000
735000
Isolate 71813 (IC4 susceptible)

Fig S2 – 2/3

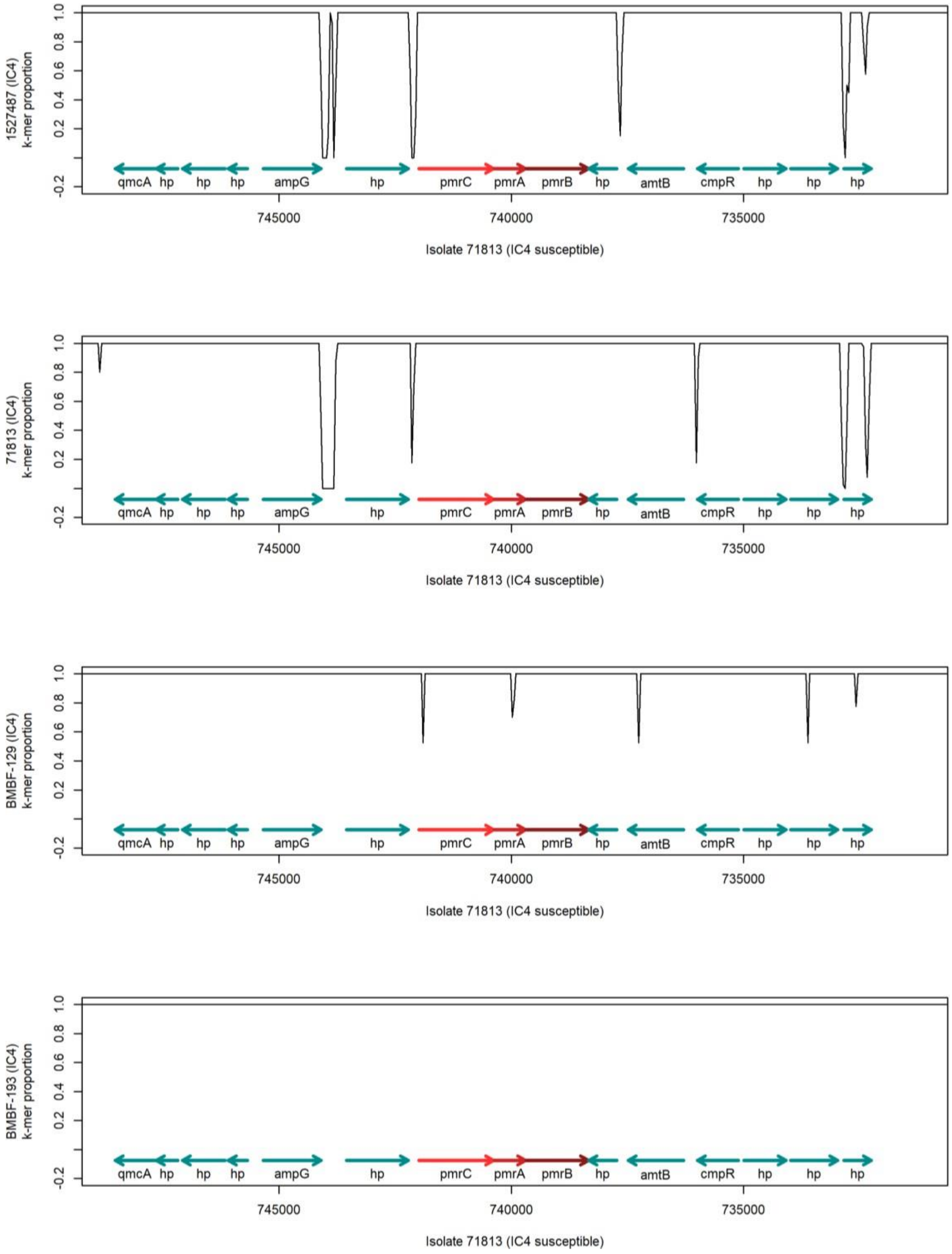

1527487 (IC4)
k-mer proportion
-0.2
0.2
0.4
0.6
0.8
1.0
qmcA
hp
hp
hp
ampG
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
745000
740000
735000
Isolate 71813 (IC4 susceptible)
71813 (IC4)
k-mer proportion
-0.2
0.2
0.4
0.6
0.8
1.0
qmcA
hp
hp
hp
ampG
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
745000
740000
735000
Isolate 71813 (IC4 susceptible)
BMBF-129 (IC4)
k-mer proportion
-0.2
0.2
0.4
0.6
0.8
1.0
qmcA
hp
hp
hp
ampG
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
745000
740000
735000
Isolate 71813 (IC4 susceptible)
BMBF-193 (IC4)
k-mer proportion
-0.2
0.2
0.4
0.6
0.8
1.0
qmcA
hp
hp
hp
ampG
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
745000
740000
735000
Isolate 71813 (IC4 susceptible)

Fig S2 – 3/3

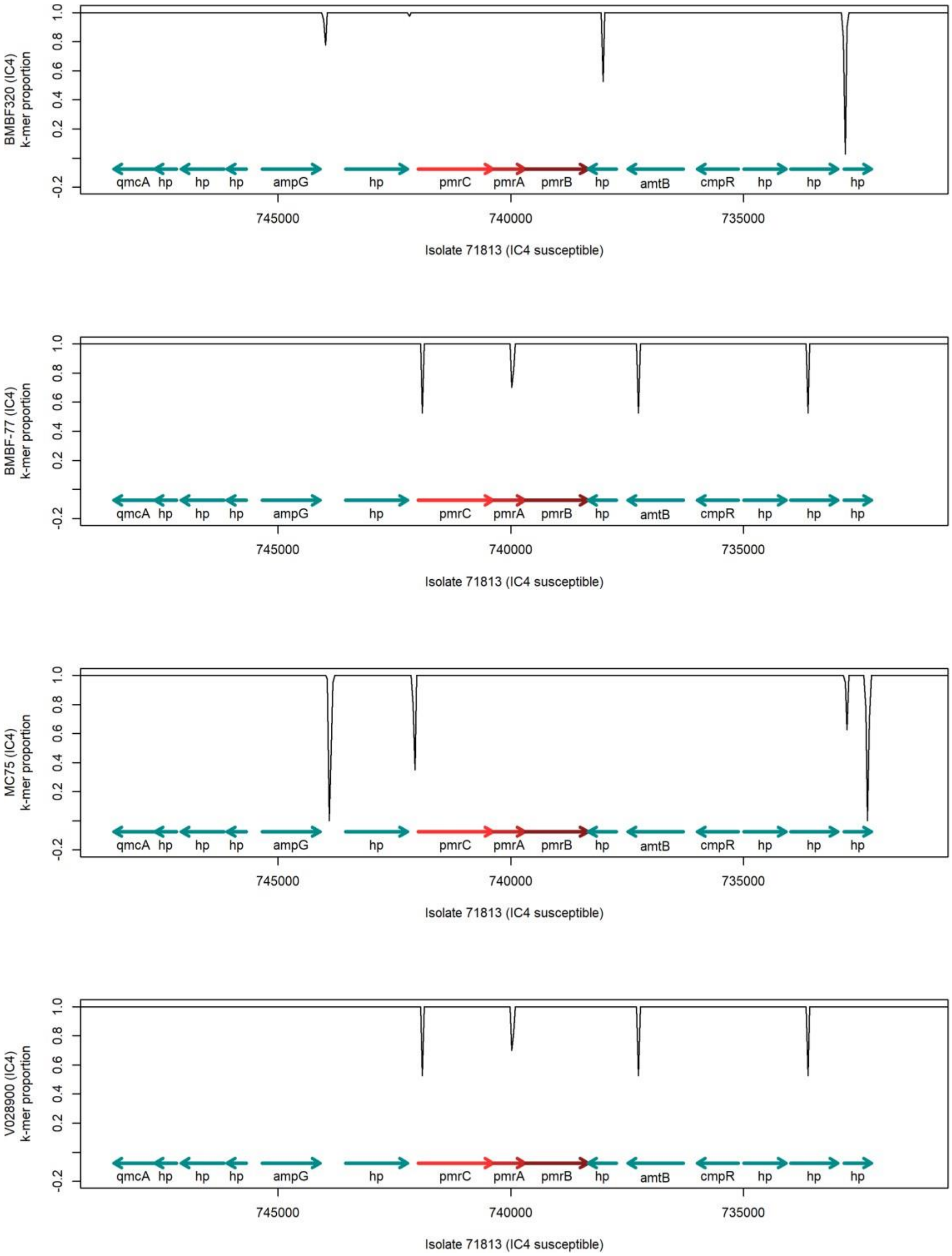


Figure S2: Spatial k-mer sharing plots of *pmrCAB* and flanking regions of isolates belonging to IC4 against 71813. The plots show spatial variations in the proportion of k-mers present in 71813 also observed in the genome of the different isolates described on the *y* axis, calculated in sliding windows of 40 bases along the genome of the 71813 and for $k = 19$. Plots are based on k-mer counts computed with Cortex and a custom R visualization script. *pmrCAB* coding regions are highlighted in red, and flanking genes are indicated in green.

Fig S3 – 1/5

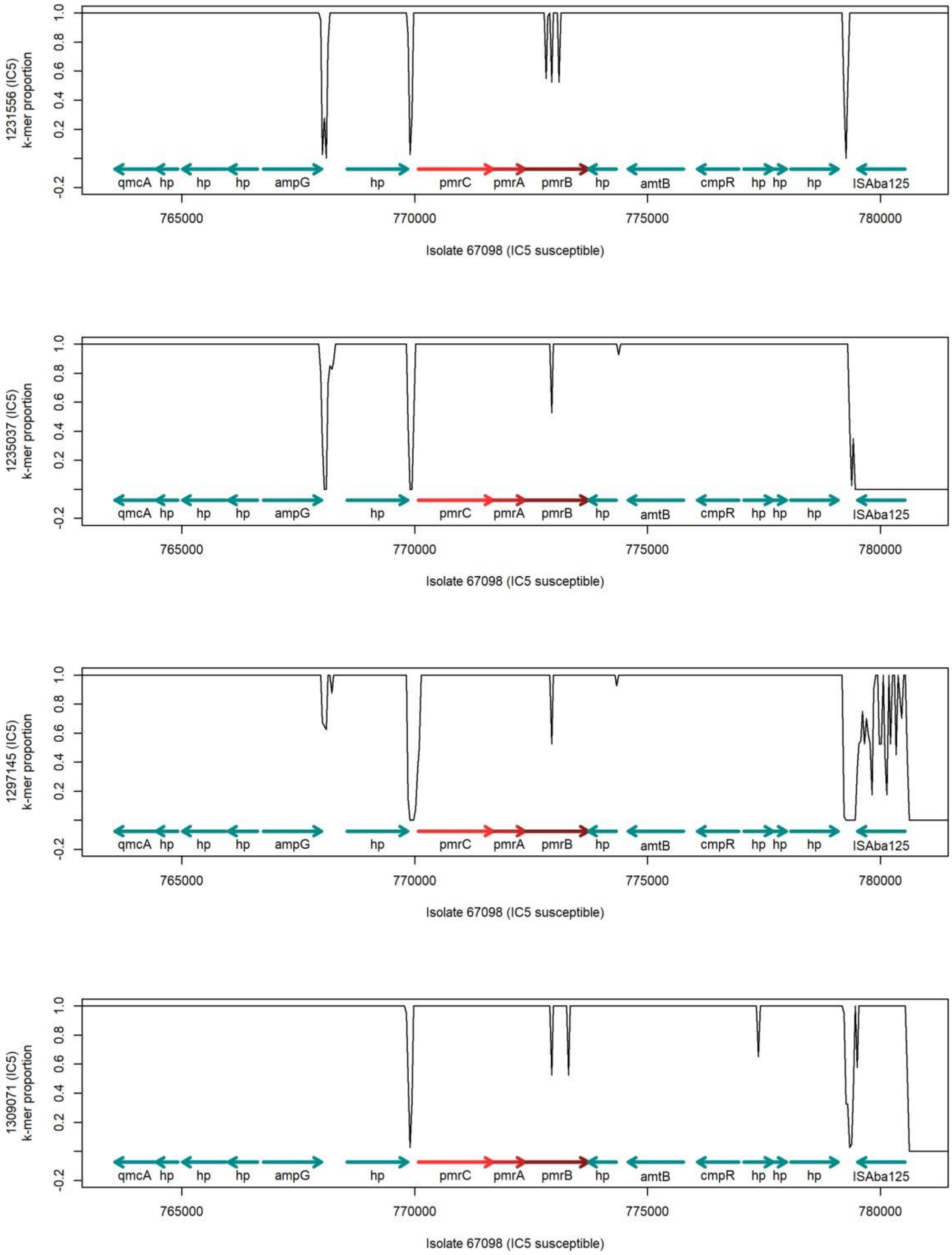

1231556 (IC5)
k-mer proportion
1235037 (IC5)
k-mer proportion
1297145 (IC5)
k-mer proportion
1309071 (IC5)
k-mer proportion
-0.2
0.2
0.4
0.6
0.8
1.0
qmcA
hp
hp
hp
ampG
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
ISAba125
765000
770000
775000
780000
Isolate 67098 (IC5 susceptible)

Fig S3 – 2/5

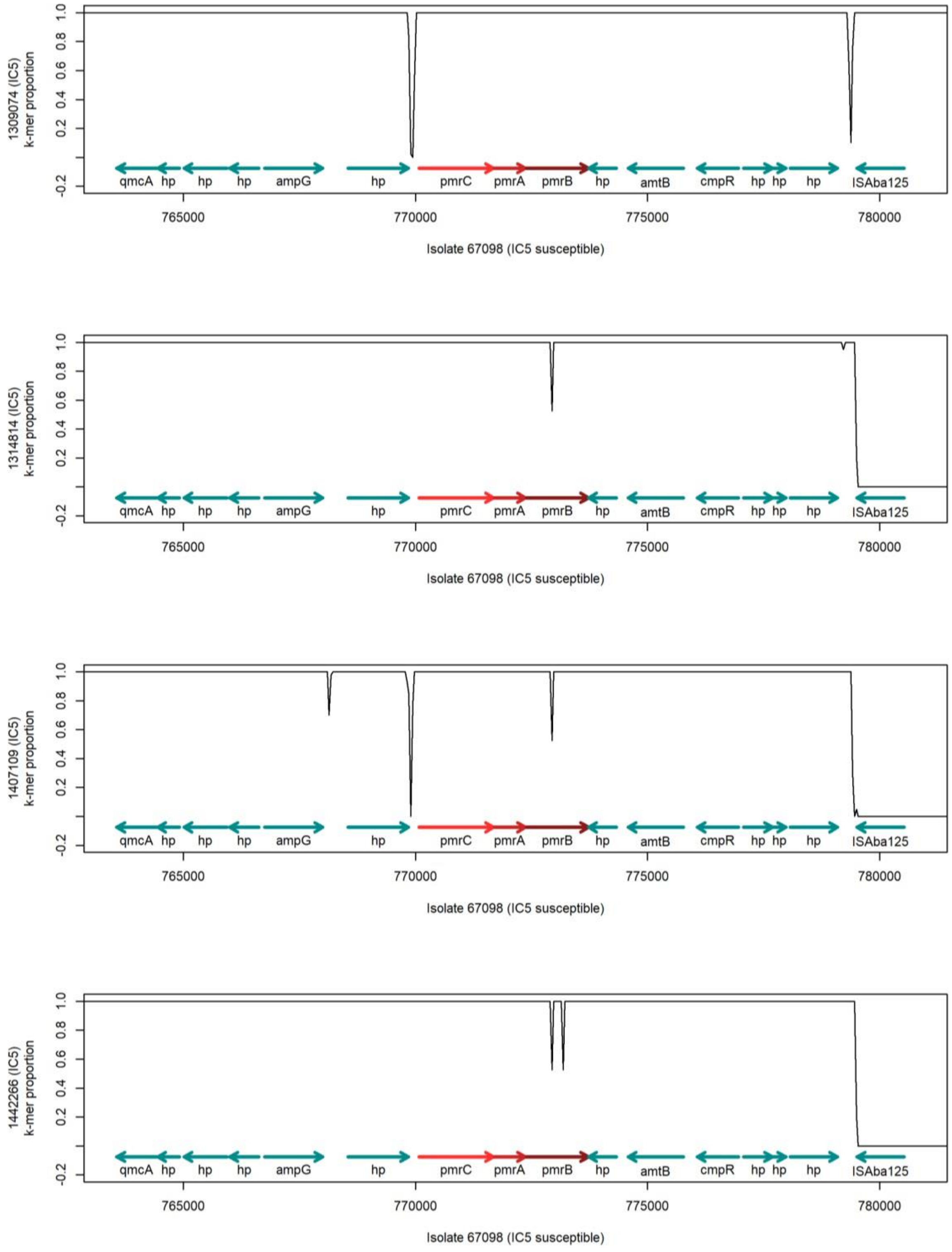

1309074 (IC5)
k-mer proportion
-0.2
0.2
0.4
0.6
0.8
1.0
qmcA
hp
hp
hp
ampG
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
ISAba125
765000
770000
775000
780000
Isolate 67098 (IC5 susceptible)
1314814 (IC5)
k-mer proportion
Isolate 67098 (IC5 susceptible)
1407109 (IC5)
k-mer proportion
Isolate 67098 (IC5 susceptible)
1442266 (IC5)
k-mer proportion
Isolate 67098 (IC5 susceptible)

Fig S3 – 3/5

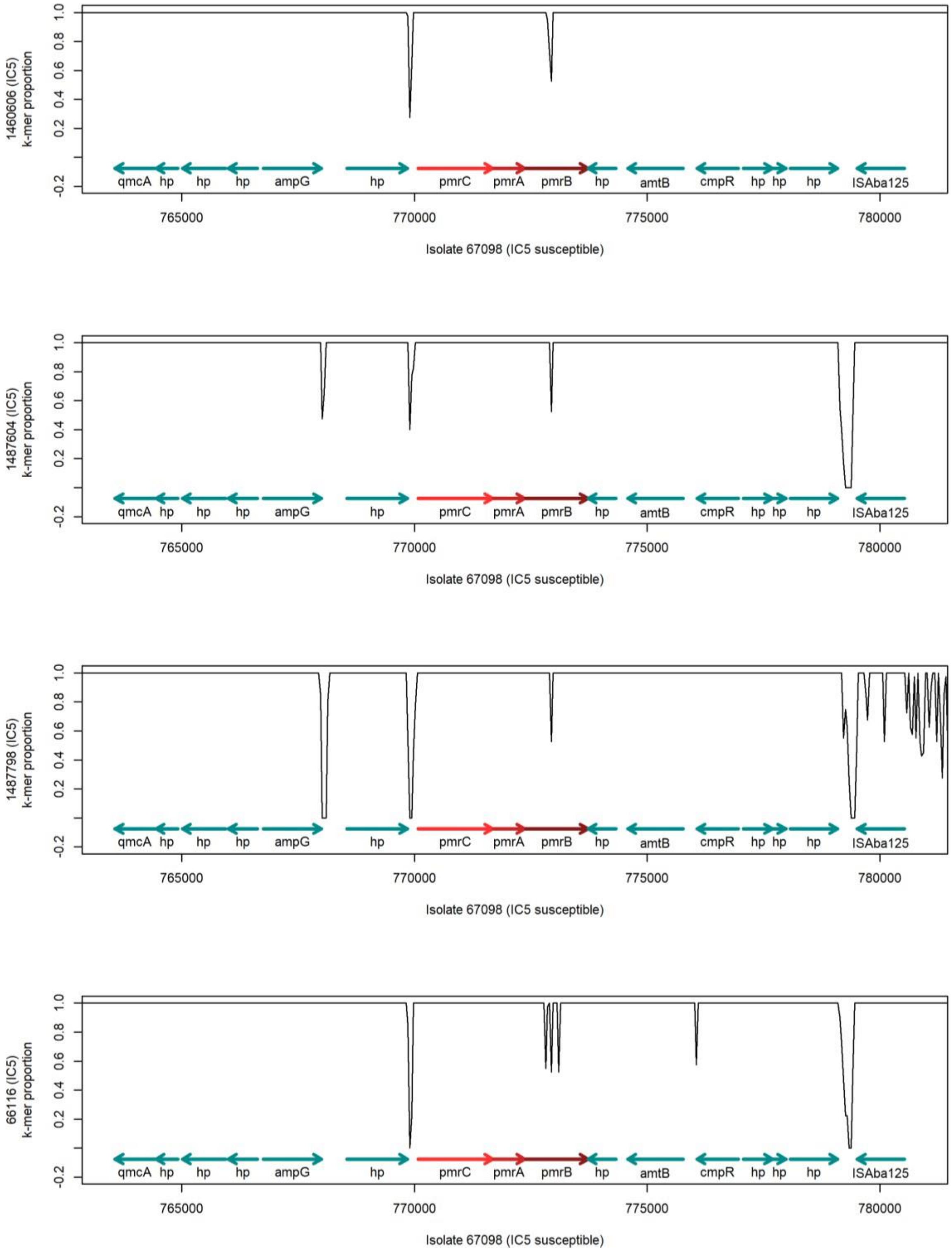

1460606 (IC5)
k-mer proportion
1487604 (IC5)
k-mer proportion
1487798 (IC5)
k-mer proportion
66116 (IC5)
k-mer proportion
-0.2
0.2
0.4
0.6
0.8
1.0
qmcA
hp
hp
hp
ampG
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
ISAba125
765000
770000
775000
780000
Isolate 67098 (IC5 susceptible)

Fig S3 – 4/5

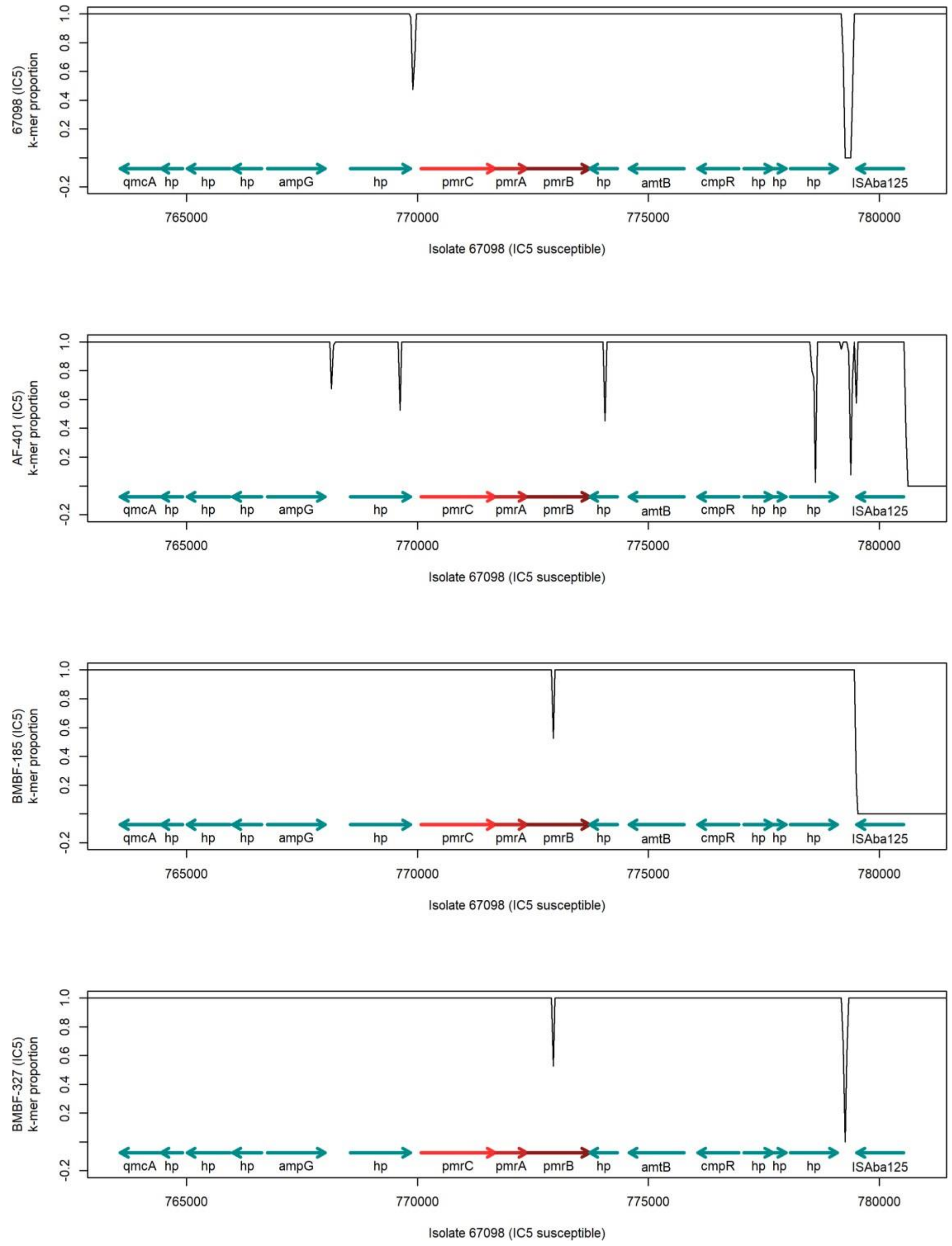

67098 (IC5)
k-mer proportion
AF-401 (IC5)
k-mer proportion
BMBF-185 (IC5)
k-mer proportion
BMBF-327 (IC5)
k-mer proportion
-0.2
0.2
0.4
0.6
0.8
1.0
qmcA
hp
hp
hp
ampG
hp
pmrC
pmrA
pmrB
hp
amtB
cmpR
hp
hp
hp
ISAba125
765000
770000
775000
780000
Isolate 67098 (IC5 susceptible)

Fig S3 – 5/5

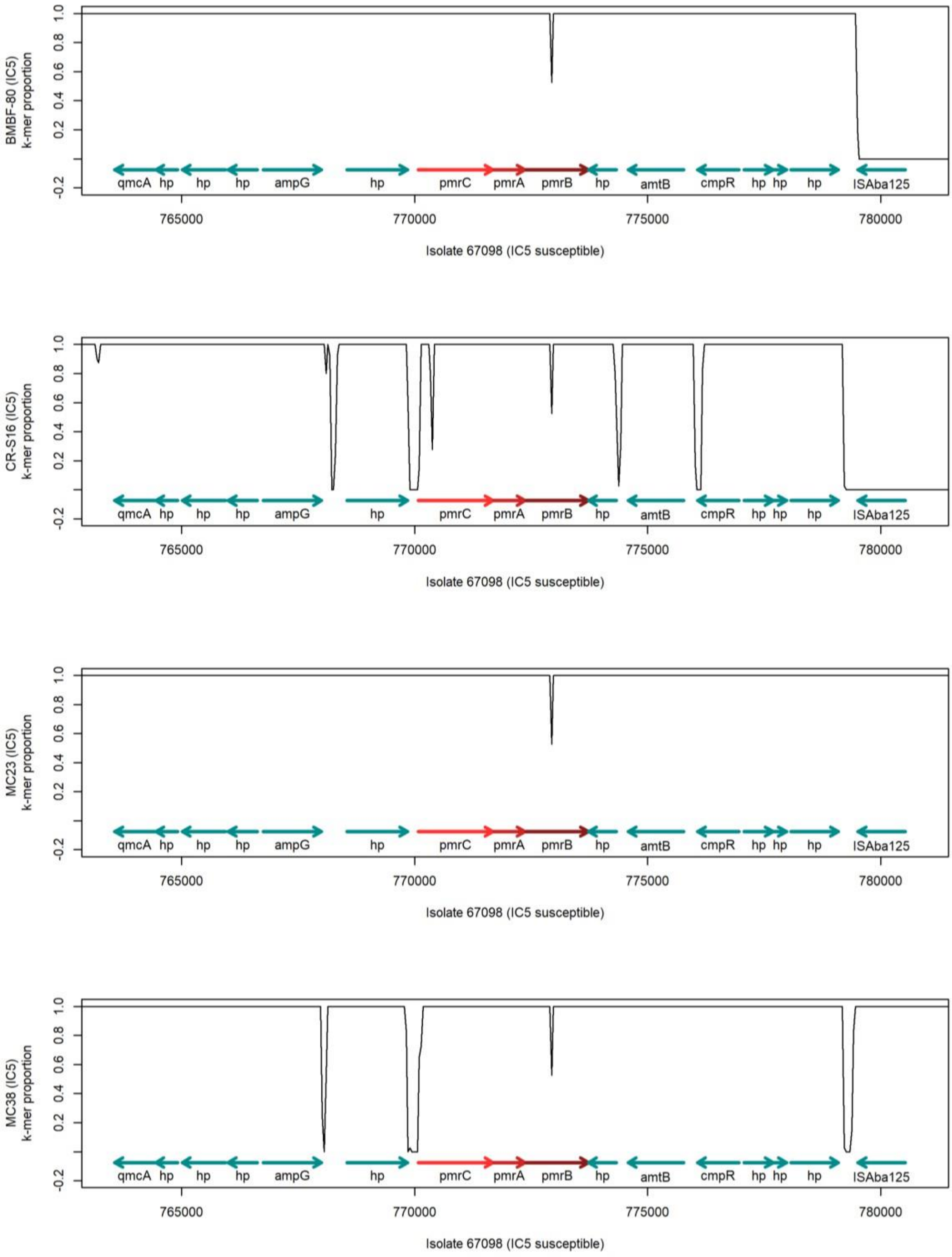


Figure S3: Spatial k-mer sharing plots of *pmrCAB* and flanking regions of isolates belonging to IC5 against 67098. The plots show spatial variations in the proportion of k-mers present in 67098 also observed in the genome of the different isolates described on the *y* axis, calculated in sliding windows of 40 bases along the genome of the 67098 and for $k = 19$. Plots are based on k-mer counts computed with Cortex and a custom R visualization script. *pmrCAB* coding regions are highlighted in red, and flanking genes are indicated in green.

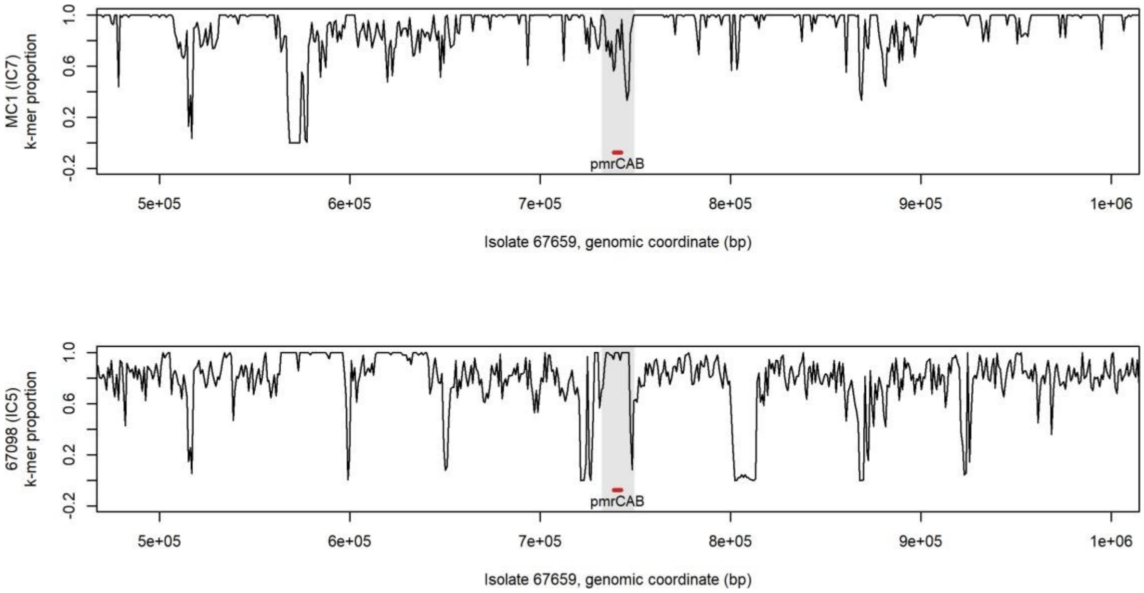


Figure S4: Spatial k-mer sharing plots of a 500-kb genomic region encompassing *pmrCAB* and flanking genes of isolate 67659 against isolates MC1 (IC7, top) and 67098 (IC5, bottom). The plots show spatial variations in the proportion of k-mers present in the genomes described on the *x* axis also present in the genome of the different references described on the *y* axis, calculated in sliding windows of 40 bases along the genome of the first isolate and for *k* = 19. Plots are based on k-mer counts computed with Cortex and a custom R visualization script. *pmrCAB* coding regions are highlighted in red. The shaded grey area highlights the genomic region depicted in Fig. 1A.

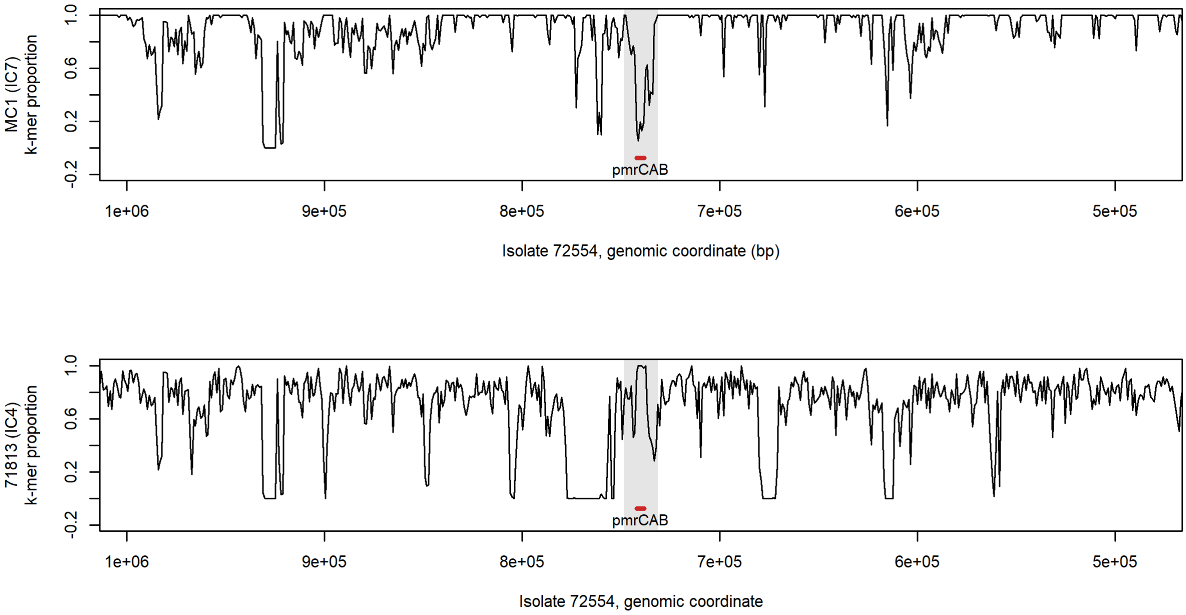
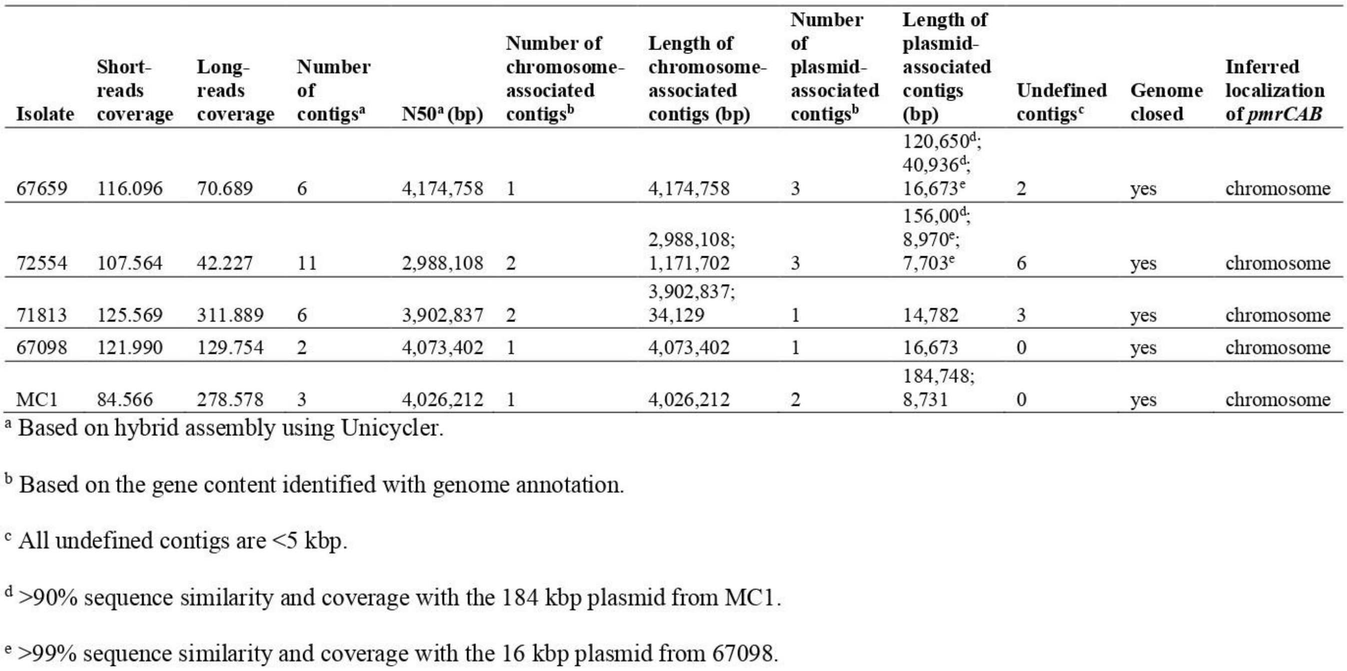


Figure S5: Spatial k-mer sharing plots of a 500-kb genomic region encompassing *pmrCAB* and flanking genes of isolate 72554 against isolates MC1 (IC7, top) and 71813 (IC4, bottom). The plots show spatial variations in the proportion of k-mers present in the genomes described on the *x* axis also present in the genome of the different references described in the *y* axis, calculated in sliding windows of 40 bases along the genome of the first isolate and for *k* = 19. Plots are based on k-mer counts computed with Cortex and a custom R visualization script. *pmrCAB* coding regions are highlighted in red. The shaded grey areas highlight the genomic region depicted in Fig. 1B.

Table S1: Genome assembly statistics and inferred location of *pmrCAB* in *A. baumannii* clinical isolates included in the study.



Isolate	Short-reads coverage	Long-reads coverage	Number of contigs[a]	N50[a] (bp)	Number of chromosome-associated contigs[b]	Length of chromosome-associated contigs (bp)	Number of plasmid-associated contigs[b]	Length of plasmid-associated contigs (bp)	Undefined contigs[c]	Genome closed	Inferred localization of *pmrCAB*

67659	116.096	70.689	6	4,174,758	1	4,174,758	3	120,650[d];
								40,936[d];
								16,673[e]
									2	yes	chromosome

72554	107.564	42.227	11	2,988,108	2	2,988,108;
						1,171,702
							3	156,00[d];
								8,970[e];
								7,703[e]
									6	yes	chromosome

71813	125.569	311.889	6	3,902,837	2	3,902,837; 34,129	1	14,782	3	yes	chromosome
67098	121.990	129.754	2	4,073,402	1	4,073,402	1	16,673	0	yes	chromosome
MC1	84.566	278.578	3	4,026,212	1	4,026,212	2	184,748; 8,731	0	yes	chromosome

[a] Based on hybrid assembly using Unicycler.

[b] Based on the gene content identified with genome annotation.

[c] All undefined contigs are <5 kbp.

[d] >90% sequence similarity and coverage with the 184 kbp plasmid from MC1.

[e] >99% sequence similarity and coverage with the 16 kbp plasmid from 67098.

.

3.5 Publication 3: "NanoCore: Core-genome-based bacterial genomic surveillance and outbreak detection in healthcare facilities from Nanopore and Illumina data"

Manuscript submitted to: ASM Journals, mSystems (pending review process)

Impact factor: 7.800 (2023)

DOI: not applicable yet

Authorship: First author

Contribution to the publication according to CRediT classification:

- Conceptualization (supporting)
- Data curation (equal)
- Formal analysis (lead)
- Investigation (lead)
- Methodology (equal)
- Project administration (supporting)
- Software (lead)
- Validation (lead)
- Visualization (lead)
- Writing - original draft preparation (equal)
- Writing - review and editing (lead)

NanoCore: Core-genome-based bacterial genomic surveillance and outbreak detection in healthcare facilities from Nanopore and Illumina data

Sebastian A. Fuchs[1*], Lisanna Hülse[1], Teresa Tamayo[1,2], Susanne Kolbe-Busch[1,3], Klaus Pfeffer[1], Alexander T. Dilthey[1*]

[1] Institute of Medical Microbiology and Hospital Hygiene, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[2] Current address: German Consulting Centre for Infection Prevention and Control, Freiburg, Germany
[3] Current address: Institute of Hygiene, Hospital Epidemiology and Environmental Medicine, University of Leipzig Medical Center, Leipzig, Germany
[*] Corresponding authors: SebastianAlexander.Fuchs@med.uni-duesseldorf.de and dilthey@hhu.de

Abstract

Genomic surveillance can enable the early detection of pathogen transmission in healthcare facilities and contribute to the reduction of substantial patient harms. Fast turnaround times, flexible multiplexing schemes, and low capital requirements make Nanopore sequencing well-suited for genomic surveillance purposes; the analysis of Nanopore sequencing data, however, can be challenging. We present NanoCore, a user-friendly method for Nanopore-based genomic surveillance in healthcare facilities, which enables the calculation and visualization of cgMLST-like sample distances directly from raw Nanopore sequencing reads. NanoCore implements a mapping, variant calling and multi-level filtering strategy and also supports the analysis of Illumina-sequenced isolates. We validated NanoCore on two 24-isolate datasets of methicillin-resistant Staphylococcus aureus (MRSA) and vancomycin-resistant Enterococcus faecium (VRE). In Nanopore-only mode, NanoCore-based pairwise distances between closely related isolates were near-identical to distances calculated with SeqSphere+, a gold-standard commercial method (average differences of 0.75 alleles for MRSA and of 0.81 alleles for VRE), and gave an identical clustering into closely related and non-closely-related isolates. In "hybrid" mode, in which only Nanopore data is used for some isolates and only Illumina data for others, increased average pairwise isolate distance differences were observed (average differences of 3.44 and 1.95 for MRSA and VRE, respectively), while clustering results remained identical. NanoCore is computationally efficient (<15 hours of wall time for the analysis of a 24-isolate dataset on a modern workstation), available as free software, and supports user-friendly installation via bioconda. In conclusion, NanoCore enables the effective use of the Nanopore technology for bacterial pathogen surveillance in healthcare facilities.

Importance

Genomic surveillance involves sequencing the genomes and measuring the relatedness of bacteria from different patients or locations in the same healthcare facility, enabling an improved understanding of pathogen transmission pathways and the detection of "silent" outbreaks that would otherwise go undetected. It has become an indispensable tool for the detection and prevention of healthcare-associated infections and is routinely applied by many healthcare institutions. The earlier an outbreak or transmission chain is detected, the better; in this context, the Oxford Nanopore sequencing technology has important potential advantages over traditionally used short-read sequencing technologies, because it supports "real-time" data generation and the cost-effective "on demand" sequencing of small numbers of bacterial isolates. The analysis of Nanopore sequencing data, however, can be challenging. We present NanoCore, a user-friendly software for genomic surveillance that works directly based on raw Nanopore sequencing reads and demonstrate that its accuracy is equivalent to traditional gold-standard short-read-based analyses.

Keywords: microbial genomics, bacterial outbreak analysis, Nanopore sequencing, hybrid approaches, MLST, healthcare pathogen surveillance

Introduction

Genomic pathogen surveillance has become an essential tool for the detection, characterization and prevention of healthcare-associated infections (1,2) and for improved infection control (3–5). Genomic surveillance can be applied retrospectively to investigate epidemiologically indicated potential outbreaks or prospectively as part of "sequence first" regimes (6), involving the routine sequencing of indicator organisms of nosocomial importance (i.e. those spreading quickly, exhibiting multidrug-resistance and/or virulence factors) and enabling the detection of cryptic transmissions and silent outbreaks. Key factors for the successful implementation of genomic surveillance include linking epidemiological data to genomic analyses, the speed at which sequencing data is generated and analyzed (7,8), and the accuracy of calculated genetic distances between samples.

While most sequencing for genomic pathogen surveillance purposes in healthcare facilities has traditionally relied on the Illumina technology (9), the Oxford Nanopore technology (10) has become increasingly attractive. Advantages of Nanopore sequencing include rapid turnaround times, "real-time" data generation and output, the ability to sequence long fragments of DNA, and low capital costs; for healthcare facility pathogen surveillance, these may translate into reduced outbreak investigation times or the ability to implement genomic surveillance in resource-limited settings. In addition, throughput and error rates, previously limitations of the Nanopore technology (11,12), have improved rapidly (13,14), and Nanopore sequencing is widely used for the assembly of bacterial genomes (13,15). During the COVID-19 pandemic, tens of thousands of viral genomes were sequenced with the Nanopore technology (16–18), demonstrating the potential of the technology for large-scale surveillance.

Challenges for the introduction of Nanopore sequencing in the healthcare pathogen surveillance context, however, include i) the sensitivity of important established bacterial strain typing methods, such as MLST (19,20), cgMLST, or cgSNP (21) to sequencing errors, which may, despite recent progress, remain a concern for Nanopore sequencing data; and ii) the potential requirement that newly generated isolate sequencing data should remain comparable to that of existing, typically Illumina-based, isolate sequencing data, for example to enable the detection of low-intensity unrecognized outbreaks that may span several years.

Multiple studies on the use of Nanopore sequencing for the determination of bacterial sequence types and bacterial genomic epidemiology have shown encouraging results (22,23). Larger-scale studies include Oh et al. (24), who reported mostly consistent, but non-identical, results between Nanopore- and Illumina-based analyses of 23 isolates of vancomycin-resistant *Enterococcus* (VRE); Hall et al. (25), who reported largely consistent results between Nanopore and Illumina for *Mycobacterium tuberculosis*; Liou et al. (26) and Liao et al. (27), who presented a Nanopore-based MLST typing approach for *Staphylococcus aureus*; Ferreira et al. (28), who demonstrated Nanopore-based sequence typing and phylogenetic analysis of methicillin-resistant *Staphylococcus aureus* (MRSA), obtaining results generally consistent with an Illumina-based analysis; and a number of studies on the successful application of Nanopore sequencing to sequence typing in *Salmonella* (29–32). Xian et al. (29), in particular, presented a homopolymer error reduction approach and explicitly considered the case of combining Illumina and Nanopore data in the same analysis. These results are complemented by a number of smaller-scale studies: Linde et al. (33) found consistent results between Illumina and

Nanopore sequencing for two out of three evaluated species of highly pathogenic bacteria, represented by two isolates each; Greig et al. (34) compared the two technologies on two isolates of *Escherichia coli* and found largely concordant results; Tarumoto et al. (35) found that Illumina- and Nanopore-based sequence of two VRE isolates produced concordant results; Both et al. (36) applied Nanopore sequencing to improve the resolution of hospital VRE isolates; and Duc Cao et al. (37) reported successful strain typing for three *Klebsiella pneumoniae* isolates.

With the exception of nanoMLST (26,27), however, no tools have been presented for the user-friendly, integrated analysis of putative bacterial outbreaks directly from raw Oxford Nanopore sequencing reads. NanoMLST was designed for the analysis of multiplex PCR data and implements a classical 7-gene MLST scheme, the resolution of which is often not sufficient for the fine-scale analysis of bacterial transmission chains (38). In addition, the important "hybrid" use case, in which only Nanopore data is used for some isolates and only Illumina data for others, and which enables, for example, the fast investigation of urgent cases with Nanopore sequencing against a background of Illumina-sequenced other isolates, was only considered in Hall et al. (25) and Xian et al. (29).

Here, we present NanoCore, a user-friendly tool developed specifically to enable the effective use of the Oxford Nanopore technology for the genomic surveillance of bacteria and outbreak detection in healthcare facilities. NanoCore works directly based on raw (i.e. unassembled) Nanopore sequencing reads, while also supporting the analysis of Illumina-sequenced isolates. We demonstrate the accuracy of NanoCore on two datasets of MRSA and VRE, comprising two species that are highly relevant in the hospital infection control and genomic epidemiology context (39,40) and which exhibit a medium (MRSA) as well as high degree of genome plasticity (VRE) (41,42). For validation, we compared NanoCore against Illumina-based analyses of the same samples with Ridom SeqSphere+ (43), a commercial "gold standard" software used by many hospital hygiene and infection control departments.

Results

Overview of NanoCore

NanoCore enables the investigation of putative bacterial outbreaks from Nanopore sequencing data, while also supporting the integrated analysis of Illumina-sequenced isolates (Figure 1).

In NanoCore, input reads are mapped to a species-specific core genome reference, followed by variant calling, the calculation of pairwise isolate distances, and the visualization of the analyzed sample using a Minimum Spanning Tree (MST). The robust computation of isolate distances from Nanopore data alone as well as in "hybrid" analysis mode is enabled by a tailored multi-level filtering strategy, accounting for e.g. copy number variation in the utilized core genome reference in individual isolates.

The pairwise isolate distance metric employed by NanoCore is similar, but not identical, to cgMLST: Isolate distances in NanoCore are based on the number of species-specific core genome genes for which a difference in allelic state can confidently be asserted; however, no attempt is made to assign a fixed allele identifier to each analyzed gene in each isolate.

NanoCore, which is implemented in R and Perl, is freely available from bioconda (44) or GitHub.

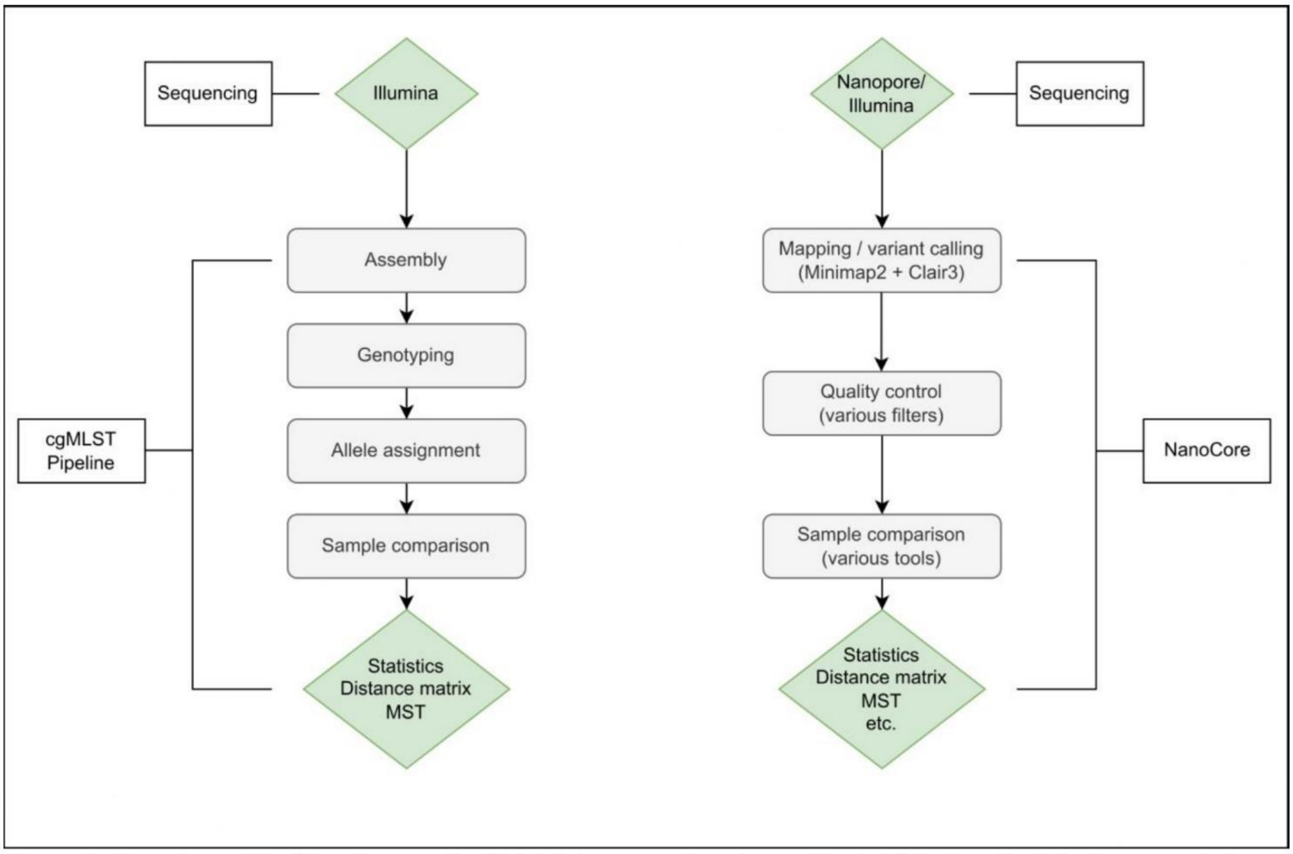


Figure 1: Overview of the NanoCore method (right), in comparison to a well-established method for the computation of cgMLST-based distances (SeqSphere+ (43), left).

Validation experiment 1: *S. aureus* in Nanopore-only mode

In the first experiment, we benchmarked the Nanopore-only analysis mode of NanoCore on MRSA, representing a species of key relevance in the hospital outbreak context. Briefly, we assembled a 24-isolate benchmark dataset from the biobank of University Hospital Düsseldorf's Institute of Medical Microbiology and Hospital Hygiene, consisting of isolates collected between April 2017 and February 2022 and comprising three clusters of closely related isolates as determined by cgMLST analysis before. Per-sample Nanopore sequencing data were generated in a single multiplexed MinION R10 flow cell run (see Methods) and coverages ranged from 74x to 246x with an average of 120x (Supplementary Figure 1). NanoCore was benchmarked against an Illumina-based analysis of the same isolates with SeqSphere+, with per-sample coverages ranging from 33x to 187x (average: 101x).

Pairwise isolate distances computed by NanoCore (Supplementary Table 1) were based on an average number of 1856 compared genes per isolate pair, out of 1864 genes present in the utilized *S. aureus* core genome dataset (45). The gene-level filters affecting the largest number of genes were the "coverage and mapping quality" and "low coverage" filters, leading to the exclusion of 66 and 29 genes over all isolates, respectively (see Supplementary Table 2 and Supplementary Figure 2). Furthermore, 629 genomic positions were globally excluded from all pairwise distance calculations (most often due to the global positional heterozygosity filter; Supplementary Table 3), and an additional 1537 genomic positions were removed from individual pairwise comparisons (identified by the individual-variant filter; Supplementary Table 4). By comparison, SeqSphere-computed distances were based on an average number of 1832 analyzed genes per isolate and on an average number of 1799 analyzed genes per isolate pair (Supplementary Table 5 and Supplementary Table 6).

NanoCore-computed pairwise distances (Supplementary Table 1) were highly concordant with SeqSphere+

(Supplementary Table 5; Pearson's r = 1.000); for 47 out of 276 isolate pairs, the computed pairwise distances were identical. For the 19 pairs of closely related isolates with SeqSphere+ distances of ≤ 15 (i.e. covering the important use case of identifying pairs of isolates potentially related due to an infection chain context), NanoCore -computed pairwise distances were identical in 4 cases, and the average difference in pairwise distances was 0.75 (Figure 2 panel A).

We carried out an in-depth investigation of the observed differences between NanoCore and SeqSphere+ in the set of 19 pairs of closely related isolates with SeqSphere+ distances ≤ 15. First, we focused, across all included isolate pairs, on the 34,729 instances of pairwise gene comparisons present in both the NanoCore and SeqSphere+ analyses; of those, NanoCore and SeqSphere+ disagreed on only 23 instances (Figure 2 panel B). Manual investigation showed that the SeqSphere+ calls were likely correct in 8 of these 23 cases; 5 cases were classified as SeqSphere+ false-positive calls; and 10 cases remained ambiguous. The 8 false-negative calls by NanoCore were exclusively due to the positions of the missed variants being close to the 5´ or 3´ends of a gene (Supplementary Table 7)., However, the detection of such variants is a known issue with the Clair3 variant caller used within NanoCore (GitHub issue: https://github.com/HKU-BAL/Clair3/issues/135). Next, we investigated the 15 out of 19 closely related isolate pairs for which a difference between the NanoCore- and SeqSphere+-computed distances was observed, independent of whether the gene pairs responsible for the observed differences were analyzed by both NanoCore and SeqSphere. In 6 cases, the observed differences in pairwise isolate distances could be attributed to a failure to detect true-positive allelic differences by NanoCore (usually driven by false-negative calls of variants close to the 5´ or 3´end of a gene); in 3 cases to likely false-positive variant calls by SeqSphere+; and in 6 instances the manual investigation showed that the distances calculated by neither approach were likely fully correct (see Supplementary Table 8 for a full list of investigated pairwise differences).

Finally, clustering the isolates using a genetic distance threshold of 10 (consistent with recommendations by Schürch et al. (21)) produced the same sets of related isolates for NanoCore and SeqSphere^{+} (Figure 2 panel C), further demonstrating the high degree of concordance between NanoCore and SeqSphere+.

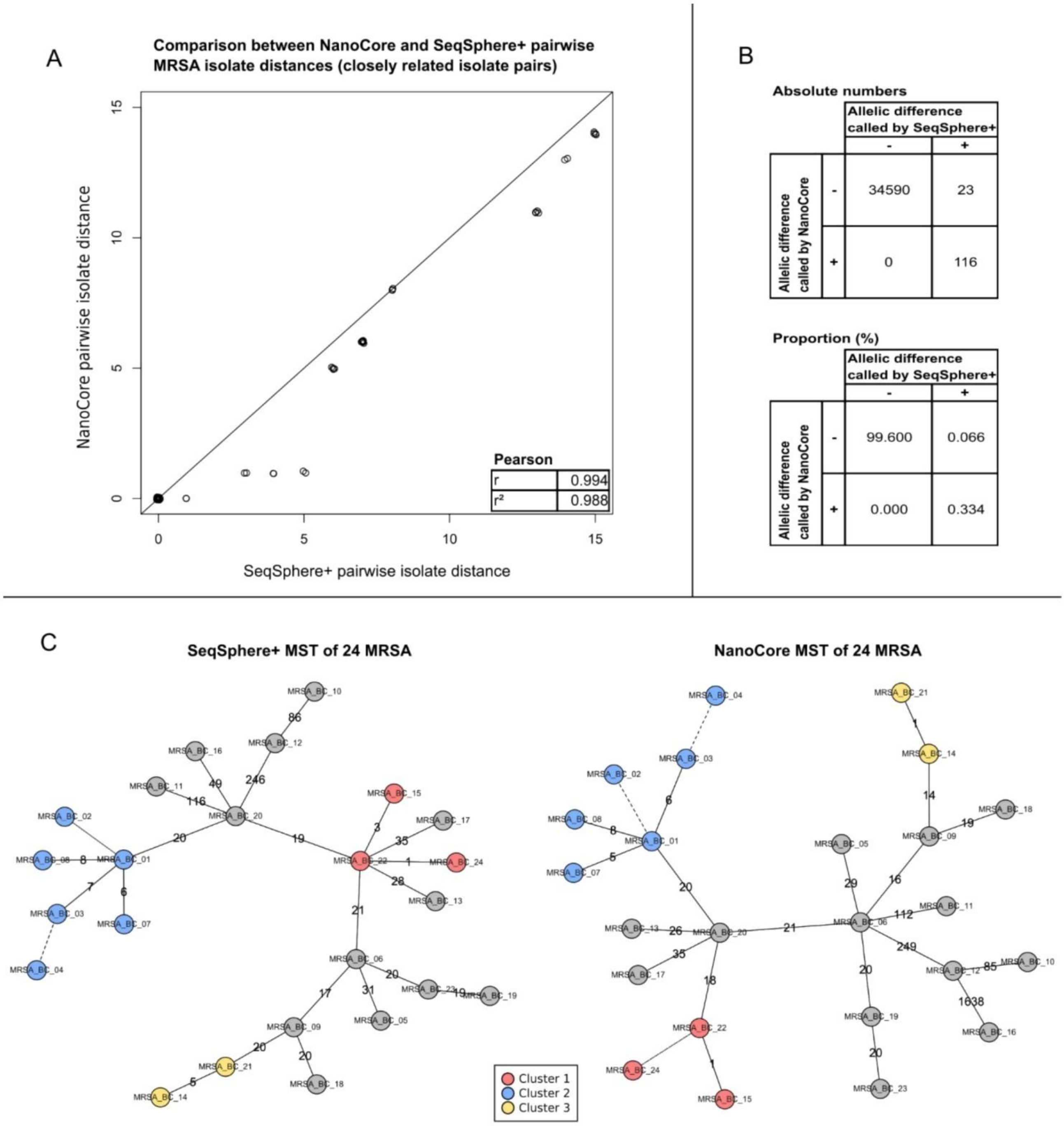


Figure 2: Analysis of 24 MRSA isolates. A. Comparison of NanoCore- and SeqSphere+-based pairwise isolate distances for pairs of closely related isolates (SeqSphere+ distance ≤ 15), with Pearson correlation shown in the inset. B. Comparison of individual-gene NanoCore- and SeqSphere+ results across closely related isolate pairs (SeqSphere+ distance ≤ 15). Shown are results from genes that were analyzed by both NanoCore and SeqSphere+. C. Minimum spanning trees (MSTs) of the analyzed isolates based on SeqSphere+ (left) and NanoCore (right); clusters of closely related isolates, computed independently based on the output of SeqSphere+ and NanoCore, are shown as red, blue and yellow circles.

Validation experiment 2: *E. faecium* in Nanopore-only mode

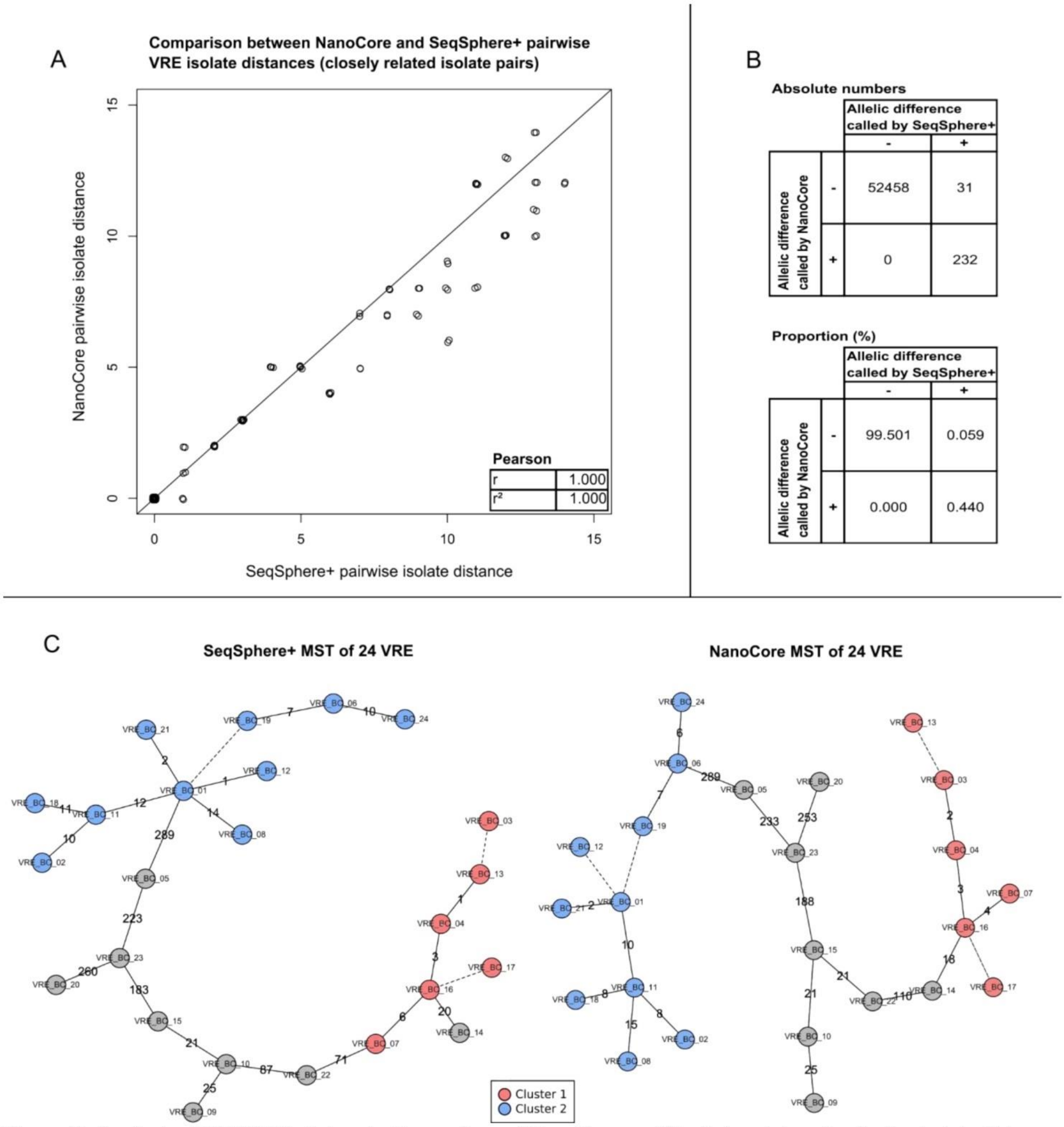
In the second experiment, we benchmarked the Nanopore-only mode of NanoCore on VRE, which may, due to a higher degree of genome plasticity, represent a challenge for the variant calling and filtering strategies employed by NanoCore. The selected 24 VRE isolates were taken from the biobank of University Hospital Düsseldorf's Institute of Medical Microbiology and Hospital Hygiene, comprising two clusters of closely related isolates, and were collected between August and October 2021. Nanopore sequencing data were generated in two multiplexed MinION runs and genome coverages ranged from 96x to 563x (mean: 273x; Supplementary Figure 3), compared to 51x to 108x (mean: 87x) for the Illumina data that were used for the comparative SeqSphere+ analysis.

In the case of VRE, we observed an increased number of genes removed by NanoCore's default filters; pairwise distances (Supplementary Table 9) were based on an average number of 1397 compared genes, out of 1423 genes present in the core genome (46). Consistent with an assumed effect of genome plasticity, the filter affecting the highest number of genes was the gene-level "heterozygosity" filter (607 genes removed in individual isolates; Supplementary Table 2 and Supplementary Figure 4), which is sensitive to variations in genome structure. Furthermore, 877 genomic positions were globally excluded from all pairwise distance calculations (most often due to the global positional "heterozygosity" filter; Supplementary Table 3) and 468 genomic positions were removed from individual pairwise comparisons (identified by the individual-variant filter; Supplementary Table 4). By comparison, SeqSphere+-computed distances were based on an average number of 1404 analyzed genes per isolate and on an average number of 1385 analyzed genes per isolate pair (Supplementary Table 10 and Supplementary Table 11).

As was the case for MRSA, the NanoCore-computed pairwise distances for VRE (Supplementary Table 9) exhibited a high degree of concordance with SeqSphere+ (Supplementary Table 10; Pearson's r = 0.998 for all isolate pairs); for the 39 pairs of closely related isolates (SeqSphere+ distances ≤ 15), the degree of concordance for computed distances was higher (r = 1.000) and exhibited an average difference of 0.81 (Figure 3 panel A).

Furthermore, within the set of pairwise gene comparisons conducted by both SeqSphere+ and NanoCore in the set of closely related isolates, the two methods disagreed on only 31 out of 55,497 instances of pairwise gene comparisons (Figure 3 panel B), driven by differences in allelic state called by SeqSphere+. Manual investigation showed that SeqSphere+ was likely correct in 24 of these 31 cases; 5 cases were classified as SeqSphere false-positive calls and 2 cases remained ambiguous. False-negative calls by NanoCore were either due to low coverage (9 cases) or the positions of the missed variants being close to the 5′ or 3′ end of a gene (15 cases; Supplementary Table 7). Finally, manual adjudication of the 25 out of 39 closely related isolate pairs for which a difference between the NanoCore- and SeqSphere+-computed distances was observed showed that 11 of these instances were due to false-negative calls by NanoCore (typically driven by exclusion of the variant-containing genes by the gene-level heterozygosity filter); 4 were due to likely false-positive calls by SeqSphere; and in 10 instances, the manual investigation showed that the distances calculated by neither approach were likely fully correct (Supplementary Table 8).

Finally, isolate clusters computed using a genetic distance threshold of 15 (consistent with recommendations by Schürch et al. (21)) were identical between NanoCore and SeqSphere+ (Figure 3 panel C), demonstrating the high degree of consistency between the two methods.



Absolute numbers

		Allelic difference called by SeqSphere+	
		-	+
Allelic difference called by NanoCore	-	52458	31
	+	0	232

Proportion (%)

		Allelic difference called by SeqSphere+	
		-	+
Allelic difference called by NanoCore	-	99.501	0.059
	+	0.000	0.440

Figure 3: Analysis of 24 VRE isolates. A. Comparison of NanoCore- and SeqSphere+-based pairwise isolate distances for pairs of closely related isolates (SeqSphere+ distance ≤ 15), with Pearson correlation shown in the inset. B. Comparison of NanoCore- and SeqSphere+-based results on the level of individual genes across closely related isolates (SeqSphere+ distance ≤ 15). Shown are results from genes that were analyzed by both NanoCore and SeqSphere+. C. Minimum spanning trees of the analyzed isolates based on SeqSphere+ (left) and NanoCore (right); clusters of closely related isolates, computed independently from the output of SeqSphere+ and NanoCore, are shown as red and blue circles.

Validation experiment 3: Evaluation of the "hybrid" mode of NanoCore on MRSA and VRE

To evaluate the "hybrid" analysis mode of NanoCore, we first assembled synthetic hybrid MRSA and VRE datasets for benchmarking purposes based on the sequencing data analyzed in the first two experiments; to assemble the hybrid datasets, the Nanopore and Illumina data from each isolate were not combined but treated as if they emanated from biologically different isolates, yielding two MRSA and VRE datasets with 48 isolates each.

We first evaluated the impact of NanoCore's multi-level filtering strategy. For the "Nanopore" component of the hybrid datasets, gene-level filtering, which is applied to each isolate independently, produced the same results as in the first two experiments; for the "Illumina" component, gene-level filtering led to the exclusion of a median number of 374 and 79 genes for MRSA and VRE, respectively (Supplementary Table 12) and Supplementary Table 13). The filters leading to the largest numbers of genes for the "Illumina" component were the gene-level "low coverage" filter (approximately 4000 genes over all isolates in both datasets, Supplementary Table 2) and the gene-level "coverage and mapping quality" filter, which had a particularly large effect in the MRSA dataset (almost 6500 genes over all isolates in both datasets, Supplementary Table 2); correlations between the different filters are visualized in Supplementary Figure 5 and Supplementary Figure 6. Furthermore, 11442 and 3048 genomic positions were excluded by global positional filters for MRSA and VRE, respectively (Supplementary Table 3), as well as 3533 (MRSA) and 1272 (VRE) positions from individual pairwise isolate distance calculations (identified by the individual-variant filter; Supplementary Table 4).

Within the two benchmarking datasets, we compared, for each pair of biological isolates (276 pairs in total per species), hybrid with single-technology pairwise isolate distances (Supplementary Table 14 and Supplementary Table 15). Specifically, for two isolates X and Y, we compared $\text{distance}_{\text{NanoCore}}(X_{\text{Nanopore}},Y_{\text{Illumina}})$ and $\text{distance}_{\text{NanoCore}}(X_{\text{Illumina}},Y_{\text{Nanopore}})$ (the "hybrid" distances) with $\text{distance}_{\text{NanoCore}}(X_{\text{Nanopore}},Y_{\text{Nanopore}})$ and $\text{distance}_{\text{SeqSphere}}(X_{\text{Illumina}},Y_{\text{Illumina}})$ (the "single-technology" distances); the first subscript indicates the utilized pairwise distance computation method and the subscripts of X and Y indicate the sequencing technology data type. We found that hybrid isolate distances were generally highly concordant with single-technology isolate distances; specifically, over 552 evaluated hybrid distances for each species, hybrid distances and the Nanopore-based distances exhibited a correlation (Pearson's r) of 1.000 (MRSA) and 0.985 (VRE); hybrid distances and Illumina-based SeqSphere+ distances exhibited a correlation of 0.981 (MRSA) and 0.985 (VRE). When considering only pairs of closely related isolates (SeqSphere+ distance ≤ 15), we observed an average difference between NanoCore- and SeqSphere+-based distances of 3.44 and a correlation of 0.866 for MRSA (19 isolate pairs; Figure 4 panel A and Supplementary Table 14), and an average difference of 1.95 and a correlation of 0.997 for VRE (39 isolate pairs; Figure 4 panel C and Supplementary Table 15).

Next, to investigate the accuracy of isolate clustering in "hybrid" mode, we created three "hybrid" scenarios for both MRSA and VRE, which all comprised the full set of 24 biological isolates of the corresponding species, and in which only Nanopore data was used for one randomly assigned half of the biological isolates and only Illumina data for the other half. Each "hybrid" scenario was analyzed as an independent NanoCore run, and clustering was carried out using the same distance thresholds as in the first two validation experiments. Within each "hybrid" scenario and for both species, we found perfect agreement between the computed clusters and the single-technology clustering results from the first two experiments (Figure 4 panels B and D).

To further characterize potential error modes of the "hybrid" analysis mode of NanoCore, we carried out an in-depth analysis of the first "hybrid" scenario for each species, focusing on the 18 out of 19 (MRSA) and on the 30 out of 39 (VRE) closely related isolate pairs (SeqSphere+ distance ≤ 15) for which a difference between SeqSphere+ and NanoCore (in "hybrid" mode) distances was observed within the respective first "hybrid" scenario. For MRSA, 12 of the 18 differences were accounted for by "hybrid" distances; for VRE, 22 of 30. Further manual investigation showed that 5 of the 18 differences for MRSA were due to false-positive or false-negative calls by NanoCore, 2 were errors by SeqSphere+, and in 11 cases neither distance was likely fully correct (Supplementary Table 7). For VRE, 11 of the observed 30 discrepancies were likely driven by false-positive or false-negative calls by NanoCore; 2, by errors by SeqSphere+; and in 17 cases neither distance was likely fully correct. Across both species, false calls by NanoCore were often due to exclusion of the variant-containing genes by the gene-level "low coverage" filter, due to the corresponding variants being close to the 5′ or 3′ borders of a gene, or due to variant calling artifacts in low-coverage regions that were not removed by any of the coverage-related filters (Supplementary Table 7 and Supplementary Table 8).

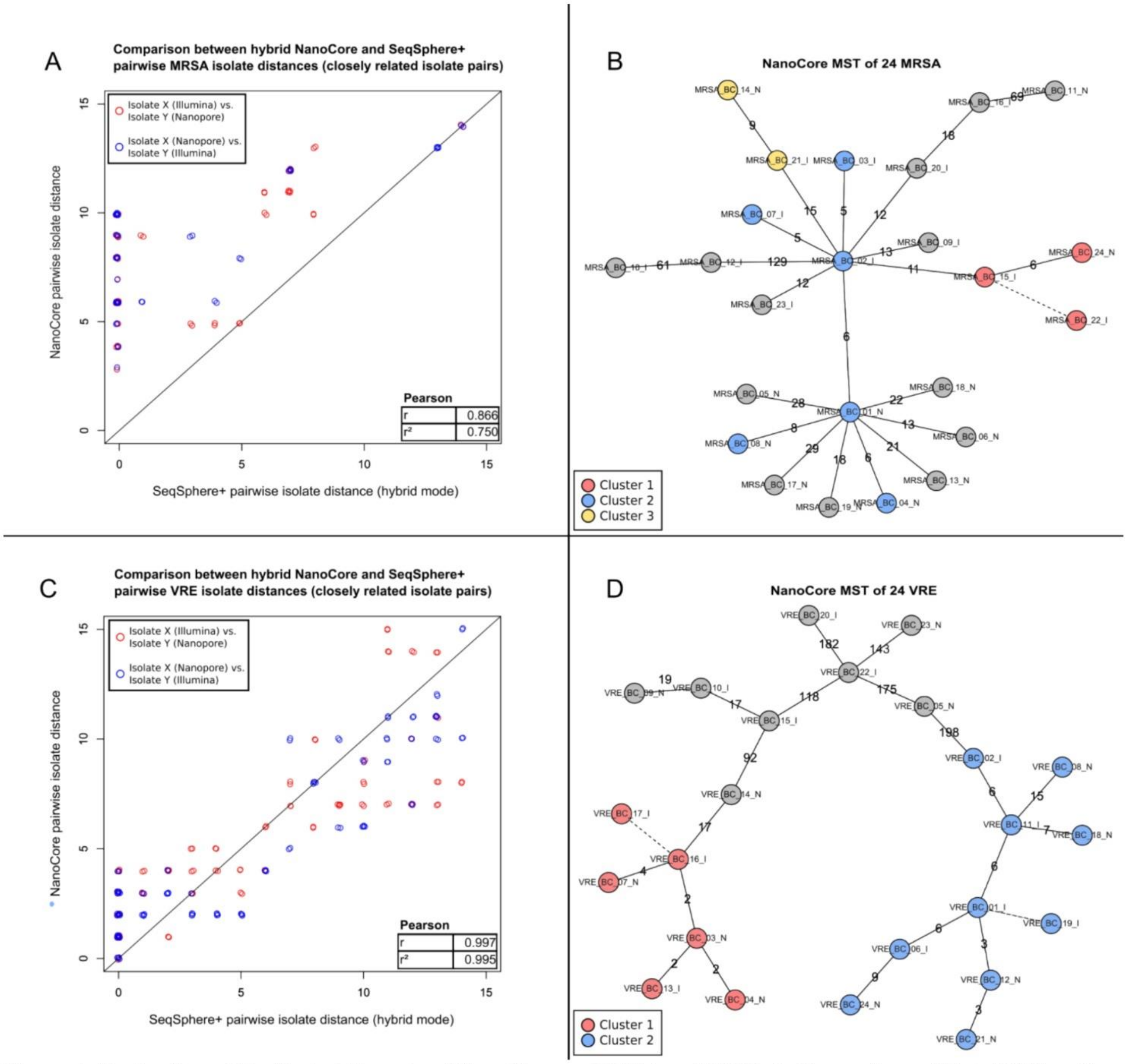


Figure 4: Evaluation of the "hybrid" mode of NanoCore on MRSA and VRE. A. Comparison of "hybrid" NanoCore and SeqSphere+ pairwise isolate distances for pairs of closely related MRSA isolates (SeqSphere+ distance ≤ 15), with Pearson correlation shown in the inset. B. NanoCore "hybrid" mode minimum spanning tree of the analyzed MRSA isolates, based on the first "hybrid" MRSA scenario (comprising 12 isolates for which only Nanopore data was used and 12 isolates for which only Illumina data was used; see Results); clusters of closely related isolates are shown as red, blue and yellow circles. C. Comparison of "hybrid" NanoCore- and SeqSphere+-based pairwise isolate distances for pairs of closely related VRE isolates (SeqSphere+ distance ≤ 15), with Pearson correlation shown in the inset. D. NanoCore "hybrid" mode minimum spanning tree of the analyzed VRE isolates, based on the first "hybrid" VRE scenario (comprising 12 isolates for which only Nanopore data was used and 12 isolates for which only Illumina data was used; see Results); clusters of closely related isolates are shown as red and blue circles.

Computational performance

Analysis of the 24-isolate datasets described above with NanoCore (8 threads) took <15 hours of wall time and <5 Gb of RAM per experiment on an AMD Ryzen Threadripper 3970X system with 3.7Ghz. Detailed runtime and computational requirements statistics are reported in Supplementary Table 16.

Discussion

We have presented NanoCore, a user-friendly method for Nanopore-based genomic surveillance of bacteria and outbreak detection in healthcare facilities. NanoCore does not require any preprocessing of the Nanopore read data, accepting raw sequencing reads as input. In addition to Nanopore sequencing, NanoCore also supports the analysis of Illumina-sequenced isolates. Important use cases of this include the selective application of Nanopore sequencing to urgent cases, leveraging the technology's rapid data generation capabilities, as well as the complete transition of a hospital's surveillance platform from Illumina to Nanopore sequencing without having to exclude or re-sequence older isolates for which only Illumina data are available.

We validated NanoCore on two independent 24-isolate datasets of MRSA and VRE, species highly relevant to the field of hospital hygiene and infection control. The validation experiments demonstrated identical clustering results between NanoCore in both evaluated modes (Nanopore-only and "hybrid") and SeqSphere+, a commercial gold standard method, for both species. Pairwise isolate distances for closely related isolates based on NanoCore in Nanopore-only mode were near-identical to those of SeqSphere+ (average differences of 0.75 for MRSA and of 0.81 for VRE); for NanoCore in "hybrid" mode, the average difference in pairwise isolate distances between NanoCore and SeqSphere+ was found to be increased (average differences of 3.44 and 1.95 for MRSA and VRE, respectively), but remained at a low level. Hospital outbreak investigations typically focus on distinguishing between related and non-related isolates and on the fine-scale structure of relatedness within the set of related isolates. By contrast, the determination of accurate pairwise isolate distances for more distantly related isolates can be relevant in the context of phylogenetics, but typically not in the context of outbreak investigations. The validation experiments thus demonstrated the near-equivalence between NanoCore and SeqSphere+ for the use case of bacterial genomic surveillance and outbreak detection in healthcare facilities.

NanoCore employs a multi-level filtering strategy to heuristically reduce the potential impact of false variant calls on computed pairwise sample distances. First, gene-level filters are applied at a per-isolate level to detect read mapping ambiguities as well as duplications or deletions of individual genes, which are associated with variant calling artifacts and which were occasionally observed in the analyzed isolates (Supplementary Figure 7), the classification of the analyzed genes as "core" notwithstanding. Consistent with the higher genomic plasticity of *E. faecium,* gene-level filters and the "heterozygosity" filter in particular had a substantially larger effect in VRE than in MRSA (Supplementary Table 2). Second, positional filters capture technical artifacts of the variant calling process and base contexts that pose challenges for Nanopore-based variant calling, as well as drops in coverage. Positional filtering is implemented in a way that initially identifies potentially problematic positions on a per-sample basis, which are subsequently propagated across the complete dataset (i.e., excluded from all distance calculations); this is based on the rationale that the properties that render individual positions challenging are typically shared between isolates, even if the heuristics employed to detect these positions are not activated in every individual isolate. Last, individual isolate-distinguishing variant calls are filtered based on the allele frequency of the called variant in the involved isolates; this step reduces the impact of false-negative variant calls. Because of the increased rate of homopolymer errors in Nanopore sequencing, INDEL calls are generally ignored by NanoCore; of note, Xian et al. similarly proposed a heuristic approach for homopolymer correction (29).

Our in-depth investigation of differences between NanoCore and SeqSphere+ for pairs of closely related isolate pairs showed that these were almost exclusively driven by false-negatives (i.e., NanoCore failing to detect a true isolate-distinguishing variant), which were often caused by a known variant calling issue of the Clair3 variant caller in the case of MRSA, and often related to the gene-level "heterozygosity" filter in the case of VRE. Improvements to the Clair3 variant caller, or integration of another variant calling algorithm, may reduce these errors in the future. In "hybrid" mode, we also observed false-positive calls by NanoCore (i.e., NanoCore erroneously calling an isolate-distinguishing variant that is not really present); these could be addressed by integration of an Illumina-optimized variant calling approach (47) in future releases of NanoCore. In addition, the filtering strategy of NanoCore could be optimized for short-read data, for example with respect to the increased coverage fluctuations (Supplementary Figure 6,) and lower mapping qualities (Supplementary Figure 6) observed in short-read data; such potential for optimization was particularly apparent for the MRSA dataset, in which increased coverage fluctuations in the short-read data led to the exclusion of a comparably high number of genes (Supplementary Table 2), contributing to increased discrepancies between "hybrid" NanoCore and SeqSphere+ for this species. Importantly, while most observed differences between SeqSphere+ and NanoCore were due to NanoCore, we also observed false-positive calls by SeqSphere+ in all experiments.

NanoCore has a number of limitations. First, NanoCore requires a core genome reference; while these are available (https://www.cgmlst.org/ncs) for the large majority of clinically important species, there are still microbial species for which a core genome dataset has not been defined yet. Second, by design, NanoCore will only detect isolate-distinguishing variants in the core genome; in some instances, whole-genome based approaches also accounting for extrachromosomal genome information (i.e., from plasmids) may offer increased resolution for the fine-scale analysis of otherwise closely related isolates (38). Third, NanoCore does not assign a standardized allele identifier to the analyzed genes; NanoCore does thus not enable the comparison of isolates based on allele identifiers alone (48), which can be important e.g. in the context of inter-institutional outbreak investigations in which the sharing of raw sequencing data is not possible. Fourth, in the current implementation, NanoCore may not scale to the analysis of very large datasets; in future releases, this could be addressed by limiting the computation of full pairwise distances to closely related isolates while relying on an approximate distance metric, e.g. based on Mash (49), otherwise. Fifth, NanoCore does not support the analysis of isolates based on *de novo* assembly. While limiting, as discussed above, the resolution of NanoCore to the core genome, the advantage of this approach is that NanoCore can also be applied to lower-coverage datasets. For example, we obtained virtually identical results for the MRSA dataset after downsampling the Nanopore input data to 50% of its original size (data not shown); in addition to demonstrating robustness, this result indicates that NanoCore may also support Nanopore multiplexing schemes with more than 24 isolates per flow cell.

Conclusion

NanoCore is a user-friendly method for genomic surveillance and outbreak detection in healthcare facilities based on the Oxford Nanopore sequencing technology. In two independent validation experiments based on MRSA and VRE, we demonstrated consistency between NanoCore and SeqSphere+, a gold-standard commercial method. NanoCore also supports the analysis of Illumina-sequenced samples. In conclusion, NanoCore enables the effective use of the Nanopore technology for bacterial pathogen surveillance in healthcare facilities, the potential advantages of which include low capital costs and reduced sample-to-result turnaround times.

Material and Methods

Analyzed bacterial isolates and core genome references

24 VRE and 24 MRSA isolates were selected from the isolate collection of the Institute of Medical Microbiology and Hospital Hygiene of Düsseldorf University Hospital. All isolates had been previously sequenced with Illumina and analyzed with SeqSphere+ as part of the Institute's routine surveillance activities; the analyzed isolates were selected to represent different degrees of genetic relatedness (see Results). For the generation of the Nanopore data, DNA was obtained from cryostocks of the selected isolates that were thawed and re-cultured.

For the analysis of these samples with NanoCore, we selected well-established core genome references for *Staphylococcus aureus*, comprising 1864 core genes and 1.70 Mbp of sequence (45), as well as for *Enterococcus faecium*, comprising 1423 core genes and 1.35 Mbp of sequence (46) .

Bacterial culture and DNA extraction

Bacterial isolates were cultured employing routine overnight LB culture protocols at 37°C. DNA was extracted using the Qiagen DNeasy UltraClean Microbial Kit according to the manufacturer's instructions. DNA concentrations and quality were checked with NanoDrop and 100ng of DNA were diluted to fit the desired concentration of 5ng/µl.

Nanopore sequencing and demultiplexing

Nanopore sequencing was carried out on the Oxford Nanopore MinION device. DNA concentrations were measured using Qubit. Sequencing libraries for MRSA were prepared using the Oxford Nanopore ligation sequencing gDNA native barcoding kit SQK-NBD112-24 and sequenced on "FLO-MIN112" R10 flow cell, multiplexing 24 isolates per flow cell. Sequencing data for VRE were generated in two separate MinION runs, multiplexing 13 and 11 isolates per flow cell, based on the SQK-NBD112-24 kit with a "FLO-MIN112" R10 flow cell and based on the SQK-NBD114-24 kit with a "FLO-MIN114" R10.4 flow cell, respectively. Reads were basecalled and demultiplexed using Guppy (version 6.1.5). Per-isolate sequencing data statistics are shown in Supplementary Table 17.

Illumina sequencing and demultiplexing

Illumina sequencing data were generated for routine surveillance purposes and over multiple sequencing runs. DNA quality control was carried out using the Fragment Analyzer and NanoDrop instruments. Sequencing libraries were prepared using the Illumina Nextera XT DNA Library Preparation Kit "FC-131-1096" for 96 samples. Post-library-prep QC was carried out using the Fragment Analyzer and NanoDrop instruments as well as using Fluorometric Assay for concentration checks. Samples were prepared by equimolar pooling (including additional quality control) and sequenced with the MiSeq v2 500 cycle kit (251 - 8 - 8 – 251). Post-sequencing processing, quality control and demultiplexing were carried out on the instrument. Per-isolate sequencing data statistics are shown in Supplementary Table 17.

NanoCore

NanoCore is based on the following key steps: (i) For each isolate, mapping of the generated sequencing reads using minimap2 (50) to a species-specific core genome reference (using flags „-x map-ont" or "-x sr" depending on the type of sequencing reads); (ii) for each isolate, detection of variants in core genome genes using the Clair3 variant caller (51) (with flags „--include_all_ctgs" and „-m /path/to/model" set according to the type of input data); (iii) computation of

pairwise sample distances (see below); (iv) generation of a minimum spanning tree (MST), visualizing of the genetic structure of analyzed isolates, and of various results and quality control tables.
NanoCore is implemented in Perl; the MST step is implemented in R (52). BAM files are manipulated using samtools (53). NanoCore is available under the MIT license and can be installed via bioconda.

Input sequencing data are specified using a simple sample sheet in tab-separated format; in addition, the user specifies a species-specific core genome reference file. Reference files for 8 bacterial species (Supplementary Table 18) are included in the NanoCore package. In addition, the user may specify a minimum coverage threshold (default 20) and the number of threads used for components of the pipeline that support multithreading.

The genetic distance between two isolates in NanoCore is computed based on the number of genes that confidently, i.e., after application of gene-level, positional and individual-variant filters (see below), differ in allelic state. Formally, for a pair of isolates X and Y, the set of candidate pair-distinguishing variants is defined as the set of non-shared variant calls from the Clair3-generated VCF files for X and Y, where a candidate variant is defined by its location (gene and position) and the called variant allele. The set of candidate variants is filtered by (i) removing all INDEL variants; (ii) removing all variants located in genes flagged by gene-level filters as suspicious in isolates X or Y; (iii) removing all variants at positions flagged by global positional filters; and (iv) removing all variants flagged by the individual-variant filter. The genetic distance between X and Y is then defined as the number of core genome genes for which one or more variants remain in the set of candidate variant pairs post-filtering. We note that the NanoCore approach to computing genetic distances is similar, but not identical, to cgMLST, as no attempt is made by NanoCore to explicitly determine and label with an allele identifier the allelic state of individual genes.

Gene-level, positional and individual-variant filters

Gene-level filtering is carried out independently for each isolate by NanoCore; the aim of gene-level filtering is to identify specific genes in individual isolates that exhibit an increased probability of unreliable variant calling results. Gene-level filters comprise i) the gene-level "heterozygosity" filter, which marks genes in which more than 50% of Clair3 variant calls are heterozygous; (ii) the gene-level "coverage and mapping quality" filter, which flags genes that exhibit average per-read mapping qualities of <55 and average coverages that deviate by more than 25% from the average coverage of the isolate (both conditions need to be satisfied for this filter to be activated); and (iii) the gene-level "low coverage" filter, which marks genes in which more than 10% of positions exhibit a coverage below the minimum coverage threshold.

Global positional filters flag individual positions with potentially problematic variant calling results; these are ignored across the entire analyzed dataset. Global positional filters comprise (i) the positional "heterozygosity" filter, which flags positions with a heterozygous call in at least one isolate; (ii) the positional "low allele frequency" filter, which marks variant positions at which the called variant allele has <50% allele frequency in the raw sequencing reads in at least one isolate (determined using the "allele frequency" tag in the VCF produced by the variant caller); (iii) the positional "low quality" filter, which marks all positions at which a Clair3 variant call was annotated with the "LowQual" tag in at least one isolate; and (iv) the positional "low coverage" filter, which flags all positions with a coverage below the specified minimum coverage in at least one isolate.

Last, the individual-variant filter is applied to all candidate variants potentially distinguishing two isolates X and Y remaining after application of the other filters; the aim of the individual-variant filter is to remove false-positive pair-distinguishing variants that arise from false-negative variant calls in either X or Y. Let *a* be the variant allele of the candidate pair-distinguishing variant and assume without loss of generality that *a* was called in X, but not in Y; a variant passes the individual-variant filter if and only if the allele frequency of *a* in the raw reads of Y is less than 20% (determined with the "mpileup" function of samtools).

SeqSphere+ comparison

Illumina sequencing data were analyzed with Ridom SeqSphere+ (43) using default settings for the analyzed species; pairwise genetic isolate distances based on cgMLST and the sets of analyzed genes per isolate were extracted from SeqSphere+ default output using custom scripts. For the presented analyses, the cgMLST-based distance metric of SeqSphere+ was compared to the cgMLST-like distance metric of NanoCore.

Manual adjudication of differences between SeqSphere+ and NanoCore

Manual adjudication of differences between SeqSphere+ and NanoCore was based on visual inspection of the aligned Illumina and/or Nanopore sequencing reads using the Integrative Genomics Viewer (IGV) tool (version 2.11.0)(54).

Clustering of closely related isolates

For a given maximum genetic distance d, clusters of closely related isolates are defined as the connected components of the graph G = (V, E), where V are the analyzed isolates and an edge e connecting two isolates X and Y exists if and only if the pairwise genetic distance between X and Y is ≤ d. For analysis of the VRE isolates, d was set to 15; for the analysis of the MRSA isolates, d was set to 10, in line with recommendations by Schürch et al. (21).

Supplementary Information

Supplementary Files

Supplementary Figure 1.pdf	Experiment 1 (MRSA): Basic statistic plots
Supplementary Figure 2.pdf	Experiment 1 (MRSA): Heatmaps of excluded genes
Supplementary Figure 3.pdf	Experiment 2 (VRE): Basic statistic plots
Supplementary Figure 4.pdf	Experiment 2 (VRE): Heatmaps of excluded genes
Supplementary Figure 5.pdf	Experiment 3 part 1 (MRSA): Heatmaps of excluded genes
Supplementary Figure 6.pdf	Experiment 3 part 1 (VRE): Heatmaps of excluded genes
Supplementary Figure 7.pdf	IGV screenshots of possible gene duplication
Supplementary Tables.xlsx	Different supplementary tables:
- Supplementary Table 1	Experiment 1 (MRSA): Distance matrix
- Supplementary Table 2	Excluded genes per isolate and the reasons for the exclusion
- Supplementary Table 3	Excluded positions and the reasons for the exclusion
- Supplementary Table 4	Excluded positions due to mpileup statistics
- Supplementary Table 5	Experiment 1 (MRSA): Differences between SeqSphere and NanoCore
- Supplementary Table 6	Pairwise listing of important values for the MRSA experiments
- Supplementary Table 7	Condensed list of differing isolate pairs and the reasons for the false calls
- Supplementary Table 8	Differing genes per isolate pair and the reasons for the false calls
- Supplementary Table 9	Experiment 2 (VRE): Distance matrix
- Supplementary Table 10	Experiment 2 (VRE): Differences between SeqSphere and NanoCore
- Supplementary Table 11	Pairwise listing of important values for the VRE experiments
- Supplementary Table 12	Experiment 3 part 1 (MRSA): Differences between SeqSphere and NanoCore
- Supplementary Table 13	Experiment 3 part 1 (VRE): Differences between SeqSphere and NanoCore
- Supplementary Table 14	Experiment 3 part 1 (MRSA): Distance matrix
- Supplementary Table 15	Experiment 3 part 1 (VRE): Distance matrix
- Supplementary Table 16	Resource statistics of the different experiment runs
- Supplementary Table 17	Statistics of the used Oxford Nanopore and Illumina sequencing data
- Supplementary Table 18	Available core genome Multi Locus Sequence Typing schemata

Declarations

Availability of data and code

The utilized sequencing data are available under BioProject ID PRJNA1012291. NanoCore is available on GitHub (https://github.com/SebastianMeyer1989/NanoCore; DOI: https://doi.org/10.5281/zenodo.8424707) and licensed under the MIT license.

Conflicts of interest

The authors declare no conflicts of interest.

Funding

This work has been supported by the Jürgen Manchot Foundation, Deutsche Forschungsgemeinschaft (DFG) award 428994620, and German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung; Netzwerk Universitätsmedizin, GenSurv/MolTraX) award number 01KX2021.

Authors contributions

Sebastian A. Fuchs: Conceptualization (supporting); Data curation (equal); Formal analysis (lead); Investigation (lead); Methodology (equal); Project administration (supporting); Software (lead); Validation (lead); Visualization (lead); Writing - original draft preparation (equal); Writing - review and editing (lead).

Lisanna Hülse: Data curation (equal); Writing - review and editing (supporting).

Teresa Tamayo: Data curation (equal); Validation (supporting); Writing - review and editing (supporting).

Susanne Kolbe-Busch: Data curation (equal); Validation (supporting); Writing - review and editing (supporting).

Klaus Pfeffer: Funding acquisition (equal); Resources (lead); Writing - review and editing (supporting).

Alexander T. Dilthey: Conceptualization (lead); Funding acquisition (equal); Methodology (equal); Project administration (lead); Software (supporting); Supervision (lead); Validation (lead); Writing - original draft preparation (equal); Writing - review and editing (lead).

All authors read and approved the final manuscript.

Acknowledgments

Illumina sequencing was performed at the Biologisch-Medizinisches Forschungszentrum der Heinrich-Heine-Universität Düsseldorf (BMFZ).

References

1. Werner G, Couto N, Feil EJ, Novais A, Hegstad K, Howden BP, et al. Taking hospital pathogen surveillance to the next level. Microb Genom. 2023 Apr;9(4):mgen001008.
2. Quainoo S, Coolen JPM, van Hijum SAFT, Huynen MA, Melchers WJG, van Schaik W, et al. Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. Clinical Microbiology Reviews. 2017 Oct;30(4):1015–63.
3. Mellmann A, Bletz S, Böking T, Kipp F, Becker K, Schultes A, et al. Real-Time Genome Sequencing of Resistant Bacteria Provides Precision Infection Control in an Institutional Setting. J Clin Microbiol. 2016 Dec;54(12):2874–81.
4. Harris SR, Cartwright EJ, Török ME, Holden MT, Brown NM, Ogilvy-Stuart AL, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. The Lancet Infectious Diseases. 2013 Feb 1;13(2):130–6.
5. Lee XJ, Elliott TM, Harris PNA, Douglas J, Henderson B, Watson C, et al. Clinical and Economic Outcomes of Genome Sequencing Availability on Containing a Hospital Outbreak of Resistant Escherichia coli in Australia. Value

Health. 2020 Aug;23(8):994–1002.

6. Peacock SJ, Parkhill J, Brown NM. Changing the paradigm for hospital outbreak detection by leading with genomic surveillance of nosocomial pathogens. Microbiology (Reading). 2018 Oct;164(10):1213–9.
7. Reuter S, Ellington MJ, Cartwright EJP, Köser CU, Török ME, Gouliouris T, et al. Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology. JAMA Intern Med. 2013 Aug 12;173(15):1397–404.
8. Snitkin ES, Won S, Pirani A, Lapp Z, Weinstein RA, Lolans K, et al. Integrated genomic and interfacility patient-transfer data reveal the transmission pathways of multidrug-resistant Klebsiella pneumoniae in a regional outbreak. Sci Transl Med. 2017 Nov 22;9(417):eaan0093.
9. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Research. 2011;39(13):13.
10. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, et al. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif. 2015 Mar 24;3:1–8.
11. Krishnakumar R, Sinha A, Bird SW, Jayamohan H, Edwards HS, Schoeniger JS, et al. Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias. Sci Rep [Internet]. 2018 Feb 16 [cited 2020 Jan 7];8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5816649/
12. Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, et al. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. :8.
13. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods. 2022 Jul;19(7):823–6.
14. Liu C, Yang X, Duffy BF, Hoisington-Lopez J, Crosby M, Porche-Sorbet R, et al. High-resolution HLA typing by long reads from the R10.3 Oxford nanopore flow cells. Human Immunology. 2021 Apr 1;82(4):288–95.
15. Dilthey AT, Meyer SA, Kaasch AJ. Ultraplexing: increasing the efficiency of long-read sequencing for hybrid assembly with k-mer-based multiplexing. Genome Biology. 2020 Mar 14;21(1):68.
16. COVID-19 Genomics UK (COG-UK) consortiumcontact@cogconsortium.uk. An integrated national scale SARS-CoV-2 genomic surveillance network. Lancet Microbe. 2020 Jul;1(3):e99–100.
17. Michaelsen TY, Bennedbæk M, Christiansen LE, Jørgensen MSF, Møller CH, Sørensen EA, et al. Introduction and transmission of SARS-CoV-2 lineage B.1.1.7, Alpha variant, in Denmark. Genome Med. 2022 May 4;14(1):47.
18. Houwaart T, Belhaj S, Tawalbeh E, Nagels D, Fröhlich Y, Finzer P, et al. Integrated genomic surveillance enables tracing of person-to-person SARS-CoV-2 transmission chains during community transmission and reveals extensive onward transmission of travel-imported infections, Germany, June to July 2021. Euro Surveill. 2022 Oct;27(43):2101089.
19. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences. 1998 Mar 17;95(6):3140–5.
20. Maiden MCJ, van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013 Oct;11(10):728–36.
21. Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. Clin Microbiol Infect. 2018 Apr;24(4):350–4.
22. Magi A, Semeraro R, Mingrino A, Giusti B, D'Aurizio R. Nanopore sequencing data analysis: state of the art, applications and challenges. Briefings in Bioinformatics. 2018 Nov 27;19(6):1256–72.
23. Ahrenfeldt J, Skaarup C, Hasman H, Pedersen AG, Aarestrup FM, Lund O. Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods. BMC Genomics. 2017 Jan 5;18(1):19.
24. Oh S, Nam SK, Chang HE, Park KU. Comparative Analysis of Short- and Long-Read Sequencing of Vancomycin-Resistant Enterococci for Application to Molecular Epidemiology. Front Cell Infect Microbiol. 2022;12:857801.
25. Hall MB, Rabodoarivelo MS, Koch A, Dippenaar A, George S, Grobbelaar M, et al. Evaluation of Nanopore sequencing for Mycobacterium tuberculosis drug susceptibility testing and outbreak investigation: a genomic analysis. Lancet Microbe. 2023 Feb;4(2):e84–92.
26. Liao YC, Wu HC, Liou CH, Lauderdale TLY, Huang IW, Lai JF, et al. Rapid and Routine Molecular Typing Using Multiplex Polymerase Chain Reaction and MinION Sequencer. Front Microbiol. 2022;13:875347.
27. Liou CH, Wu HC, Liao YC, Yang Lauderdale TL, Huang IW, Chen FJ. nanoMLST: accurate multilocus sequence typing using Oxford Nanopore Technologies MinION with a dual-barcode approach to multiplex large numbers of samples. Microb Genom. 2020 Feb 17;6(3):e000336.
28. Ferreira FA, Helmersen K, Visnovska T, Jørgensen SB, Aamot HV. Rapid nanopore-based DNA sequencing protocol of antibiotic-resistant bacteria for use in surveillance and outbreak investigation. Microb Genom. 2021 Apr 22;7(4):000557.
29. Xian Z, Li S, Mann DA, Huang Y, Xu F, Wu X, et al. Subtyping Evaluation of Salmonella Enteritidis Using Single Nucleotide Polymorphism and Core Genome Multilocus Sequence Typing with Nanopore Reads. Appl Environ Microbiol. 2022 Aug 9;88(15):e0078522.

30. Wu X, Luo H, Ge C, Xu F, Deng X, Wiedmann M, et al. Evaluation of multiplex nanopore sequencing for Salmonella serotype prediction and antimicrobial resistance gene and virulence gene detection. Front Microbiol. 2022;13:1073057.
31. Wu X, Luo H, Xu F, Ge C, Li S, Deng X, et al. Evaluation of Salmonella Serotype Prediction With Multiplex Nanopore Sequencing. Front Microbiol. 2021;12:637771.
32. Xu F, Ge C, Luo H, Li S, Wiedmann M, Deng X, et al. Evaluation of real-time nanopore sequencing for Salmonella serotype prediction. Food Microbiol. 2020 Aug;89:103452.
33. Linde J, Brangsch H, Hölzer M, Thomas C, Elschner MC, Melzer F, et al. Comparison of Illumina and Oxford Nanopore Technology for genome analysis of Francisella tularensis, Bacillus anthracis, and Brucella suis. BMC Genomics. 2023 May 12;24(1):258.
34. Greig DR, Jenkins C, Gharbia S, Dallman TJ. Comparison of single-nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga toxin-producing Escherichia coli. Gigascience. 2019 Aug 1;8(8):giz104.
35. Tarumoto N, Sakai J, Sujino K, Yamaguchi T, Ohta M, Yamagishi J, et al. Use of the Oxford Nanopore MinION sequencer for MLST genotyping of vancomycin-resistant enterococci. J Hosp Infect. 2017 Jul;96(3):296–8.
36. Both A, Kruse F, Mirwald N, Franke G, Christner M, Huang J, et al. Population dynamics in colonizing vancomycin-resistant Enterococcus faecium isolated from immunosuppressed patients. J Glob Antimicrob Resist. 2022 Mar;28:267–73.
37. Cao MD, Ganesamoorthy D, Elliott AG, Zhang H, Cooper MA, Coin LJM. Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time MinION(TM) sequencing. Gigascience. 2016 Jul 26;5(1):32.
38. Higgs C, Sherry NL, Seemann T, Horan K, Walpola H, Kinsella P, et al. Optimising genomic approaches for identifying vancomycin-resistant Enterococcus faecium transmission in healthcare settings. Nat Commun. 2022 Jan 26;13(1):509.
39. Humphreys H, Coleman DC. Contribution of whole-genome sequencing to understanding of the epidemiology and control of meticillin-resistant Staphylococcus aureus. J Hosp Infect. 2019 Jun;102(2):189–99.
40. Egan SA, Corcoran S, McDermott H, Fitzpatrick M, Hoyne A, McCormack O, et al. Hospital outbreak of linezolid-resistant and vancomycin-resistant ST80 Enterococcus faecium harbouring an optrA-encoding conjugative plasmid investigated by whole-genome sequencing. J Hosp Infect. 2020 Aug;105(4):726–35.
41. Hyun JC, Monk JM, Palsson BO. Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. BMC Genomics. 2022 Jan 4;23(1):7.
42. Mortensen K, Lam TJ, Ye Y. Comparison of CRISPR–Cas Immune Systems in Healthcare-Related Pathogens. Frontiers in Microbiology [Internet]. 2021 [cited 2023 Aug 28];12. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2021.758782
43. Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, et al. Updating benchtop sequencing performance comparison. Nat Biotechnol. 2013 Apr;31(4):294–6.
44. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018 Jul;15(7):475–6.
45. Enright MC, Day NPJ, Davies CE, Peacock SJ, Spratt BG. Multilocus Sequence Typing for Characterization of Methicillin-Resistant and Methicillin-Susceptible Clones of Staphylococcus aureus. Journal of Clinical Microbiology. 2000 Mar;38(3):1008–15.
46. de Been M, Pinholt M, Top J, Bletz S, Mellmann A, van Schaik W, et al. Core Genome Multilocus Sequence Typing Scheme for High-Resolution Typing of Enterococcus faecium. Journal of Clinical Microbiology. 2015 Dec;53(12):3788–97.
47. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing [Internet]. arXiv; 2012 [cited 2023 Jul 20]. Available from: http://arxiv.org/abs/1207.3907
48. Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, et al. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill. 2017 Jun 8;22(23):30544.
49. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol [Internet]. 2016 Jun 20 [cited 2020 Jan 7];17. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4915045/
50. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094–100.
51. Zheng Z, Li S, Su J, Leung AWS, Lam TW, Luo R. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling [Internet]. bioRxiv; 2021 [cited 2022 Jun 29]. p. 2021.12.29.474431. Available from: https://www.biorxiv.org/content/10.1101/2021.12.29.474431v1
52. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics. 1996 Sep 1;5(3):299–314.
53. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.
54. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics. 2013 Mar 1;14(2):178–92.

3.6 Supplementary material to publication 3

Due to their size, the supplementary tables could not be inserted here. They can be accessed via the enclosed CD instead. The supplementary figures mentioned in the publication are listed on the following pages.

SupFig 1 – 1/2

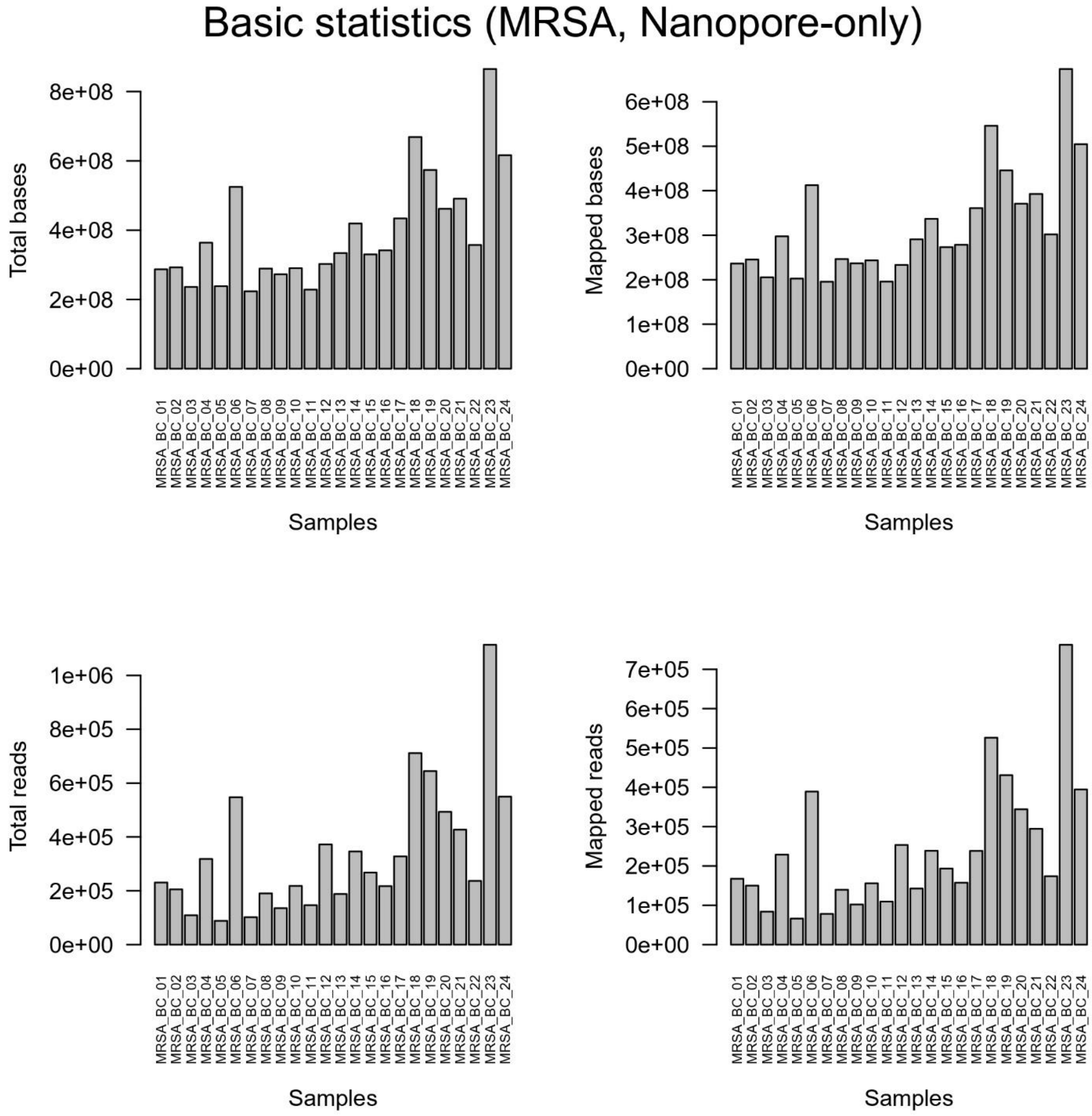
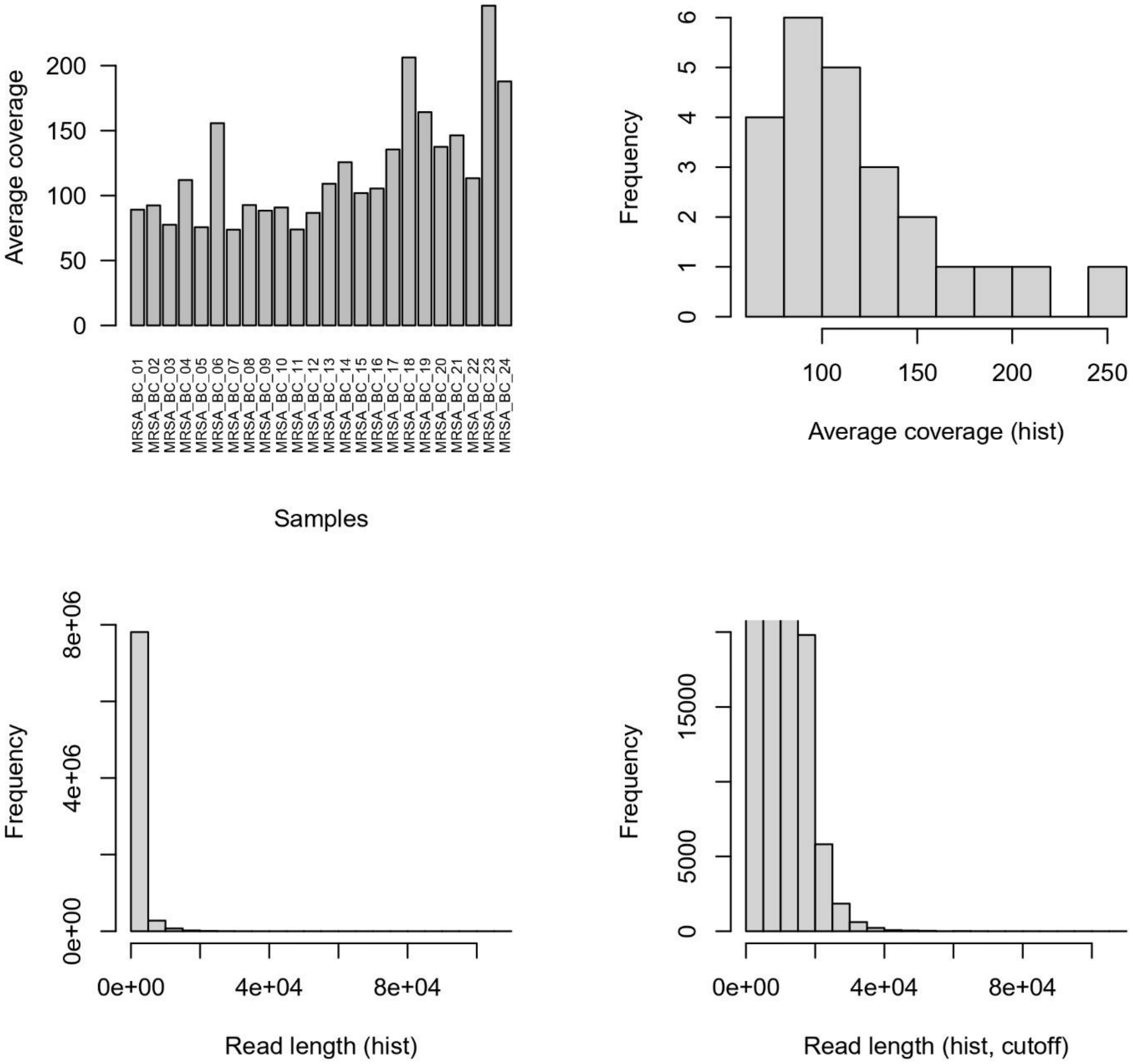


Fig S1 – 2/2



Supplementary Figure 1: Basic statistics of the Nanopore-only validation experiment 1 on MRSA data. Shown are the number of total bases, the number of mapped bases, the number of total reads, the number of mapped reads and the average coverage, each per sample, as well as frequency histograms of the average coverage and the read length.

Fig S2 – 1/3

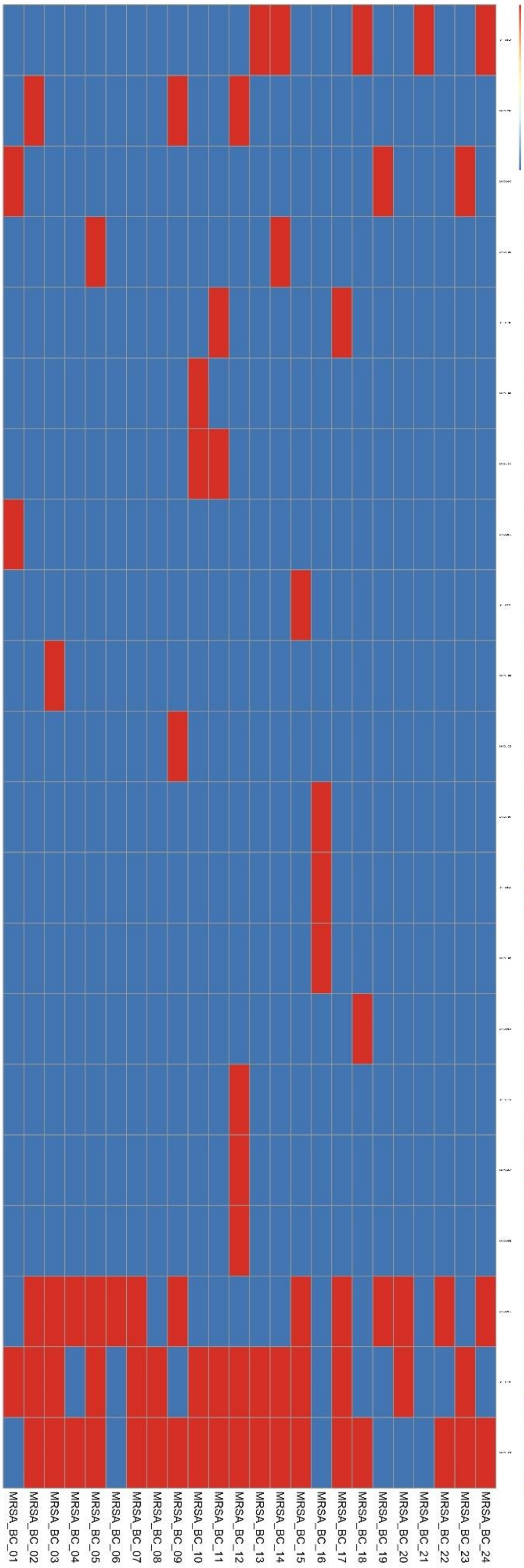


Fig S2 – 2/3

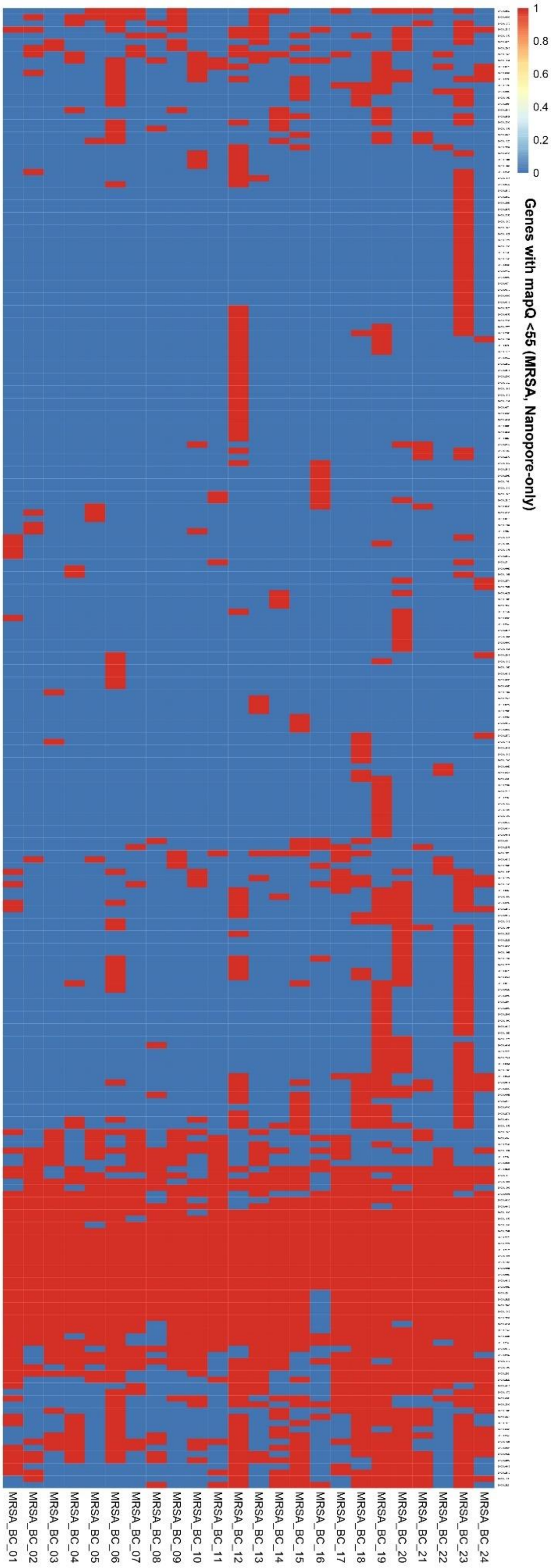
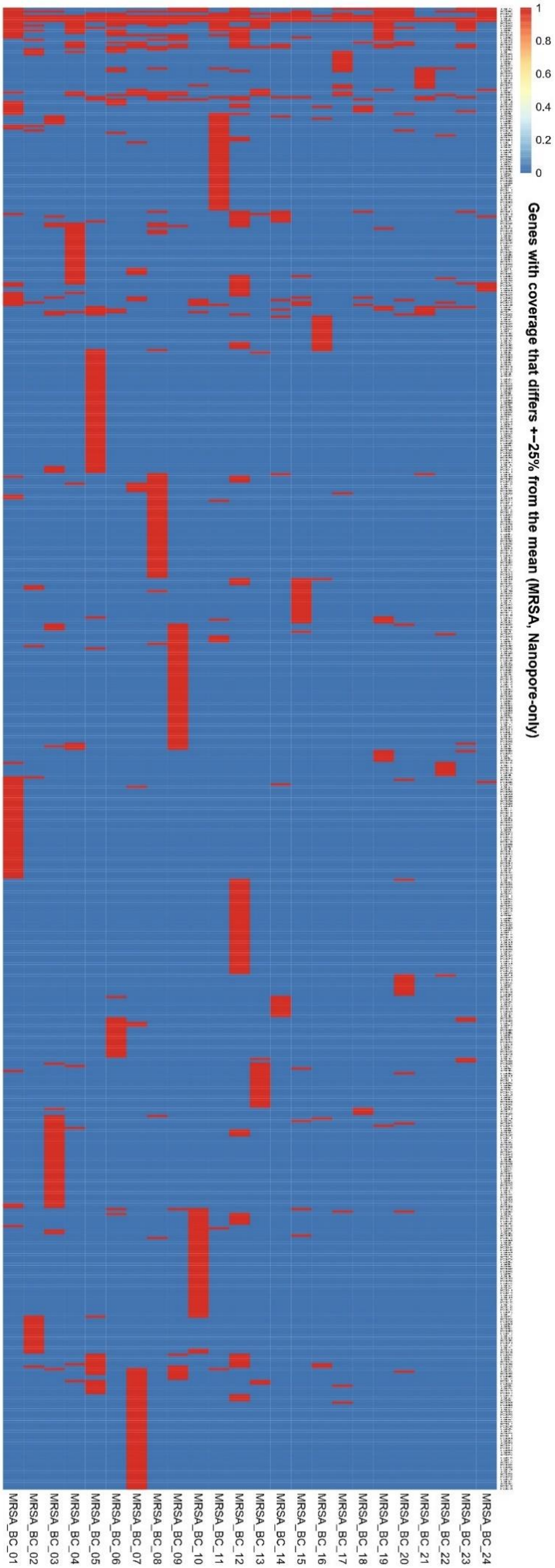


Fig S2 – 3/3



Supplementary Figure 2: Heatmaps of eye-catching and potentially due to different filters excluded genes in the Nanopore-only validation experiment 1 on MRSA data. Shown are genes with more than 50% of heterozygous variant-calls, genes with below 55 mapping quality and genes with a coverage that differs more than 25% from the mean.

Fig S3 – ½

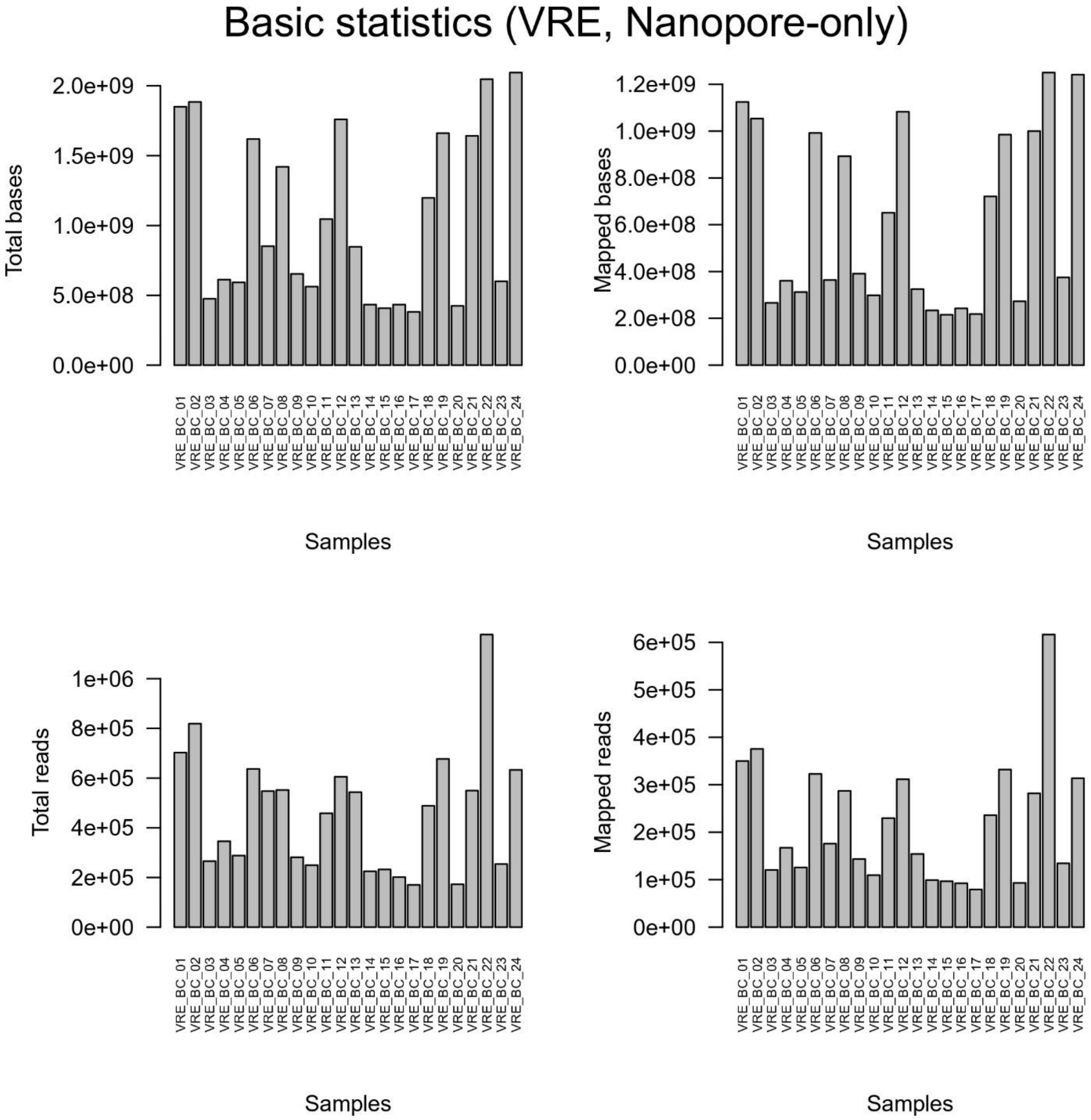
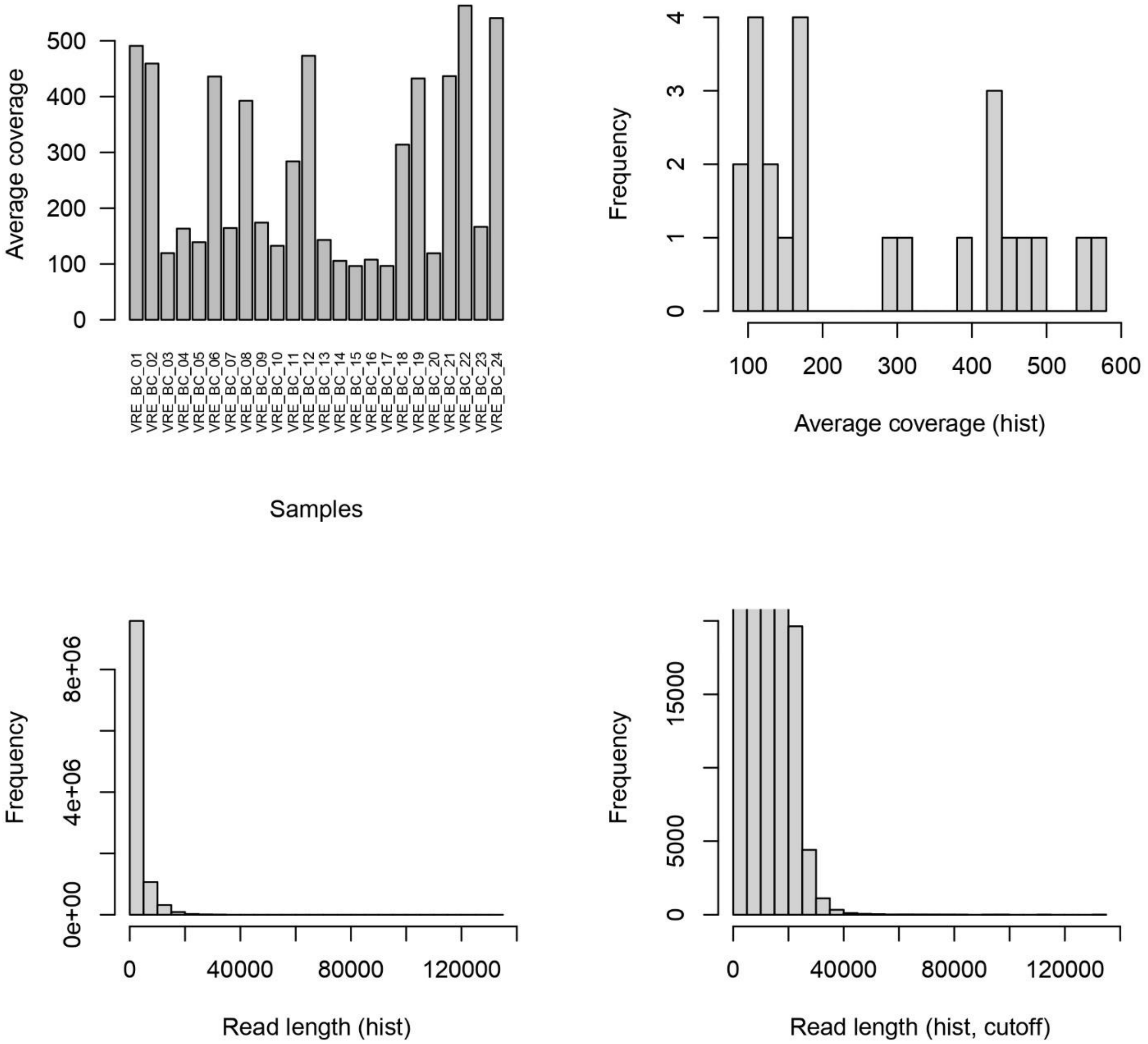

Basic statistics (VRE, Nanopore-only)
Total bases
Mapped bases
Total reads
Mapped reads
Samples
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
2.0e+08
4.0e+08
6.0e+08
8.0e+08
1.2e+09
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
VRE_BC_01
VRE_BC_02
VRE_BC_03
VRE_BC_04
VRE_BC_05
VRE_BC_06
VRE_BC_07
VRE_BC_08
VRE_BC_09
VRE_BC_10
VRE_BC_11
VRE_BC_12
VRE_BC_13
VRE_BC_14
VRE_BC_15
VRE_BC_16
VRE_BC_17
VRE_BC_18
VRE_BC_19
VRE_BC_20
VRE_BC_21
VRE_BC_22
VRE_BC_23
VRE_BC_24

Fig S3 – 2/2



Supplementary Figure 3: Basic statistics of the Nanopore-only validation experiment 2 on VRE data. Shown are the number of total bases, the number of mapped bases, the number of total reads, the number of mapped reads and the average coverage, each per sample, as well as frequency histograms of the average coverage and the read length.

Fig S4 – 1/3

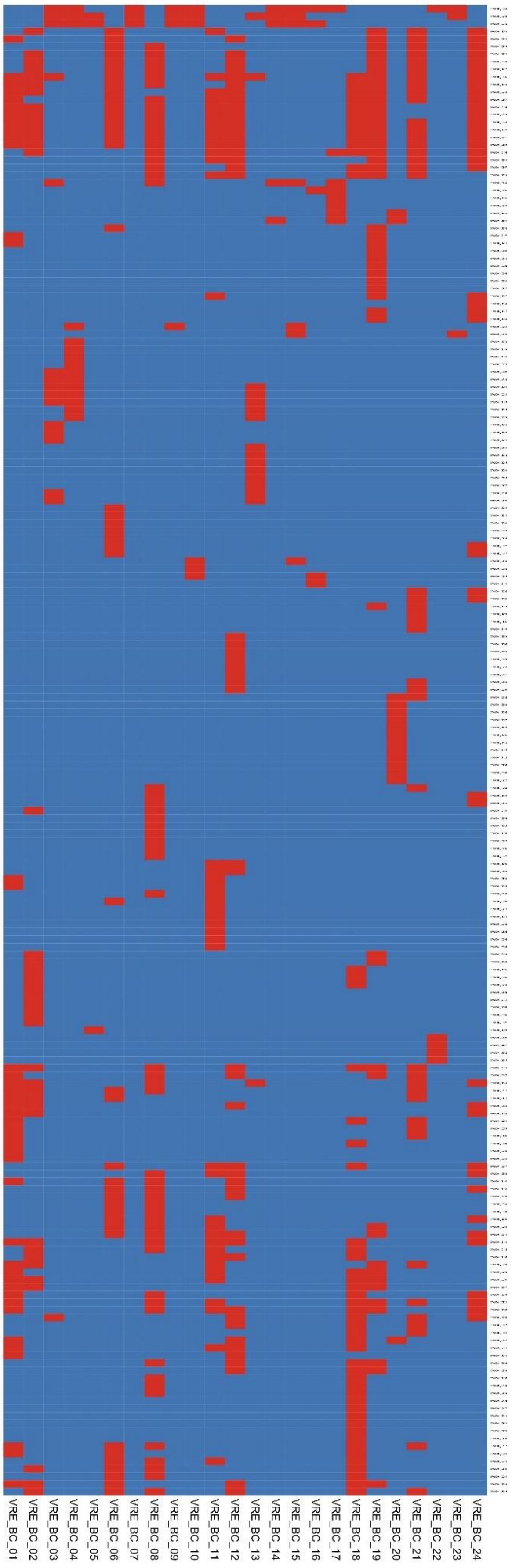


Fig S4 – 2/3

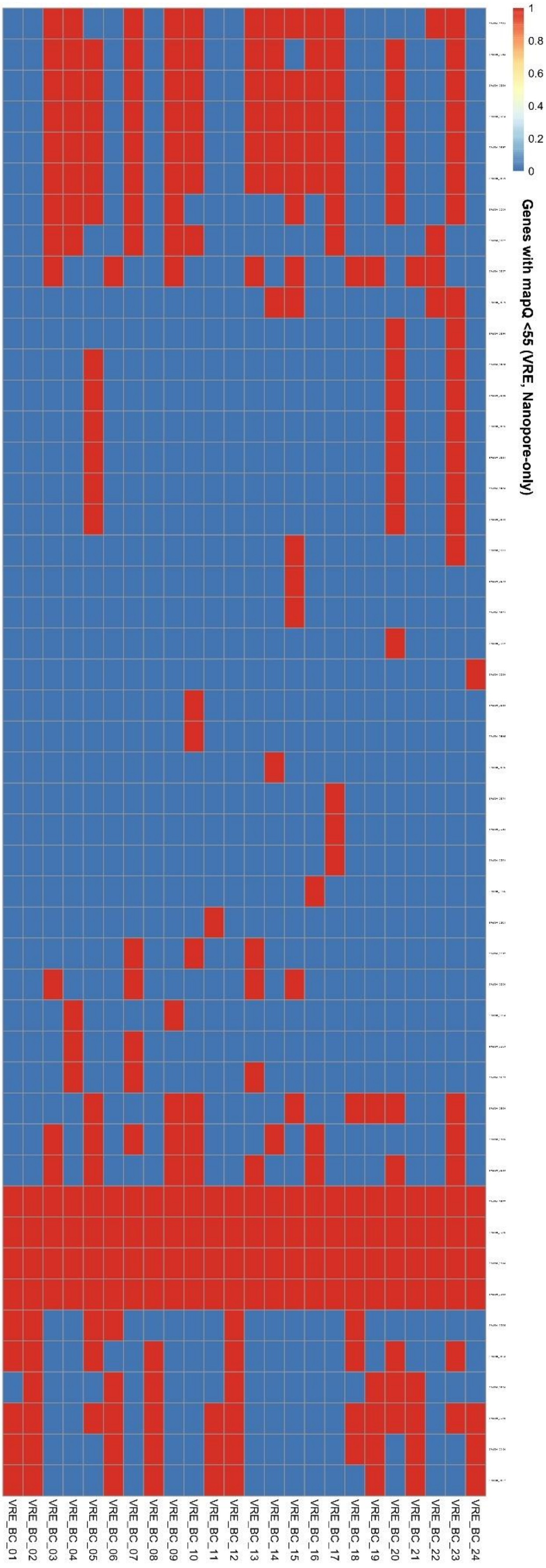
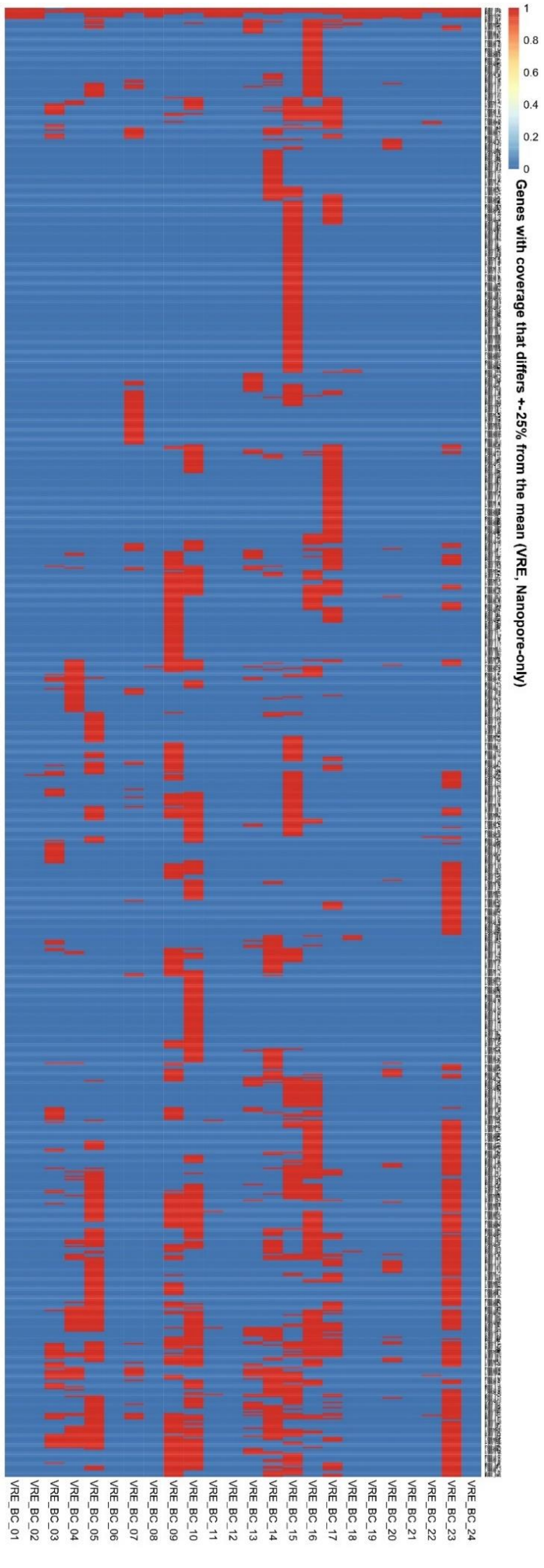


Fig S4 – 3/3



Supplementary Figure 4: Heatmaps of eye-catching and potentially due to different filters excluded genes in the Nanopore-only validation experiment 2 on VRE data. Shown are genes with more than 50% of heterozygous variant-calls, genes with below 55 mapping quality and genes with a coverage that differs more than 25% from the mean.

Fig S5 – 1/3

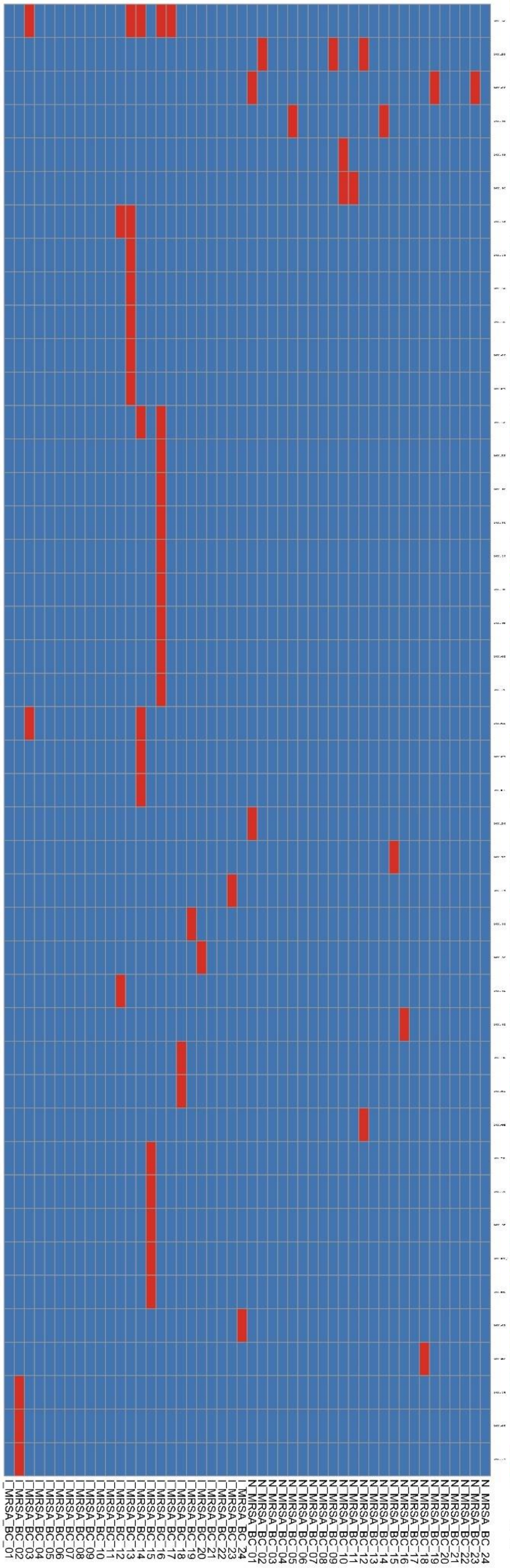


Fig S5 – 2/3

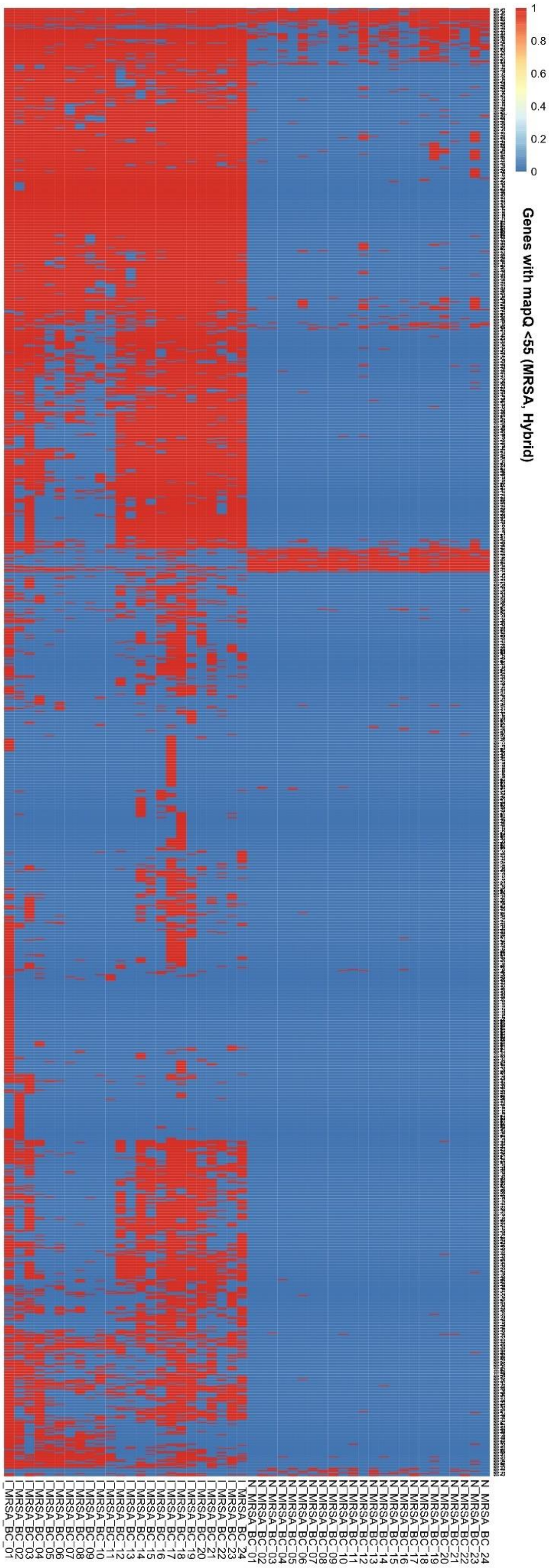
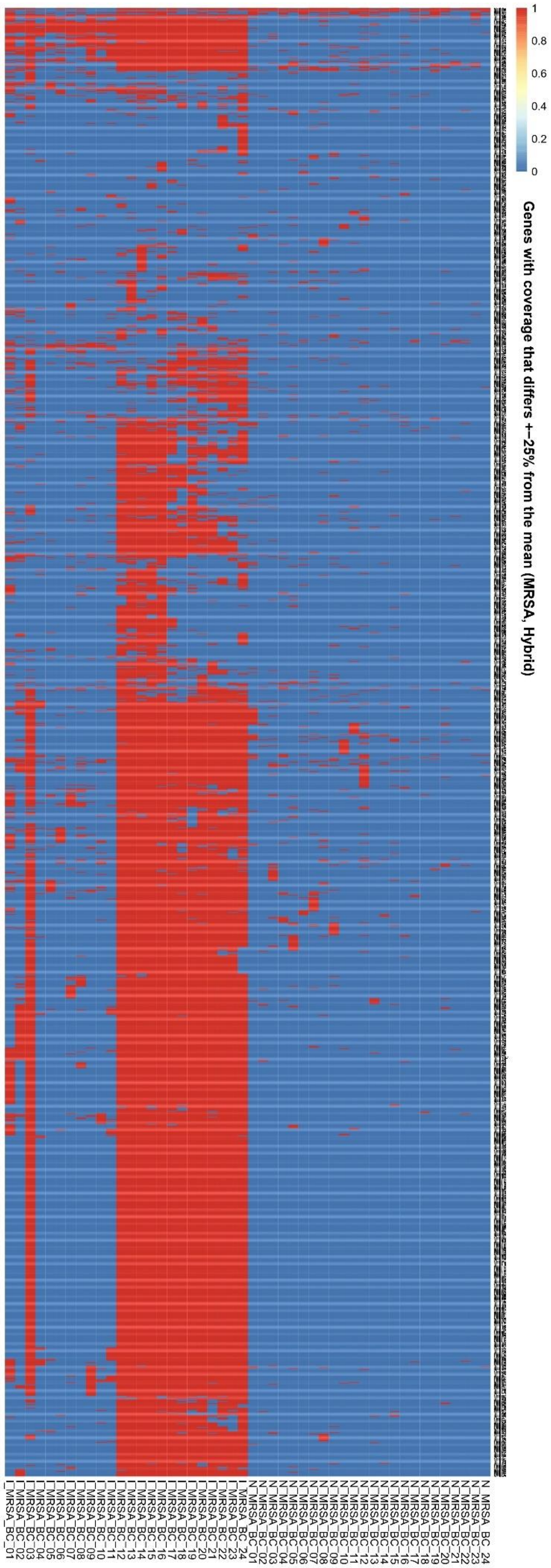


Fig S5 – 3/3



Supplementary Figure 5: Heatmaps of eye-catching and potentially due to different filters excluded genes in the hybrid validation experiment 3 part 1 on MRSA data. Shown are genes with more than 50% of heterozygous variant-calls, genes with below 55 mapping quality and genes with a coverage that differs more than 25% from the mean.

Fig S6 – 1/3

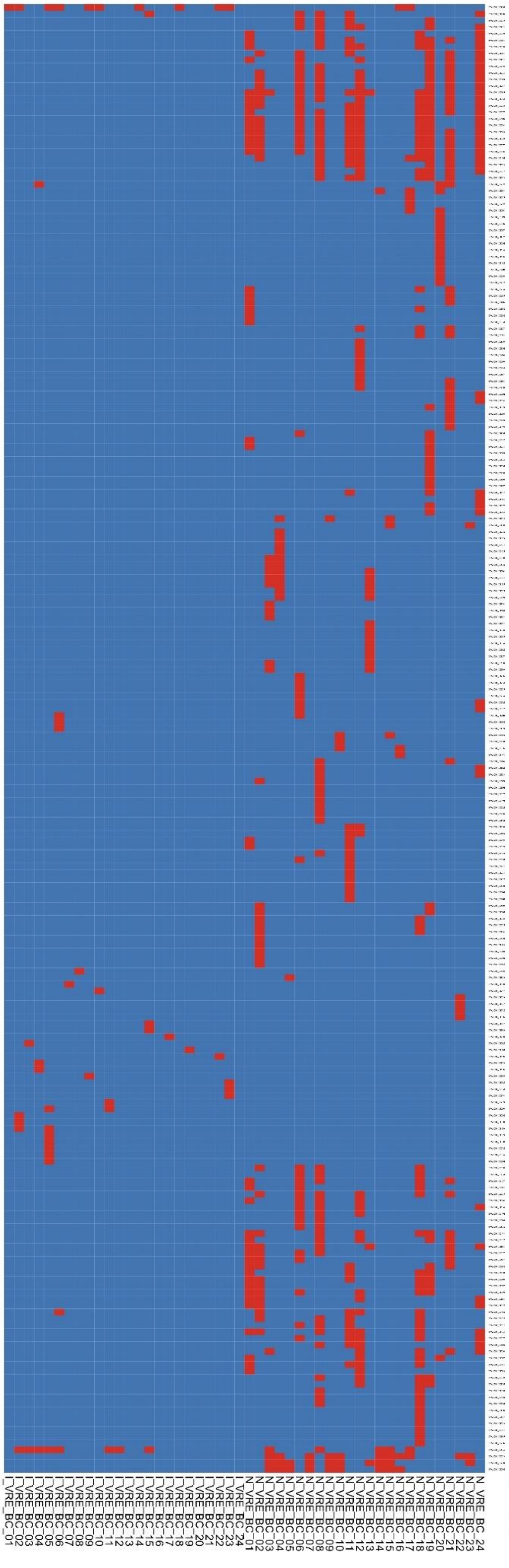


Fig S6 – 2/3

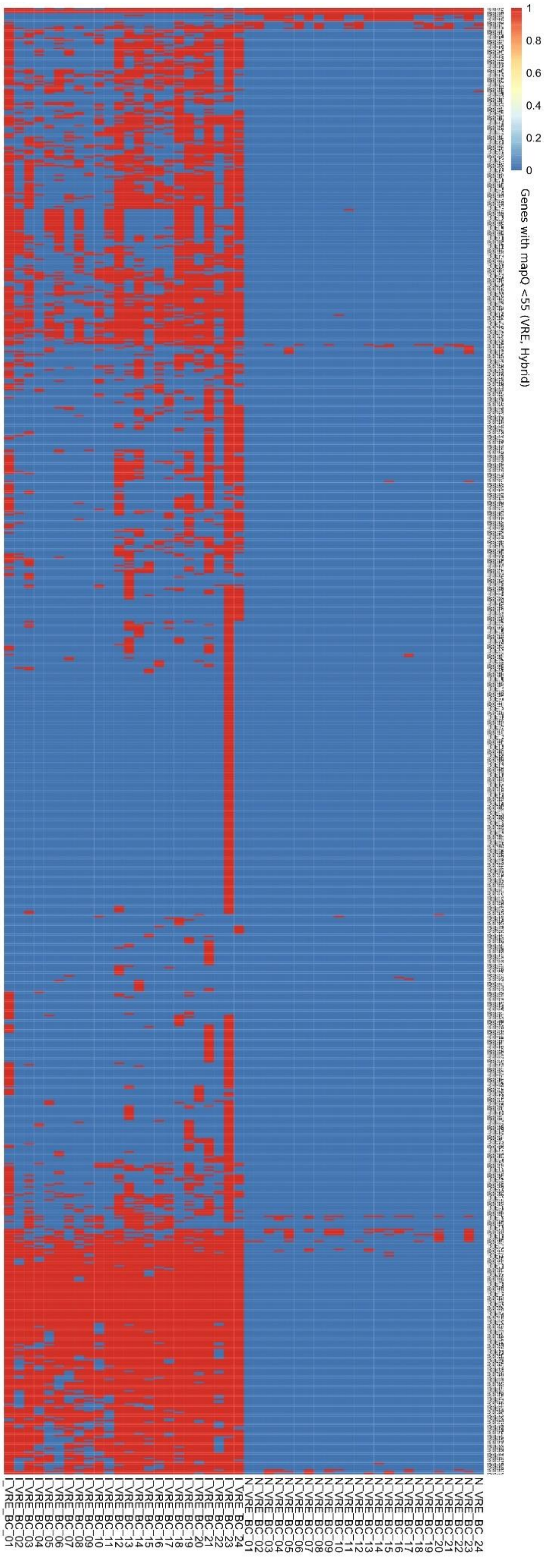
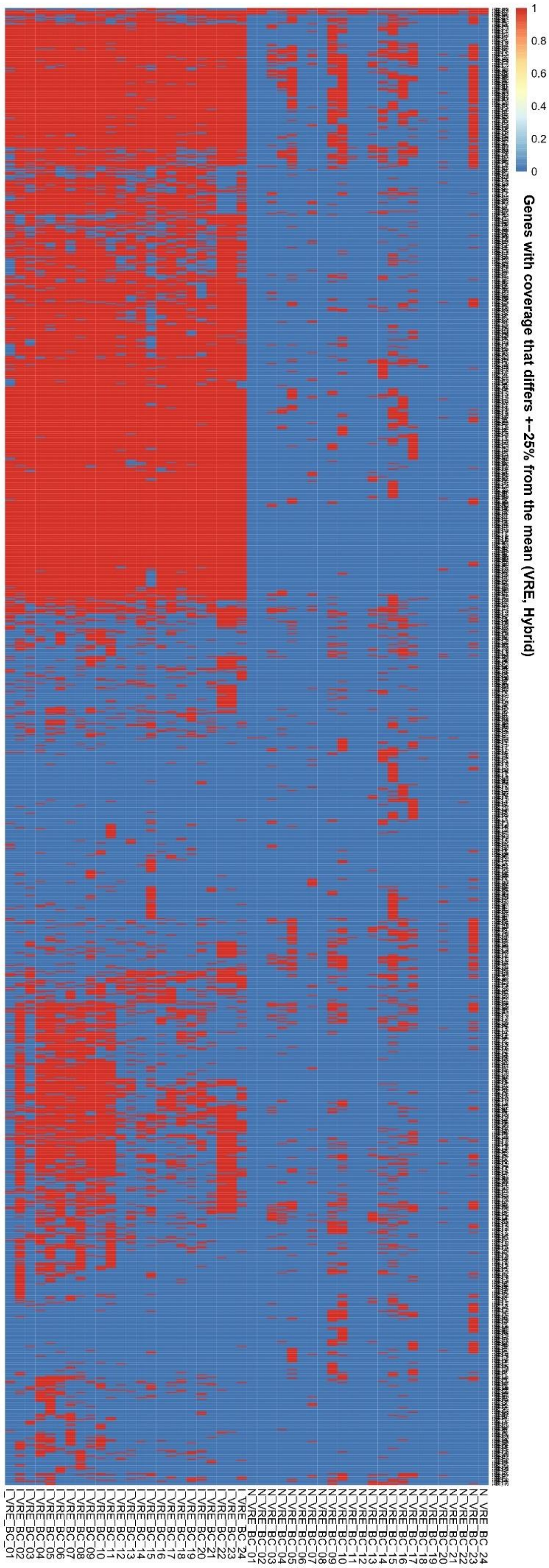
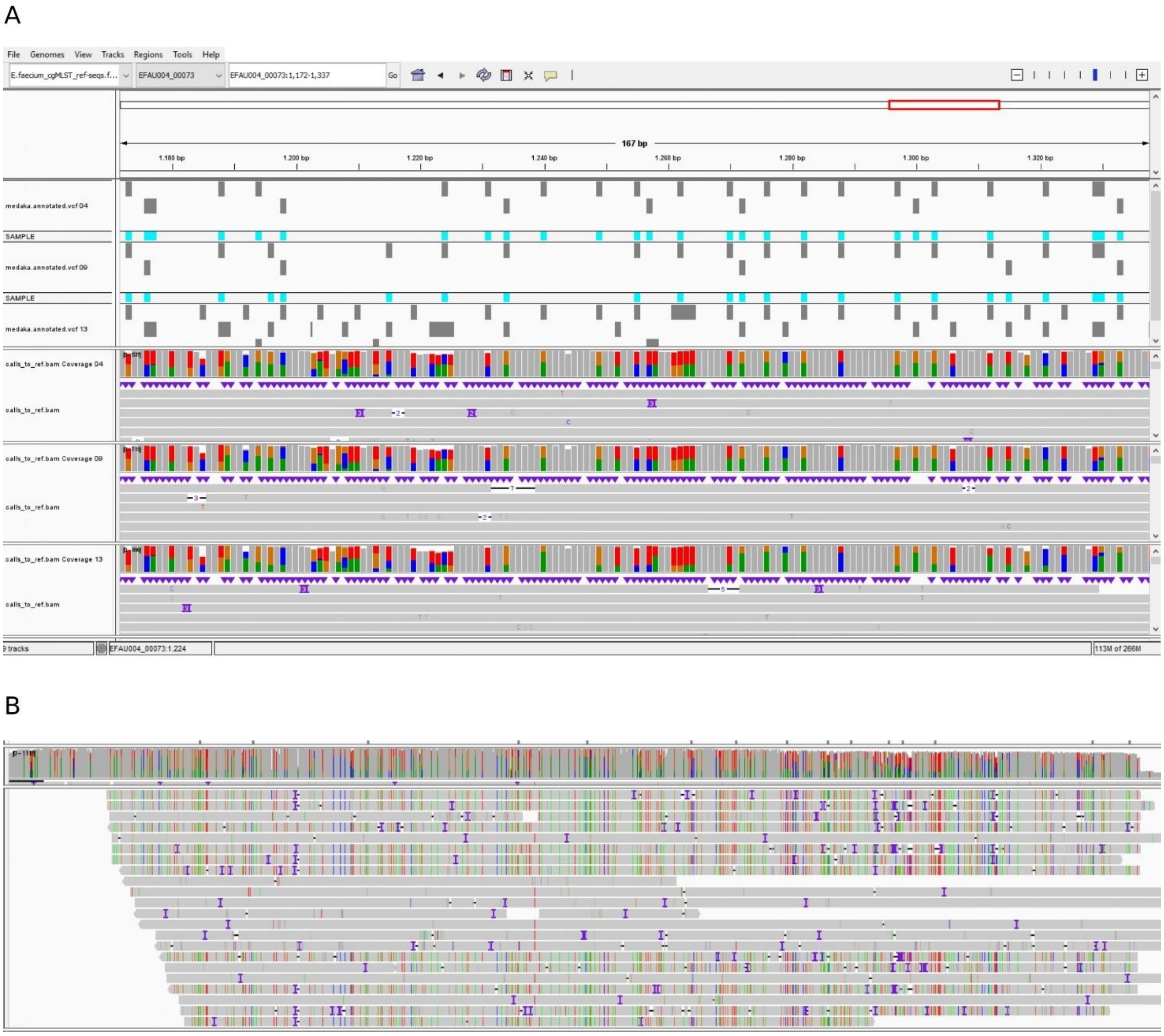


Fig S6 – 3/3



Supplementary Figure 6: Heatmaps of eye-catching and potentially due to different filters excluded genes in the hybrid validation experiment 3 part 1 on VRE data. Shown are genes with more than 50% of heterozygous variant-calls, genes with below 55 mapping quality and genes with a coverage that differs more than 25% from the mean.

IGV screenshots of possible gene duplication



Supplementary Figure 7: Integrative Genomics Viewer screenshot of a genomic area that shows reads with variant patterns of a possible gene duplication. Shown are A) a zoomed-in position of gene EFAU004_00073 that shows that most of the found variants indicate two different bases with around 50% ratio each and B) the full length of gene EFAU004_00073 that shows, that all reads exhibit one of two possible variant patterns.

Chapter 4: Integrated Discussion

In this thesis, I have demonstrated three applications aimed at [1] optimizing hybrid genome assembly, [2] advancing the understanding of genomic plasticity in the context of antibiotic resistance and [3] enhancing methods for outbreak investigations.

These applications are designed to improve our capabilities in bacterial genomics, antibiotic research, and real-time genomic surveillance in healthcare.

The results of all three projects underscore the pivotal role of ongoing developments in sequencing technologies for the improvement of various analyses and downstream processes. Whether it is the classification of sequencing data, the assembly of reads into unfragmented genomes, the localization of resistance genes, the identification of potential (horizontal) gene transfers, or the detection of outbreaks and tracking of their transmission pathways – technological advancements, particularly in Oxford Nanopore sequencing, have been instrumental.

The improvements in technologies, like developments of new flowcells and basecalling technologies within the 10.4.1 high-accuracy sequencing of Nanopore, or the new HiFi sequencing of PacBio, have opened up possibilities that were previously inconceivable. Cost-efficient sequencing and the creation of nearly perfect assemblies, without the need for additional high-quality Illumina sequencing data, are now feasible (24).

These advancements have potential implications for the future of the shown analysis and the tools developed, and some projects could have been approached differently, had the improved versions of these technologies already been available at the time of working on the projects.

Ultraplexing, a method to classify Nanopore long-reads with the help of Illumina short-reads in projects that need both datatypes could become obsolete for certain areas of application, as current Nanopore data alone is often sufficient for fully resolved assembly and Illumina data is not needed to begin with. However, it still has use-cases, like low-coverage long-read sequencing of large numbers of already Illumina-sequenced isolates.

For the identification of recombination events and potential transmission of resistance genes,the quality of the assembled genomes plays a crucial role. The process becomes even more cost-efficient if Illumina data is not needed. In addition, instead of using k-mers, a direct comparison between Nanopore assemblies and references could provide faster and more accurate results. Next to this, methods of algorithmic bioinformatics could be applied to pinpoint recombination event borders explicitly.

Regarding the third project, the accurate calculation of inter-sample genetic distances, which the application of sequencing for the characterization of transmission chains depends upon, strongly benefits from the availability of higher-accuracy sequencing chemistries. On a high level, the development of the R10.4 chemistry thus improves the applicability of Nanopore sequencing for outbreak investigation and underscores the need for specialized tools. Moreover, enhanced Nanopore read quality could have facilitated (and may continue to do so in the future) the evolution of this tool toward assembly calculations and direct alignment, rather than read mapping. A shift to a SNP-based comparison, as opposed to cgMLST-based, could have been implemented, allowing the flexibility to incorporate gene-based distance calculations through the annotation of the generated assemblies. This approach might also involve constructing a pan-genomic graph as a reference (93), encompassing reference genomes for multiple strains within a species rather than focusing on individual genes. Such an approach would enable a more comprehensive analysis that also includes the accessory genome and offers even higher-resolution insights into potential strain affiliations of an isolate.

In general, the concept of pan-genomics, considering the entire genome with its core and accessory components as one unit, is a major development din current bioinformatics and closely intertwined with the three projects discussed in this thesis.

The availability of fully resolved genomes (project one) benefits pangenomics applications, for example by accurately defining distinct components.

The unravelling of horizontal gene transfers around resistance genes (project two), directly addresses pan-genomic questions by characterising the exchange of genetic material between different members of the same species. It attempts to discern the precise locations of recombination breakpoints, elucidate the characteristics of surrounding genes, and explore the potential mechanisms through which these genes could be transferred among different isolates.

The analysis of outbreaks (project three) relies on pan-genomic concepts, such as the categorization into core and accessory genomes. NanoCore, like all other cgMLST-like schemes, relies intrinsically on the ability to define the essential genes common to all strains within a species, known as core genes. Additionally, as mentioned in the preceding paragraph, there is potential for future enhancements to NanoCore through the incorporation of a pan-genomic approach to references.

The utilization of these technological advancements is not solely dependent on improved methods but also hinges on the mindset with which these developments are approached and, on the questions, (future) users aim to address. For instance, the decision to establish a „prospective“ routine sample collection and analysis scheme, instead of only sequencing samples that have already been implicated in an epidemiologically detected outbreak.

Due to the increasing spread of antibiotic resistances among clinically relevant pathogens, there will be an increasing pressure on healthcare systems, leading to the need to better comprehend antimicrobial resistances and their dissemination. Therefore, there is a need to develop tools that facilitate the faster identification of newly emerging resistances. Moreover, the implementation of effective measures also depends on knowledge of and compliance with methods to prevent the spread of antimicrobial resistances, such as hand hygiene.

The use of genomically informed outbreak investigations and disease prevention measures has already become a gold standard in many hospitals and some areas of public health. With ongoing improvements in availability and cost-efficiency, these practices are likely to extend gradually to healthcare institutions outside of large hospitals, such as general practitioners. Noteworthy examples of such developments include the widespread adoption of testing and vaccination procedures by medically-informed physicians and the establishment of specialized test centres during the COVID-19 pandemic (94–96). Consequently, corresponding tools must be designed to be as user-friendly and versatile as possible.

Currently, a few hundred isolates can be sequenced in a time-efficient manner using gold-standard methods. With further advancements in cost-efficiency and throughput, it will become feasible to sequence not only hundreds or thousands but possibly hundreds of thousands of isolates. These advancements underscore the imperative for scalable bioinformatic methods that process sequencing data to evolve further or the development of new ones capable of handling such amounts of data.

What do these developments mean for the two main questions I aimed to address within this dissertation (refer to Introduction)?

Firstly, fully resolving bacterial genomes has become significantly more straightforward and will continue to do so. The common sufficiency of Nanopore sequencing data alone for downstream processes, without the need to produce Illumina sequencing data for quality reasons, means that sequencing more isolates concurrently within a short timeframe will progressively become more cost-efficient.

Secondly, the analysis of bacterial genome sequencing data, whether for clinical questions such as outbreak investigations or resistance tracking, or for basic research such as enhancing the database quality of available genomes, has become more efficient. This is primarily due to the fact that all tools rely on a certain level of sequencing data quality, and their performance improves dramatically with less erroneous data.

In conclusion, these advancements encourage the further utilization of high-throughput long-read sequencing and the development of new tools and methods explicitly tailored for the utilization of Nanopore data and perfectly resolved genomes derived from this data.

Chapter 5: References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463–7.

2. Canard B, Sarfati RS. DNA polymerase fluorescent substrates with reversible 3′-tags. Gene. 1994 Oct 11;148(1):1–6.

3. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015 Oct;13(5):278–89.

4. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, et al. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif. 2015 Mar;3:1–8.

5. Mardis ER. DNA sequencing technologies: 2006–2016. Nat Protoc. 2017 Feb;12(2):213–8.

6. Ng PC, Kirkness EF. Whole Genome Sequencing. In: Barnes MR, Breen G, editors. Genetic Variation: Methods and Protocols [Internet]. Totowa, NJ: Humana Press; 2010 [cited 2024 Jan 15]. p. 215–26. (Methods in Molecular Biology). Available from: https://doi.org/10.1007/978-1-60327-367-1_12

7. Magi A, Semeraro R, Mingrino A, Giusti B, D'Aurizio R. Nanopore sequencing data analysis: state of the art, applications and challenges. Brief Bioinform. 2018 Nov 27;19(6):1256–72.

8. Ahrenfeldt J, Skaarup C, Hasman H, Pedersen AG, Aarestrup FM, Lund O. Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods. BMC Genomics. 2017 Jan 5;18(1):19.

9. Liou CH, Wu HC, Liao YC, Yang Lauderdale TL, Huang IW, Chen FJ. nanoMLST: accurate multilocus sequence typing using Oxford Nanopore Technologies MinION with a dual-barcode approach to multiplex large numbers of samples. Microb Genom. 2020 Feb 17;6(3):e000336.

10. Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, et al. Updating benchtop sequencing performance comparison. Nat Biotechnol. 2013 Apr;31(4):294–6.

11. Fitzpatrick MA, Ozer EA, Hauser AR. Utility of Whole-Genome Sequencing in Characterizing Acinetobacter Epidemiology and Analyzing Hospital Outbreaks. Journal of Clinical Microbiology. 2016 Feb 25;54(3):593–612.

12. Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, et al. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology. 2023 Jul;12(7):997.

13. Reuter S, Ellington MJ, Cartwright EJP, Köser CU, Török ME, Gouliouris T, et al. Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology. JAMA Intern Med. 2013 Aug 12;173(15):1397–404.

14. Croucher NJ, Harris SR, Grad YH, Hanage WP. Bacterial genomes in epidemiology—present and future. Philosophical Transactions of the Royal Society B: Biological Sciences. 2013 Mar 19;368(1614):20120202.

15. Greub G. Genomics of medical importance. Clinical Microbiology and Infection. 2013 Sep 1;19(9):781–3.

16. Saitou N. Eukaryote Genomes. Introduction to Evolutionary Genomics. 2013 Aug 22;17:193–222.

17. Dilthey AT, Meyer SA, Kaasch AJ. Ultraplexing: increasing the efficiency of long-read sequencing for hybrid assembly with k-mer-based multiplexing. Genome Biology. 2020 Mar 14;21(1):68.

18. Nodari CS, Fuchs SA, Xanthopoulou K, Cayô R, Seifert H, Gales AC, et al. pmrCAB Recombination Events among Colistin-Susceptible and -Resistant Acinetobacter baumannii Clinical Isolates Belonging to International Clone 7. mSphere. 2021 Dec;6(6):e00746-21.

19. Oxford Nanopore Technologies [Internet]. [cited 2024 Jan 15]. Community - Technical document. Available from: https://community.nanoporetech.com/technical_documents/chemistry-technical-document/v/chtd_500_v1_revan_07jul2016/barcoding-kits

20. Menzies BE. The role of fibronectin binding proteins in the pathogenesis of Staphylococcus aureus infections. Curr Opin Infect Dis. 2003 Jun;16(3):225–9.

21. Heydari M, Miclotte G, Van de Peer Y, Fostier J. Illumina error correction near highly repetitive DNA regions improves de novo genome assembly. BMC Bioinformatics. 2019 Jun 3;20(1):298.

22. Krishnakumar R, Sinha A, Bird SW, Jayamohan H, Edwards HS, Schoeniger JS, et al. Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias. Sci Rep [Internet]. 2018 Feb 16 [cited 2020 Jan 7];8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5816649/

23. Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, et al. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. Bioinformatics. 2014 Oct 1;30(19):2709–16.

24. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods. 2022 Jul;19(7):823–6.

25. Translation: DNA to mRNA to Protein | Learn Science at Scitable [Internet]. [cited 2024 Jan 15]. Available from: https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/

26. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. Science. 2001 Feb 16;291(5507):1304–51.

27. Rocha EPC. The Organization of the Bacterial Genome. Annual Review of Genetics. 2008;42(1):211–33.

28. Barton NH, Keightley PD. Understanding quantitative genetic variation. Nat Rev Genet. 2002 Jan;3(1):11–21.

29. Edwards D, Forster JW, Chagné D, Batley J. What Are SNPs? In: Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN, editors. Association Mapping in Plants [Internet]. New York, NY: Springer; 2007 [cited 2024 Jan 16]. p. 41–52. Available from: https://doi.org/10.1007/978-0-387-36011-9_3

30. Sehn JK. Chapter 9 - Insertions and Deletions (Indels). In: Kulkarni S, Pfeifer J, editors. Clinical Genomics [Internet]. Boston: Academic Press; 2015 [cited 2024 Jan 16]. p. 129–50. Available from: https://www.sciencedirect.com/science/article/pii/B9780124047488000095

31. MAGADUM S, BANERJEE U, MURUGAN P, GANGAPUR D, RAVIKESAVAN R. Gene duplication as a major force in evolution. J Genet. 2013 Apr 1;92(1):155–61.

32. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol. 2017 Mar 28;2(4):1–5.

33. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Current Opinion in Microbiology. 2008 Oct 1;11(5):472–7.

34. cgMLST.org Nomenclature Server (h25) [Internet]. [cited 2024 Jan 22]. Available from: https://www.cgmlst.org/ncs

35. Fang Y, Quan J, Hua X, Feng Y, Li X, Wang J, et al. Complete genome sequence of Acinetobacter baumannii XH386 (ST208), a multi-drug resistant bacteria isolated from pediatric hospital in China. Genom Data. 2015 Dec 19;7:269–74.

36. Min B, Yoo D, Lee Y, Seo M, Kim H. Complete Genomic Analysis of Enterococcus faecium Heat-Resistant Strain Developed by Two-Step Adaptation Laboratory Evolution Method. Frontiers in Bioengineering and Biotechnology [Internet]. 2020 [cited 2024 Jan 22];8. Available from: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00828

37. Serres MH, Gopal S, Nahum LA, Liang P, Gaasterland T, Riley M. A functional update of the Escherichia coli K-12 genome. Genome Biol. 2001;2(9):research0035.1-research0035.7.

38. Rocha J, Henriques I, Gomila M, Manaia CM. Common and distinctive genomic features of Klebsiella pneumoniae thriving in the natural environment or in clinical settings. Sci Rep. 2022 Jun 21;12:10441.

39. Subedi D, Vijay AK, Kohli GS, Rice SA, Willcox M. Comparative genomics of clinical strains of Pseudomonas aeruginosa strains isolated from different geographic sites. Sci Rep. 2018 Oct 23;8(1):15668.

40. Sivakumar R, Pranav PS, Annamanedi M, Chandrapriya S, Isloor S, Rajendhran J, et al. Genome sequencing and comparative genomic analysis of bovine mastitis-associated Staphylococcus aureus strains from India. BMC Genomics. 2023 Jan 25;24(1):44.

41. Martinez JL. General principles of antibiotic resistance in bacteria. Drug Discovery Today: Technologies. 2014 Mar 1;11:33–9.

42. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol. 2005 Sep;3(9):722–32.

43. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile Genetic Elements Associated with Antimicrobial Resistance. Clinical Microbiology Reviews. 2018 Aug;31(4):10.1128/cmr.00088-17.

44. Introduction to Storing Genetic Information | Biology for Non-Majors I [Internet]. [cited 2024 Jan 15]. Available from: https://courses.lumenlearning.com/wm-nmbiology1/chapter/outcome-storing-genetic-information/

45. Cell nuclei contain chromosomes number changes per species [Internet]. [cited 2024 Jan 15]. Available from: https://mammothmemory.net/biology/dna-genetics-and-inheritance/dna-base-pairing/dna-in-cells.html

46. Science Learning Hub [Internet]. [cited 2024 Jan 15]. Bacterial DNA – the role of plasmids. Available from: https://www.sciencelearn.org.nz/resources/1900-bacterial-dna-the-role-of-plasmids

47. Linnarsson S. Recent advances in DNA sequencing methods – general principles of sample preparation. Experimental Cell Research. 2010 May 1;316(8):1339–43.

48. Kashima Y, Sakamoto Y, Kaneko K, Seki M, Suzuki Y, Suzuki A. Single-cell sequencing techniques from individual to multiomics analyses. Exp Mol Med. 2020 Sep;52(9):1419–27.

49. Andersen SB, Schluter J. A metagenomics approach to investigate microbiome sociobiology. Proceedings of the National Academy of Sciences. 2021 Mar 9;118(10):e2100934118.

50. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Research. 2011;39(13):13.

51. Bharti R, Grimm DG. Current challenges and best-practice protocols for microbiome analysis. Brief Bioinform. 2019 Dec 18;22(1):178–93.

52. Deamer D, Nichols J. Proton flux mechanisms in model and biological membranes. The Journal of membrane biology. 1989 Mar 1;107:91–103.

53. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, et al. The potential and challenges of nanopore sequencing. Nat Biotechnol. 2008 Oct;26(10):1146–53.

54. Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clinical Microbiology and Infection. 2018 Apr 1;24(4):335–41.

55. Udaondo Z, Sittikankaew K, Uengwetwanit T, Wongsurawat T, Sonthirod C, Jenjaroenpun P, et al. Comparative Analysis of PacBio and Oxford Nanopore Sequencing Technologies for Transcriptomic Landscape Identification of Penaeus monodon. Life (Basel). 2021 Aug 23;11(8):862.

56. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010 Jun 1;95(6):315–27.

57. Pop M. Genome assembly reborn: recent computational challenges. Briefings in Bioinformatics. 2009 Jul 1;10(4):354–66.

58. Sequencing from scratch: reference genomes and de novo sequence assembly – HudsonAlpha Institute for Biotechnology [Internet]. [cited 2024 Jan 15]. Available from: https://www.hudsonalpha.org/sequencing-from-scratch-reference-genomes-and-de-novo-sequence-assembly/

59. Cufflinks H. Reference based assembly - ppt download [Internet]. [cited 2024 Jan 15]. Available from: https://slideplayer.com/slide/14753621/

60. Compeau PEC, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. 2011 Nov;29(11):987–91.

61. Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, et al. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. Briefings in Functional Genomics. 2012 Jan 1;11(1):25–37.

62. Rizzi R, Beretta S, Patterson M, Pirola Y, Previtali M, Della Vedova G, et al. Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era. Quant Biol. 2019 Dec 1;7(4):278–92.

63. Goodman S, Hedetniemi S. Eulerian Walks in Graphs. SIAM J Comput. 1973 Mar;2(1):16–27.

64. Kapun E, Tsarev F. De Bruijn Superwalk with Multiplicities Problem is NP-hard. BMC Bioinformatics. 2013 Apr 10;14(5):S7.

65. Leggett RM, MacLean D. Reference-free SNP detection: dealing with the data deluge. BMC

Genomics. 2014 May 20;15(4):S10.

66. Caboche S, Audebert C, Lemoine Y, Hot D. Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. BMC Genomics. 2014 Apr 5;15(1):264.

67. Ahmed N, Bertels K, Al-Ars Z. A comparison of seed-and-extend techniques in modern DNA read alignment algorithms. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) [Internet]. 2016 [cited 2024 Apr 2]. p. 1421–8. Available from: https://ieeexplore.ieee.org/abstract/document/7822731?casa_token=FY4BKER6uVAAAAAA:HhYYWlPJ2Q9c3Ksv-2_U-_3J_daiZri8fZQqg9F_LX8kHRmmz2ODv_InwBaSOEbe1W_j4tFW0Ds

68. Ferragina P. Burrows-Wheeler Transform. In: Kao MY, editor. Encyclopedia of Algorithms [Internet]. Boston, MA: Springer US; 2008 [cited 2024 Jan 15]. p. 1–7. Available from: https://doi.org/10.1007/978-3-642-27848-8_59-2

69. Roberts M, Hunt BR, Yorke JA, Bolanos RA, Delcher AL. A Preprocessor for Shotgun Assembly of Large Genomes. Journal of Computational Biology. 2004 Aug;11(4):734–52.

70. Wolff J, Batut B, Rasche H. Galaxy Training Network. Galaxy Training Network; [cited 2024 Jan 15]. Hands-on: Hands-on: Mapping. Available from: https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html

71. Li B, Chen W, Zhan X, Busonero F, Sanna S, Sidore C, et al. A Likelihood-Based Framework for Variant Calling and De Novo Mutation Detection in Families. PLOS Genetics. 2012 Oct 4;8(10):e1002944.

72. Supernat A, Vidarsson OV, Steen VM, Stokowy T. Comparison of three variant callers for human whole genome sequencing. Sci Rep. 2018 Dec 14;8(1):17851.

73. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018 Nov;36(10):983–7.

74. Aminov RI, Mackie RI. Evolution and ecology of antibiotic resistance genes. FEMS Microbiology Letters. 2007 Jun 1;271(2):147–61.

75. Cetinkaya Y, Falk P, Mayhall CG. Vancomycin-Resistant Enterococci. Clinical Microbiology Reviews. 2000 Oct;13(4):686–707.

76. Lee AS, de Lencastre H, Garau J, Kluytmans J, Malhotra-Kumar S, Peschel A, et al. Methicillin-resistant Staphylococcus aureus. Nat Rev Dis Primers. 2018 May 31;4(1):1–23.

77. Savov E, Kjoseva E, Borisova N, Gergova I, Ronkova G, Trifonova A. IN VITRO STUDY OF THE RESISTANCE OF PROBLEMATIC FOR HOSPITAL INFECTIOUS PATHOLOGY MICROORGANISMS TO ANTIMICROBIAL DRUGS. Trakia Journal of Sciences. 2010;8(2).

78. Reygaert WC. An overview of the antimicrobial resistance mechanisms of bacteria. AIMS Microbiol. 2018 Jun 26;4(3):482–501.

79. Kim DH, Ko KS. A distinct alleles and genetic recombination of pmrCAB operon in species of Acinetobacter baumannii complex isolates. Diagnostic Microbiology and Infectious Disease. 2015 Jul 1;82(3):183–8.

80. Lesho E, Yoon EJ, McGann P, Snesrud E, Kwak Y, Milillo M, et al. Emergence of Colistin-Resistance in Extremely Drug-Resistant Acinetobacter baumannii Containing a Novel pmrCAB Operon During Colistin Therapy of Wound Infections. The Journal of Infectious Diseases. 2013 Oct 1;208(7):1142–51.

81. Higgins PG, Dammhayn C, Hackel M, Seifert H. Global spread of carbapenem-resistant Acinetobacter baumannii. J Antimicrob Chemother. 2010 Feb;65(2):233–8.

82. Hitchcock P, Chamberlain A, Van Wagoner M, Inglesby TV, O'Toole T. Challenges to Global Surveillance and Response to Infectious Disease Outbreaks of International Importance. Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science. 2007 Sep;5(3):206–27.

83. Gilchrist CA, Turner SD, Riley MF, Petri WA, Hewlett EL. Whole-Genome Sequencing in Outbreak Analysis. Clinical Microbiology Reviews. 2015 Apr 15;28(3):541–63.

84. Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. The Lancet Infectious Diseases. 2020 Nov 1;20(11):1263–71.

85. Harris SR, Cartwright EJ, Török ME, Holden MT, Brown NM, Ogilvy-Stuart AL, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. The Lancet Infectious Diseases. 2013 Feb 1;13(2):130–6.

86. Ypma RJF, van Ballegooijen WM, Wallinga J. Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks. Genetics. 2013 Nov 1;195(3):1055–62.

87. Maiden MCJ, van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013 Oct;11(10):728–36.

88. Deneke C, Uelze L, Brendebach H, Tausch SH, Malorny B. Decentralized Investigation of Bacterial Outbreaks Based on Hashed cgMLST. Frontiers in Microbiology [Internet]. 2021 [cited 2024 Jan 16];12. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2021.649517

89. Gunderson KL, Steemers FJ, Ren H, Ng P, Zhou L, Tsan C, et al. Whole-Genome Genotyping. In: Methods in Enzymology [Internet]. Academic Press; 2006 [cited 2024 Jan 16]. p. 359–76. (DNA Microarrays, Part A: Array Platforms and Wet-Bench Protocols; vol. 410). Available from: https://www.sciencedirect.com/science/article/pii/S0076687906100178

90. Miro E, Rossen JWA, Chlebowicz MA, Harmsen D, Brisse S, Passet V, et al. Core/Whole Genome Multilocus Sequence Typing and Core Genome SNP-Based Typing of OXA-48-Producing Klebsiella pneumoniae Clinical Isolates From Spain. Front Microbiol [Internet]. 2020 Jan 31 [cited 2024 Apr 2];10. Available from: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2019.02961/full

91. Pearce ME, Alikhan NF, Dallman TJ, Zhou Z, Grant K, Maiden MCJ. Comparative analysis of core genome MLST and SNP typing within a European *Salmonella* serovar Enteritidis outbreak. International Journal of Food Microbiology. 2018 Jun 2;274:1–11.

92. Cvijović I, Good BH, Desai MM. The Effect of Strong Purifying Selection on Genetic Diversity. Genetics. 2018 Aug 1;209(4):1235–78.

93. Colquhoun RM, Hall MB, Lima L, Roberts LW, Malone KM, Hunt M, et al. Pandora: nucleotide-resolution bacterial pan-genomics with reference graphs. Genome Biol. 2021 Sep 14;22(1):267.

94. Araz OM, Ramirez-Nafarrate A, Jehn M, Wilson FA. The importance of widespread testing for COVID-19 pandemic: systems thinking for drive-through testing sites. Health Systems. 2020 Apr 2;9(2):119–23.

95. Houwaart T, Belhaj S, Tawalbeh E, Nagels D, Fröhlich Y, Finzer P, et al. Integrated genomic surveillance enables tracing of person-to-person SARS-CoV-2 transmission chains during community transmission and reveals extensive onward transmission of travel-imported infections, Germany, June to July 2021. Euro Surveill. 2022 Oct;27(43):2101089.

96. Abdin AF, Fang YP, Caunhye A, Alem D, Barros A, Zio E. An optimization model for planning testing and control strategies to limit the spread of a pandemic – The case of COVID-19. European Journal of Operational Research. 2023 Jan 1;304(1):308–24.