Nanopore sequencing for accurate bacterial outbreak tracing

Mara Lohde[1], Gabriel E. Wagner[5], Johanna Dabernig-Heinz[5], Adrian Viehweger[4], Claudia Stein[1], Sascha D. Braun[2,3], Stefan Monecke[2,3], Mike Marquet[1], Ralf Ehricht[2,3,6], Mathias W. Pletz[1,2] & Christian Brandt[1,2]

[1] Institute for Infectious Diseases and Infection Control, Jena University Hospital, Jena, Germany
[2] InfectoGnostics Research Campus, Centre for Applied Research, Jena, Germany
[3] Leibniz-Institute of Photonic Technology (Leibniz-IPHT), Jena, Germany
[4] Institute of Medical Microbiology and Virology, University Hospital Leipzig, Leipzig, Germany
[5] Diagnostic and Research Institute of Hygiene, Microbiology and Environmental Medicine, Medical University of Graz, Graz, Austria
[6] Institute of Physical Chemistry, Friedrich-Schiller-University Jena, Jena, Germany

Keywords: Oxford Nanopore Technologies, Outbreak Tracing, cgMLST, Phylogeny

Abstract

Our study investigated the effectiveness of Oxford Nanopore Technology for accurate outbreak tracing by resequencing a three-year-long *Klebsiella pneumoniae* outbreak with Illumina short-read sequencing data as the point of reference. We detected considerable base errors through cgMLST and phylogenetic analysis of genomes sequenced with Nanopore compared to the Illumina data, leading to the false exclusion of some outbreak-related strains from the outbreak cluster. Nearby methylation sites cause these errors and can also be found in other species besides *K. pneumoniae*. Based on this data, we explored PCR-based sequencing and a masking strategy, which both successfully addressed these inaccuracies and ensured accurate outbreak tracing. Additionally, we offer a bioinformatic workflow to identify and mask problematic genome positions in a reference-free manner. Without further technological developments, our study recommends

PCR-based sequencing for outbreak tracing to avoid spurious base calls from Nanopore data.

Background/Introduction

Whole genome sequencing is essential for analysing outbreaks, pandemics, or phylogenetic relationships [1], [2]. The recent SARS-CoV-2 pandemic has thus led to a leap in the integration and expansion of sequencing capacities in many laboratories and hospitals, predominantly using Illumina for short-read sequencing or Oxford Nanopore Technologies for long-read sequencing (approx. 78% and 18%, respectively) [3]. Beyond viral pandemic tracking, bacterial pathogen outbreaks, particularly those linked to antibiotic resistance, continue to impose a significant global public health burden. Gram-negative bacteria, in particular, rapidly acquire antibiotic resistance *via* horizontal gene transfer from other species [4]–[6]. This mechanism complicates tracking outbreaks or identifying their origin, as a single specific plasmid or mobile element can be responsible for a persistent outbreak or multiple outbreaks across unrelated species [5], [7]–[9].

Effectively tracking these complex molecular mechanisms requires careful strategic monitoring and sequencing-based investigation. Consequently, the accuracy and continuity of the genome data is paramount. Illumina, a short-read sequencing method with an error rate of less than 0.8% in raw data, is frequently used as its complementary genome reconstruction precision exceeds 99.997% [10]. However, repetitive elements, such as transposons, present a substantial challenge for short reads when reconstructing closed bacterial genomes and their accompanying plasmids. Long-read sequencing technologies like Pacific Bioscience (PacBio) and Oxford Nanopore Technologies (ONT) can resolve such elements, e.g. plasmids, as they achieve longer read lengths averaging around 10-20 kb and even up to 3.85 Mb in the case of ONT [11]–[14].

Real-time sequencing allows data collection and analysis, while sequencing positions Oxford Nanopore Technologies as an appealing choice for hospital surveillance and outbreak

control [15]. Owing to their recently launched Flowcells (R10.4.1) and Chemistry (SQK-NBD114.24), they have achieved raw read accuracy that now exceeds 99.1% [16]. Several studies have shared their findings and reported accuracy levels similar to those from short-read data [17], [18]. However, significant discrepancies between Illumina and Nanopore genomes were also observed for some organisms [19].

When investigating outbreaks, these contradictions can lead to inaccurate conclusions. In addition, genomes are usually stored in open public databases such as NCBI or ENA, which other scientists use as references for their work. Therefore, we used ONT to reevaluate a well-documented, three-year-long outbreak initially analysed with Illumina data to address these contradictory statements [20]. *K. pneumonia* is an ideal microorganism for this topic, as it is a common pathogen linked to hospital-wide outbreaks carrying plasmids with multidrug resistance genes [21]. When using ONT-only data, we identified a few critical issues leading to false basecalls for *K. pneumonia*. We noticed similar problems and clear patterns in other organisms, which need to be considered during outbreak identification, even though we could resolve them.

Results

Erroneous basecalls occur in some strains but not others and vary by basecaller and sequencing kits

We resequenced the genomes of 33 randomly selected *K. pneumonia* samples (from a total of 114 outbreak-related isolates) using R10.4 and R10.4.1 flowcells, along with the corresponding library preparation kits (henceforth “Kit 12”: SQK-NBD112.24 (early access) and “Kit 14”: SQK-NBD114.24 (successor)), to examine if the previously documented conflicting statements could be replicated [17]–[19]. We used cgMLST analysis to compare the sequenced genomes against Illumina data. The comparison revealed 11 outliers,

showing high deviations to Illumina genomes and not matching the outbreak cluster out of 33 samples in the Nanopore data (Supplementary Figure 1). While the remaining samples closely matched the Illumina genomes, the outliers fit the inconsistencies between ONT and Illumina, as reported in some literature. To assess whether either the basecaller or their models might be responsible, we re-basecalled and compared an outlier sample (UR2602) in detail to three samples, where we assume error-free genomes since matching Illumina genomes in cgMLST typing. To investigate the influence of Kit 14 and Kit 12, their associated flowcells, different basecallers (Guppy and Dorado) and model combinations on the basecalling error(Figure 1; see method section “Basecalling and Assembly” for further details).

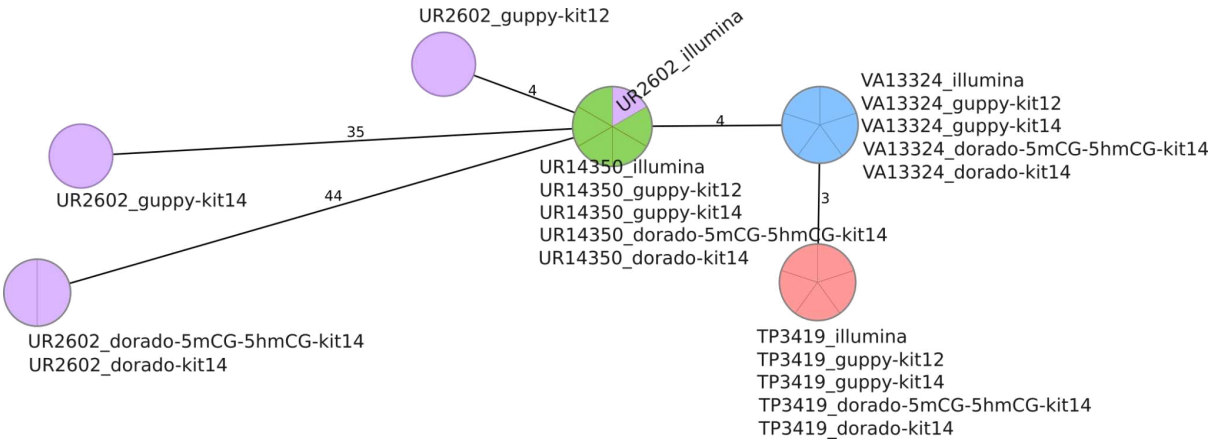


Figure 1: CgMLST typing reveals allelic differences between genomes utilising different Basecaller and Sequencing Kits. The Minimum spanning tree pictures four *K. pneumonia* samples based on 2365 loci to compare the allelic variations. Nodes (samples) are connected by lines depicting the distance by numbers of allelic differences. Loci are considered different if one or more bases change between the samples. Loci without allelic differences are described as being the same. All isolates were prepared with Kit 14 and Kit 12 and basecalled with each respective Guppy “super accurate” basecalling model (see methods “Basecalling and Assembly”). We basecalled all Kit 14 - prepared samples with Dorado using the default and a modification-aware model (see methods “Basecalling and Assembly”).

Based on 2359 loci for the cgMLST, no allelic differences, regardless of kit, basecaller or basecaller model, were identified for the isolates UR144350, VA13324 and TP3419. In contrast, the outlier sample (UR2602) revealed allelic variations to Illumina for each kit and basecaller. Despite both basecaller using the same raw signal data, 35 allelic differences by Guppy and 44 by Dorado without accordances were detected.

By cgMLST typing, the outlier sample prepared with Kit 14 would not be included as an outbreak isolate due to its 35 or 44 allelic differences, even though it is part of the outbreak. Conversely, when prepared with early access Kit 12, the same isolate would be considered as only four loci could be observed (adhering to the recommended allelic difference cutoff of 15). Since the basecalling models disagreed on the allelic differences, we suspected more issues within the raw data (reads and raw signals) and conducted a comprehensive analysis of all possible affected positions.

Ambiguities in purine or pyrimidine discrimination for a subset of genome positions can cause erroneous basecalls

The first visual inspection of mapped reads to the assembly revealed ambiguous positions with varying base ratios. For further characterisation of these positions, we examine our data on the sequence, nucleotide and raw signal level (Figure 2). For each ambiguous position on the chromosomal DNA for 33 *K. pneumoniae* outbreak samples, we determined the ratio between the two bases by counting their occurrences within the read data at that position (Figure 2 A). Searching for characteristic “indicator” sequence motifs, we explored the surrounding base for each detected ambiguous position and plotted the observed pattern as a sequence logo (Figure 2 B and Supplementary Table 1). Additionally, we compared the methylated and unmethylated raw signals around these ambiguous positions (Figure 2 C).

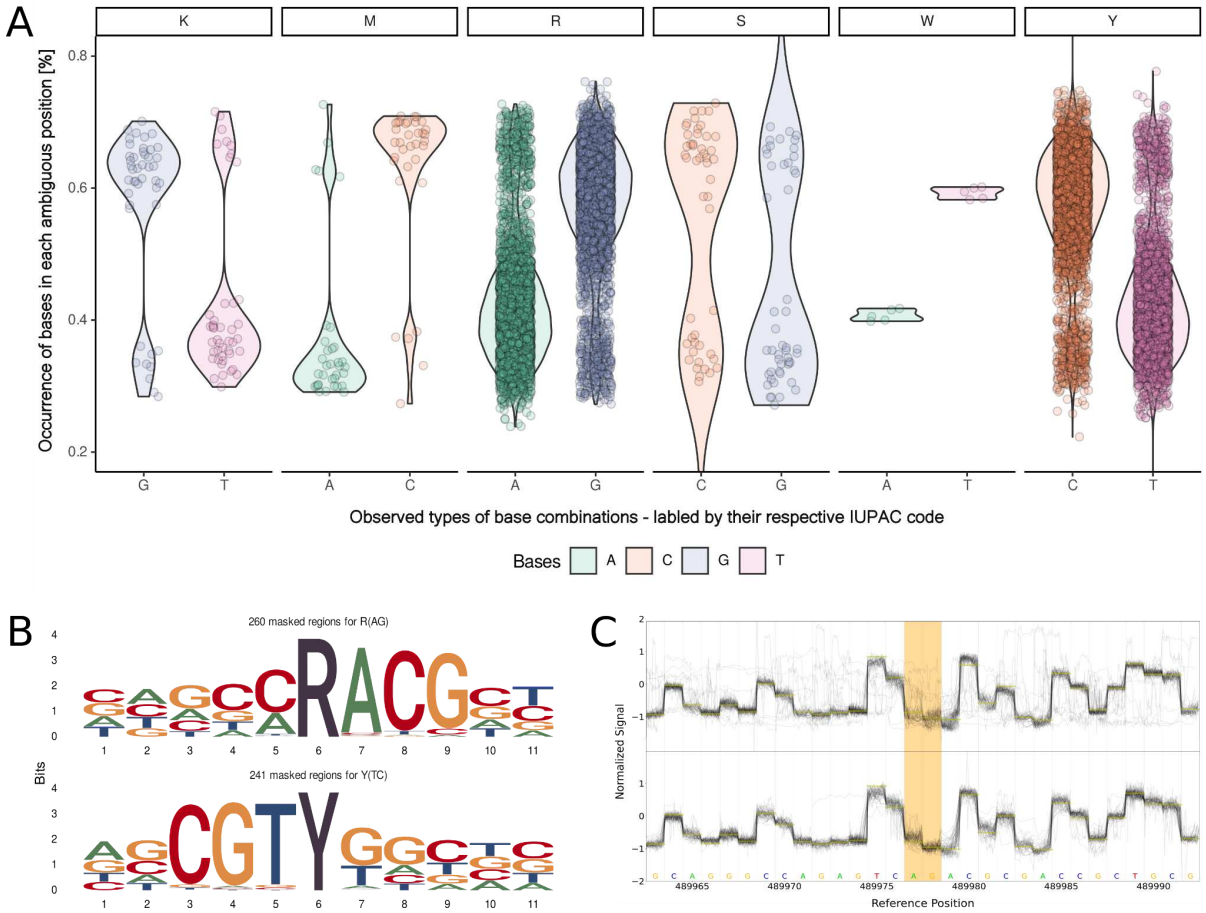


Figure 2: Overall investigation for ambiguous positions. A: Violin chart showing the ratio between two bases within the read data for 6,556 ambiguous positions in 33 *K. pneumonia* samples. Every ambiguous position is distinguished by which two bases appear and labelled by their respective degenerative base (IUPAC nucleotide code). For example, "R" stands for a combination where either A or G is found at that position. Each dot represents a base occurrence within the respective base combination at the ambiguous position. **B:** Sequence logo of observed sequence pattern around the ambiguous bases R and Y on the chromosomal contig of *K. pneumonia* for one sample. **C:** Raw signal level (fast5/pod5) of ambiguous positions (yellow) for Kit 14 (above) with methylated bases and SQK-RPB114.24 without modifications (below). Less clear signals are observable in ambiguous positions (yellow) for Kit 14. Signal plots were generated with remora (v.2.1.3; github.com/nanoporetech/remora).

As highlighted in Figure 2 A, in some positions, the basecaller can not determine between either of two bases, expressed by specific base ratios varying per position and resulting in erroneous assemblies. We could not observe ambiguous positions containing three different bases. For clarity, we assign the IPUAC nucleotide code for degenerate bases (K, M, R, S, W, Y) to each ambiguous position varying between two bases. Accordingly, we will refer, e.g. to "R" when the positions contain A or G in the read data.

Out of our 33 *K. pneumonia* outbreak isolates analysis, we have discovered 6,556 positions that exhibit ambiguity (Figure 2A). The ambiguity mainly resolved around 3,311 positions for R and 3,111 for Y. In 5442 of 6455 R and Y positions (84,31%), the basecalled reads lean towards the stronger base (C or G). We detected other ambiguous positions in K (44), M (34), S (51) and W (5), but with comparable lower occurences. It is essential to acknowledge that not all identified ambiguous positions result in errors in the final genome, which explains the varying error profile of the same sample. Errors in these ambiguous positions mainly arise when deciding between purine bases (A or G) or pyrimidine bases (T or C).

In the error-prone genomes of the *K. pneumoniae* outbreak, we detected preserved patterns around the ambiguous positions R and Y (Figure 2B). These sequence motifs are reverse-complement patterns (RACG/CGTY), pointing to a singular issue. Compared to other isolates of *K. pneumonia*, we also observed additional patterns. These motifs are likely specific to particular strains.

Furthermore, we examined and collected additional genomes that utilised the Kit 14 library preparation for sequencing (264 isolates across 32 species) to investigate whether the ambiguous positions are *K. pneumoniae* exclusive (Table 1). We determine the fewest ambiguous positions (0 to 1) in *B. pertussis* compared to all other species samples. In contrast, all 10 *Enterococcus faecalis* isolates had over 200 ambiguous positions. If we look at all isolates, over 40% of 264 screened samples have more than 50 ambiguous positions. Across all species, the most minor shared sequence motif was at least RA/TY. Certain

species, such as *Acinetobacter junii*, *Acinetobacter radioresistens*, *Chryseobacterium gleum*, *Enterobacter cloacae*, *Micrococcus luteus* and *Stenotrophomonas maltophilia*, exhibited a considerable number of ambiguous positions. This suggests that many species may be impacted, but not necessarily within all strains.

As methylated bases are probably liable for ambiguous positions, we compared sequencing data with methylations (Kit 14) (Figure 2 C above) and without (SQK-RPB114.24) (Figure 2 C below) on the raw signal level from fast5/pod5 files before basecalling occurs. For native sequencing, less clear signals at these positions are observable, which might cause these ambiguous basecalls. These noisy signals could explain the frequencies of bases we detected (Figure 2A) and, thus, the basecaller's difficulty in deciding on a specific base for that position.

We found no coherent methylation motifs in the literature that would fit the observed pattern. Nevertheless, it has been reported that methylated bases can affect the raw signal in the surrounding region [22]. Thus, we can not determine whether multiple methylation motifs are the cause or if an unknown motif is present. Accordingly, to these findings, we evaluated whether PCR-based sequencing or a bioinformatic masking strategy for ambiguous positions can reliably remove these errors for outbreak analysis.

Table 1: Overview of R (A-G) and Y (T-C) base ambiguity for 264 isolates from 32 various species, sequenced with Oxford Nanopore Technologies using Kit 14. Only chromosomal contigs were analysed, and only super accurate basecalling models were used. Genomes were coverage masked by N if below a read depth of 10x. N positions were not considered for the table to avoid overestimating one base ambiguity.

Species	Total samples	Samples with > 50 ambig. P.	Mean R (A-G) ambiguity per sample (min/max)	Mean Y (T-C) ambiguity per sample (min/max)	Motif type	Reference
Achromobacter xylosoxidans	1	0 (0%)	6	4	NA	*)
Acinetobacter baumannii	15	5 (33,33%)	22,53 (0/109)	19,87 (0/83)	NA	*)
Acinetobacter junii	1	1 (100%)	279	223	CA**R**ATG CAT**Y**TG	*)
Acinetobacter mesopotamicus	1	1 (100%)	53	28	NA	*)
Acinetobacter radioresistens	1	1 (100%)	175	150	**R**A T**Y**	*)
Acinetobacter soli	1	0 (0%)	12	11	NA	*)
Bordetella pertussis	40	0 (0%)	0,1 (0/1)	0,2 (0/1)	NA	[17]
Chryseobacterium arthrosphaerae	1	0 (0%)	13	7	NA	*)
Chryseobacterium gleum	1	1 (100%)	218	217	**R**ACGC GCGT**Y**	*)
Citrobacter freundii	3	3 (100%)	145,67 (49/329)	137,67 (39/319)	C**R**ATGTC GACAT**Y**G	*)

Citrobacter portucalensis	2	2 (100%)	29 (26/32)	29 (28/30)	**R**A T**Y**	*)
Enterobacter cloacae	1	1 (100%)	153	162	NA	*)
Enterobacter hormaechei	5	1 (20%)	15,60 (4/45)	15,60 (6/42)	NA	*)
Enterococcus faecalis	10	10 (100%)	250,40 (223/275)	249,00 (210/270)	T**R**AG CT**Y**A	+)
Enterococcus faecium	19	2 (10,53%)	15,74 (0/27)	14,63 (0/28)	**R**ACC GGT**Y**	#)
Escherichia coli	8	3 (37,50%)	31,38 (2/118)	29,31 (4/111)	NA	*)
Escherichia flexneri	10	9 (90%)	49,10 (19/88)	44,70 (18/88)	**R**AT AT**Y**	*)
Klebsiella aerogenes	1	0 (0%)	22	17	NA	*)
Klebsiella michiganensis	1	1 (100%)	56	42	NA	*)
Klebsiella pneumonia	70	38 (54,29%)	97,04 (3/835)	92,47 (3/847)	**R**ACG CGT**Y**	[20] *) #)
Klebsiella oxytoca	1	0 (0%)	5	9	NA	*)
Listeria monocytogenes	17	3 (17,65%)	37,94 (0/514)	40,82 (0/557)	NA	#)
Micrococcus luteus	1	1 (100%)	172	220	C**R**AC GT**Y**G	*)

Proteus mirabilis	1	1 (100%)	45	43	C**R**AC GT**Y**G	*)
Pseudomonas aeruginosa	19	6 (33,33%)	389,17 (0/2251)	387,56 (0/2290)	AA**R**ACC GGT**Y**TT	*)
Pseudomonas asiatica	2	1 (50%)	145 (0/290)	144,50 (0/289)	CC**R**A T**Y**GG	*)
Pseudomonas stutzeri	1	0 (0%)	14	16	NA	*)
Salmonella enterica	2	1 (50%)	41,5 (7/76)	42,5 (7/78)	NA	*)
Stenotrophomonas maltophilia	1	1 (100%)	229	171	TAC**R**AC GT**Y**GTA	*)
Serratia marcescens	4	4 (100%)	107,75 (68/178)	99,5 (67/171)	CC**R**A T**Y**GG	*)
Shewanella algae	2	2 (100%)	71,5 (37/106)	50 (32/68)	NA	*)
Staphylococcus aureus	20	8 (40%)	23,35 (0/99)	23,35 (0/97)	**R**ACC GGT**Y**	#)

*) Sequenced strains received from Leibniz-Institute of Photonic Technology, Optisch-Molekulare Diagnostik und Systemtechnologie

#) Diagnostic and Research Institute of Hygiene, Microbiology and Environmental Medicine, Medical University of Graz

+) Own samples from the Jena University Hospital

Strategies to mitigate methylation-induced basecalling errors

To solve methylation-induced basecalling errors in ambiguous base positions, we evaluated two strategies: 1) We resequenced eight *K. pneumoniae* outbreak samples using the Nanopore Rapid PCR Barcoding Kit (SQK-RPB114.24) to remove methylated bases prior to sequencing and analysed the genomes using cgMLST typing and phylogenetic analysis (Figure 3 A and B). 2) We masked ambiguous positions for Kit 14 prepared genomes with our bioinformatic workflow (see method section “Workflow for detecting and masking of ambiguous positions”). It is important to mention that these masked assemblies cannot be used for cgMLST analysis because allelic differences cannot be accurately determined for genes with masked bases. Therefore, the masked genomes were only used for phylogenetic analysis (Figure 3 B).

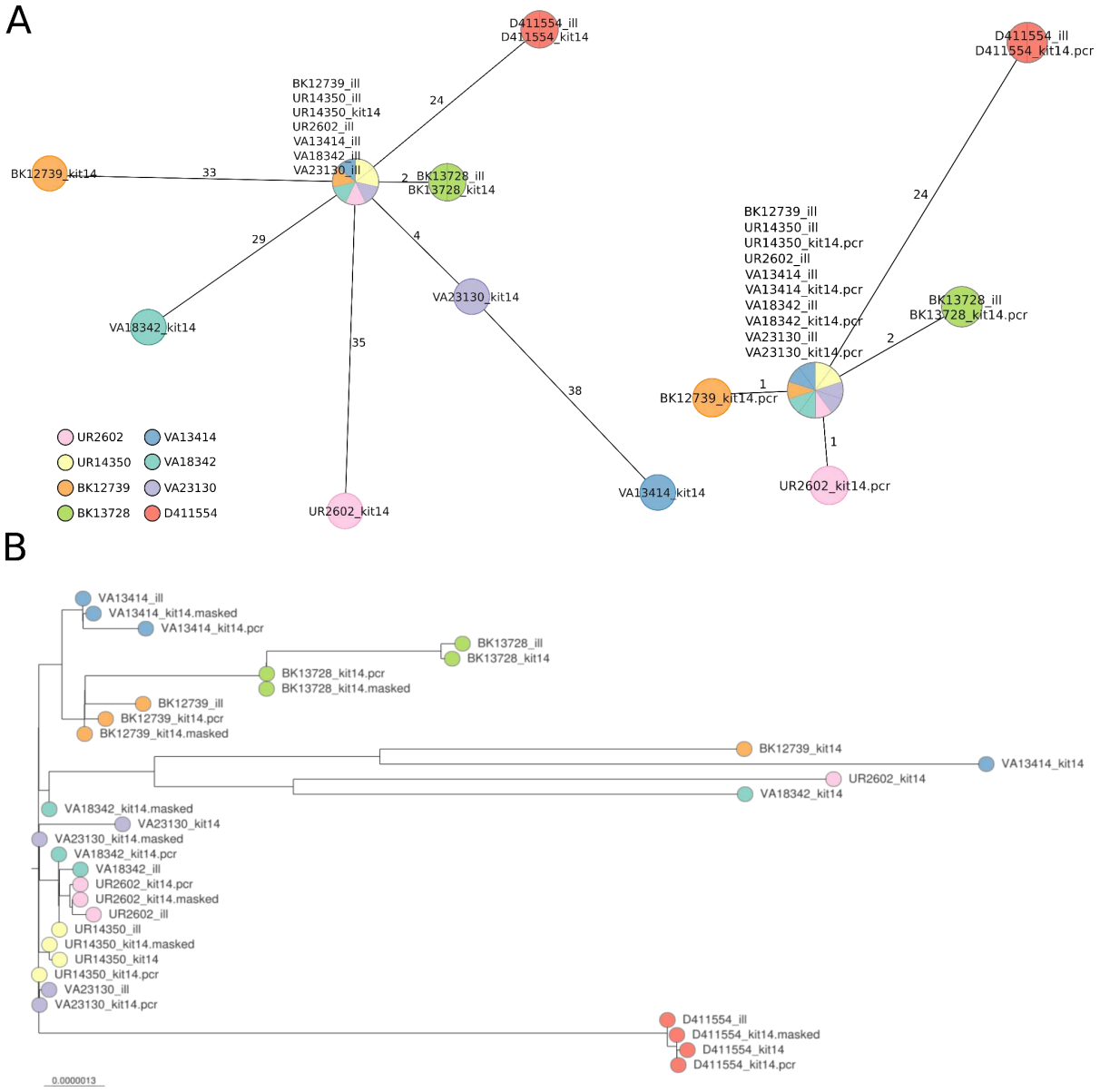


Figure 3: PCR-based sequencing or masking of ambiguous positions reduces allelic or phylogenetic distances between Illumina and nanopore genomes. A: Minimum spanning trees of each of eight *K. pneumoniae* outbreak samples based on 2365 loci to compare the allelic variations between Illumina genomes to Nanopore SQK-NBD114.24 (kit14; left) and SQK-RPB114.24 (pcr; right) showing a reduction in allelic differences. Nodes (samples) are connected by lines depicting the distance by numbers of allelic differences. Loci are considered different whether one or more bases change between the samples. Loci without allelic differences are described as being the same. **B:** Phylogenetic tree to figure the genetic distances between eight *K. pneumoniae* outbreak samples (coloured nodes),

prepared with Illumina (ill), Nanopore SQK-NBD114.24 (kit14) and SQK-RPB114.24 (pcr) compared to the masked Kit 14 assemblies (masked).

When comparing the Native Barcoding with the PCR-based Kit, the genome assemblies significantly reduced ambiguous positions from 2,316 to just 14 for R & Y across the eight resequenced *K. pneumoniae* samples (Supplementary Table 2). Both the minimal spanning trees and the phylogenetic tree also show this significant improvement in genome quality for Nanopore (Figure 3). According to the cgMLST typing, the outlier samples UR2602 and BK12739 now closely match the Illumina genome, down to only one allele difference from 35/33 (Figure 3 A). When comparing phylogenetic distances within the phylogenetic tree, an increased convergence with the Ilumina genomes, particularly for the outlier samples, was observed, too (Figure 3 B). Additionally, masked and PCR-based assemblies have almost no phylogenetic divergence.

Further, we analysed the phylogenetic tree containing native Kit 14, masked native Kit 14, and Illumina genomes for all 33 *K. pneumoniae* samples (Supplementary Figure 2). These include 11 Kit 14 outliers (average of 492 ambiguous positions) and 22 Kit 14 genomes with an average of less than 52 ambiguous positions. We observed two types of phylogenic distances between Nanopore Kit 14 and Illumina: The expected considerable distances between the outlier and Illumina genomes are due to ambiguity and, in some cases, a phylogenetic distance for which ambiguity is not the causation.

By masking ambiguous bases, we observed that 8 of 11 outlier genomes now closely align with their respective Illumina genome. The remaining three outlier samples changed their tree positions after masking, now closely aligning with other Illumina genomes but still diverged from their corresponding Illumina genome due to other non-ambiguity-related differences. For the other 22 masked Kit 14 genomes with less ambiguity than the outliers, we did not observe any substantial changes in their tree positions, as fewer positions were masked.

In summary, 22 out of 33 masked Nanopore genomes align with their respective Illumina genomes, and the remaining 11 do not. In these cases, the remaining distances do not result from ambiguous positions within the Nanopore assemblies. For instance, the Illumina and Nanopore genomes of TP3870 matched perfectly in the minimal spanning tree, but they exhibited some distance from each other in the phylogenetic tree (Supplementary Figure 2). We identified reconstruction issues in these short-read assemblies, primarily manifesting in non-coding regions. Since cgMLST typing is performed comparing coding sequences only, these errors do not affect the result analysis. Therefore, we recommend using only one technology when performing whole genome comparison for outbreak analysis.

Discussion

Over the past few years, Oxford Nanopore Technologies has been effectively used to monitor and track the SARS-CoV-2 pandemic and its viral lineages. Despite this, contradictory reports have emerged regarding the consistency of Nanopore-sequenced bacterial genomes compared to Illumina-based. Our research examined whether Oxford Nanopore Technology could accurately analyse bacterial outbreaks.

For our investigation, we resequenced a well-documented 3-year *K. pneumoniae* outbreak using the Nanopore Native Barcoding Kit 14 for Library preparation. Our analyses demonstrated that the raw signals were impacted by methylated bases, creating ambiguous positions through basecalling and leading to erroneous exclusions of certain outbreak-associated strains. Despite focusing on *K. pneumoniae* initially, other prokaryotic organisms are also impacted.

Based on our in-depth investigation, we recommend using the Nanopore Rapid PCR Barcoding Kit for sequencing to eliminate these ambiguities in the final genome assemblies. However, this method decreases the read length to roughly 3,500 bp, posing difficulties in achieving closed plasmids and genomes, similar to other short-read approaches but to a

way lesser extent. To obtain error-free and closed genomes, we suggest utilising both sequencing kits for library preparation and pooling their libraries in approximately 30/70 ratio (native to PCR-based) prior to flowcell loading. This should offset the imbalances in ambiguous positions while maintaining reasonable cost-effectiveness. For samples already sequenced without any involvement of PCR, we propose using the provided MPOA workflow (https://github.com/replikation/MPOA) to assess the quality of each genome. This workflow offers information about the frequency of ambiguous positions and masks them without needing another reference. Though these masked assemblies cannot be used for cgMLST, they remain suitable for constructing phylogenetic trees for outbreak tracking.

Given the notable strides made in direct methylation calling techniques, Oxford Nanopore Technology might overcome the issues with ambiguous positions. If available in high enough quantities, Duplex reads (connecting and sequencing both strands) might provide better raw signal data for accurate basecalling. Nevertheless, we encourage careful testing and reevaluation of sequencing chemistry and basecalling algorithms with more prokaryotic samples to avoid erroneous conclusions based on these ambiguous positions.

Conclusions

Our research highlights the drawback of using Oxford Nanopore Technologies for sequencing prokaryotic organisms, particularly in the context of outbreak investigation. We have outlined how uncertainties induced by methylated bases in genome positions can falsely exclude strains related to outbreaks. To remove these errors, we have proposed solutions, such as using PCR-based sequencing kits and our bioinformatics workflow. Additionally, advancements in direct methylation calling and the possible use of duplex reads might enhance the precision of Nanopore-based genome sequencing. However, these advancements will necessitate thorough validation and comprehensive testing across varied prokaryotic samples to ensure their reliability. Therefore, using Oxford Nanopore sequencing for tracing infectious disease outbreaks in the future requires a thorough comprehension of this current drawback and a continuous commitment to validate the associated methodology.

Methods

Isolates and Genomic Data

Nanopore sequencing data from three institutes have been collected and analysed. The sequencing data includes 264 isolates from 32 species, provided by the Leibniz-Institute of Photonic Technology Jena, Medical University of Graz and University Hospital Jena. Additionally, a set of 80 samples containing *K. pneumonia*, *Enterococcus faeces*, *Listeria monocytogenes*, *and Staphylococcus aureus* from a ring trail were used for analysis. University Hospital Leipzig provided 33 *K. pneumonia* outbreak isolates and Illumina sequencing data.

Genomic DNA Isolation

Isolates from 10% glycerin cryo culture streaked out on Columbia Agar with 5% Sheep Blood (Becton Dickinson). After overnight incubation, a single colony was selected and cultured overnight in liquid MH-Broth. Genomic DNA was isolated via ZymoBIOMICS DNA Microprep Kit (D4301 & D4305) from ZymoResearch with modifications to enhance the output yield. Qubit dsDNA BR Assay-Kit (Thermo Fisher Scientific) was employed to quantify DNA concentrations obtained from each isolate accurately. This kit uses fluorescent dyes to measure double-stranded DNA to ensure reliable results.

Whole Genome Sequencing

To prepare the library for sequencing using Oxford Nanopore Technologies' GridIon system, we used the Native Barcoding Kit 24 V12 (SQK-NBD112.24, Oxford Nanopore Technologies) and Native Barcoding Kit 24 V14 (SQK-NBD114.24, Oxford Nanopore Technologies) with R10.4 and R10.4.1 flowcells, respectively. Both sequencing protocols were optimised regarding prolonged incubation times. Additionally, one library was prepared with Rapid PCR

Barcoding Kit 24 (SQK-RPB114.24, Oxford Nanopore Technologies) for sequencing on R10.4.1 flowcell. Sequencing of libraries prepared with SQK-NBD112.24 and SQK-NBD114.24 was conducted with 4Hz and 260bp/s instead of 5Hz and 400bp/s for SQK-RPB114.24. The DNA fragments minimum length for all sequencing runs was set to 200bp in MinKNOW (v22.12.5) software.

Basecalling and Assembly

Basecalling and barcode demultiplexing were performed on the Gridion deploying Guppy (v6.4.6) using super accurate mode models associated with the different used sequencing kits (dna_r10.4_e8.1_sup.cfg, dna_r10.4.1_e8.2_260bps_sup.cfg, dna_r10.4.1_e8.2_5khz_400bps_sup.cfg). For further analysis, Dorado (v0.3.0) was used (dna_r10.4.1_e8.2_260bps_sup.cfg and dna_r10.4.1_e8.2_260bps_modbases_5mc_cg_sup.cfg)

De novo assembly was conducted using Flye (v2.9) [23]. The assembly was polished by minimap2 [24] (v2.18), racon[1] (v1.4.20) and medaka[2] (v1.5.0) using following models: r104_e81_sup_g5015, dna_r10.4.1_e8.2_260bps_sup@v3.5.2, r1041_e82_260bps_sup_g632.

Core genome multilocus sequence typing of *K.penumonia*

Core genome multilocus sequence typing (cgMLST) by Ridom SeqSphere+ [25] was utilised to compare Illumina and Nanopore genomes. Analysis was performed in a set of 2365 core loci that were present in all genomes, ensuring that only conserved genomic regions were included in the analysis.

[1] github.com/lbcb-sci/racon
[2] github.com/nanoporetech/medaka

Phylogenetic tree

The phylogenetic tree visualises the evolutionary relationship among *K. pneumonia* outbreak strains and is constructed based on variant calling using snippy[3]. The tree illustrates how closely or distantly related these strains are, providing insights into the patterns of divergence or clustering. The phylogenetic tree was built using FastTree[4] and toytree [26] and visualised with microreact [27].

Workflow for detection and masking of ambiguous positions

We developed a standardised nextflow workflow for de novo quality validation of all species, which is publicly available at https://github.com/replikation/MPOA, licenced under GNU General Public License v3.0. The workflow only needs the genome file (FASTA) and the associated reads (FASTQ) (Figure 4). The workflow provides reproducible quality control by counting and summarising ambiguous bases for the user, masking low coverage regions (0-10x depth) with BEDTools [28] (v2.31.0), and providing an assembly with these positions masked by the IUPAC nucleotide code for subsequent analysis. The workflow utilises docker and is compatible with Google Cloud. Identification and masking of ambiguous positions were conducted using samtools consensus [29] (v1.17) and minimap2 [24] (v.2.26). PlasFlow [30] (v1.1.0) extracts chromosome contigs for consideration without plasmid sequences. R was utilised to plot the sequence motif with ggseqlogo [31] and a violin chart comparing base frequencies with ggplot [28].

[3] https://github.com/tseemann/snippy
[4] http://www.microbesonline.org/fasttree

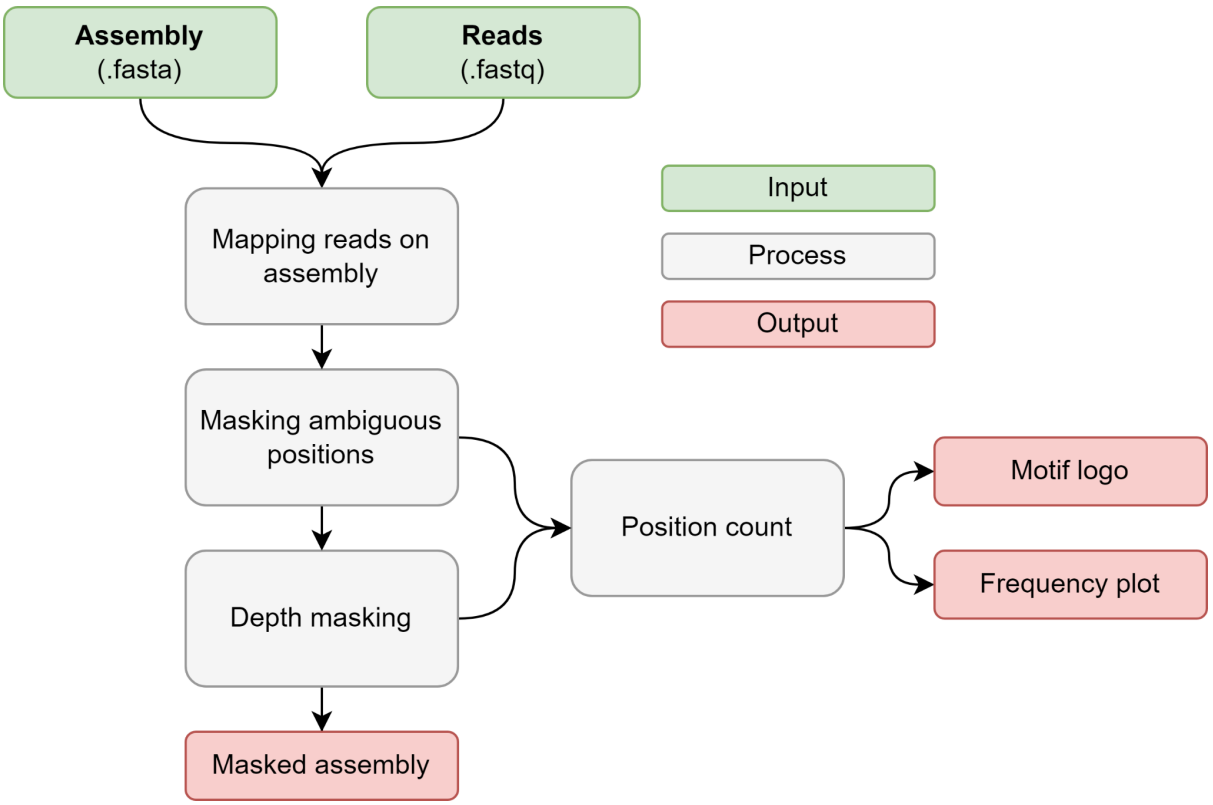


Figure 4: MPOA workflow to mask ambiguous and low coverage positions in genome files. (https://github.com/replikation/MPOA).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All parties consent to this publication.

Availability of data and materials

Workflow is available at https://github.com/replikation/MPOA.

Upload to ENA pending.

Competing interests

Not applicable

Funding

This work received financial support from the Ministry for Economics, Sciences and Digital Society of Thuringia (TMWWDG) under the framework of the Landesprogramm ProDigital (DigLeben-5575/10-9).

Authors' contributions

Sample collection and preparation, M.L., A.V.; sequencing, M.L.; workflow development and testing, M.L. and C.B.; bioinformatic analysis, M.L. and C.B.; literature research, M.L.; writing first draft, M.L.; reviewing and editing manuscript, M.L., C.B., A.V., C.S, M.M., G.W.L, J.D.H., R.E, S.D.B., S.M. and M.W.P.; supervision, C.B.; project administration C.B.; cgMLST analysis, C.S.; funding acquisition, C.B.; providing genomes: M.L., S.D.B., S.M., A.V., G.W.L, J.D.H; All authors have read and agreed to the published version of the manuscript.

References

[1] K. L. Wyres, M. M. C. Lam, and K. E. Holt, “Population genomics of Klebsiella pneumoniae,” *Nat. Rev. Microbiol.*, vol. 18, no. 6, pp. 344–359, Jun. 2020, doi:

10.1038/s41579-019-0315-1.

[2] C. Chewapreecha *et al.*, “Global and regional dissemination and evolution of Burkholderia pseudomallei,” *Nat. Microbiol.*, vol. 2, p. 16263, Jan. 2017, doi: 10.1038/nmicrobiol.2016.263.

[3] C. Brandt *et al.*, “poreCov-An Easy to Use, Fast, and Robust Workflow for SARS-CoV-2 Genome Reconstruction via Nanopore Sequencing,” *Front. Genet.*, vol. 12, p. 711437, 2021, doi: 10.3389/fgene.2021.711437.

[4] J. Moura de Sousa, M. Lourenço, and I. Gordo, “Horizontal gene transfer among host-associated microbes,” *Cell Host Microbe*, vol. 31, no. 4, pp. 513–527, Apr. 2023, doi: 10.1016/j.chom.2023.03.017.

[5] L. Hadjadj *et al.*, “Outbreak of carbapenem-resistant enterobacteria in a thoracic-oncology unit through clonal and plasmid-mediated transmission of the bla OXA-48 gene in Southern France,” *Front. Cell. Infect. Microbiol.*, vol. 12, p. 1048516, 2022, doi: 10.3389/fcimb.2022.1048516.

[6] N. A. Lerminiaux and A. D. S. Cameron, “Horizontal transfer of antibiotic resistance genes in clinical environments,” *Can. J. Microbiol.*, vol. 65, no. 1, pp. 34–44, Jan. 2019, doi: 10.1139/cjm-2018-0275.

[7] R. Abe *et al.*, “Hospital-wide outbreaks of carbapenem-resistant Enterobacteriaceae horizontally spread through a clonal plasmid harbouring blaIMP-1 in children’s hospitals in Japan,” *J. Antimicrob. Chemother.*, vol. 76, no. 12, pp. 3314–3317, Nov. 2021, doi: 10.1093/jac/dkab303.

[8] A. Sivertsen *et al.*, “A multicentre hospital outbreak in Sweden caused by introduction of a vanB2 transposon into a stably maintained pRUM-plasmid in an Enterococcus faecium ST192 clone,” *PloS One*, vol. 9, no. 8, p. e103274, 2014, doi: 10.1371/journal.pone.0103274.

[9] M. W. Pletz *et al.*, “A Nosocomial Foodborne Outbreak of a VIM Carbapenemase-Expressing Citrobacter freundii,” *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.*, vol. 67, no. 1, pp. 58–64, Jun. 2018, doi: 10.1093/cid/ciy034.

[10] Y. Wang, Y. Zhao, A. Bollas, Y. Wang, and K. F. Au, “Nanopore sequencing technology, bioinformatics and applications,” *Nat. Biotechnol.*, vol. 39, no. 11, pp. 1348–1365, Nov. 2021, doi: 10.1038/s41587-021-01108-x.

[11] J. C. Dohm, P. Peters, N. Stralis-Pavese, and H. Himmelbauer, “Benchmarking of long-read correction methods,” *NAR Genomics Bioinforma.*, vol. 2, no. 2, p. lqaa037, Jun. 2020, doi: 10.1093/nargab/lqaa037.

[12] J. Eid *et al.*, “Real-time DNA sequencing from single polymerase molecules,” *Science*, vol. 323, no. 5910, pp. 133–138, Jan. 2009, doi: 10.1126/science.1162986.

[13] J. R. Tyson, N. J. O’Neil, M. Jain, H. E. Olsen, P. Hieter, and T. P. Snutch, “MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome,” *Genome Res.*, vol. 28, no. 2, pp. 266–274, Feb. 2018, doi: 10.1101/gr.221184.117.

[14] M. A. Grohme *et al.*, “The genome of Schmidtea mediterranea and the evolution of core cellular mechanisms,” *Nature*, vol. 554, no. 7690, pp. 56–61, Feb. 2018, doi: 10.1038/nature25473.

[15] R. Spott, B. T. Schleenvoigt, B. Edel, M. W Pletz, and C. Brandt, “A Rare Case of Periprosthetic Streptobacillosis - Rapid Identification via Nanopore Sequencing after Inconclusive VITEK MS Results,” *Arch. Clin. Med. Case Rep.*, vol. 06, no. 04, 2022, doi: 10.26502/acmcr.96550529.

[16] Y. Ni, X. Liu, Z. M. Simeneh, M. Yang, and R. Li, “Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing,” *Comput. Struct. Biotechnol. J.*, vol. 21, pp. 2352–2364, 2023, doi: 10.1016/j.csbj.2023.03.038.

[17] G. E. Wagner *et al.*, “Real-Time Nanopore Q20+ Sequencing Enables Extremely Fast and Accurate Core Genome MLST Typing and Democratizes Access to High-Resolution Bacterial Pathogen Surveillance,” *J. Clin. Microbiol.*, vol. 61, no. 4, p. e0163122, Apr. 2023, doi: 10.1128/jcm.01631-22.

[18] N. D. Sanderson *et al.*, “Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction,” *Microb. Genomics*, vol. 9, no. 1, p. mgen000910, Jan. 2023, doi: 10.1099/mgen.0.000910.

[19] J. Linde *et al.*, “Comparison of Illumina and Oxford Nanopore Technology for genome analysis of Francisella tularensis, Bacillus anthracis, and Brucella suis,” *BMC Genomics*, vol. 24, no. 1, p. 258, May 2023, doi: 10.1186/s12864-023-09343-z.

[20] A. Viehweger *et al.*, “Context-aware genomic surveillance reveals hidden transmission of a carbapenemase-producing Klebsiella pneumoniae,” *Microb. Genomics*, vol. 7, no. 12, p. 000741, Dec. 2021, doi: 10.1099/mgen.0.000741.

[21] C. Brandt *et al.*, “Assessing genetic diversity and similarity of 435 KPC-carrying plasmids,” *Sci. Rep.*, vol. 9, no. 1, p. 11223, Aug. 2019, doi: 10.1038/s41598-019-47758-5.

[22] A. Tourancheau, E. A. Mead, X.-S. Zhang, and G. Fang, “Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing,” *Nat. Methods*, vol. 18, no. 5, pp. 491–498, May 2021, doi: 10.1038/s41592-021-01109-3.

[23] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner, “Assembly of long, error-prone reads using repeat graphs,” *Nat. Biotechnol.*, vol. 37, no. 5, pp. 540–546, May 2019, doi: 10.1038/s41587-019-0072-8.

[24] H. Li, “Minimap2: pairwise alignment for nucleotide sequences,” *Bioinforma. Oxf. Engl.*, vol. 34, no. 18, pp. 3094–3100, Sep. 2018, doi: 10.1093/bioinformatics/bty191.

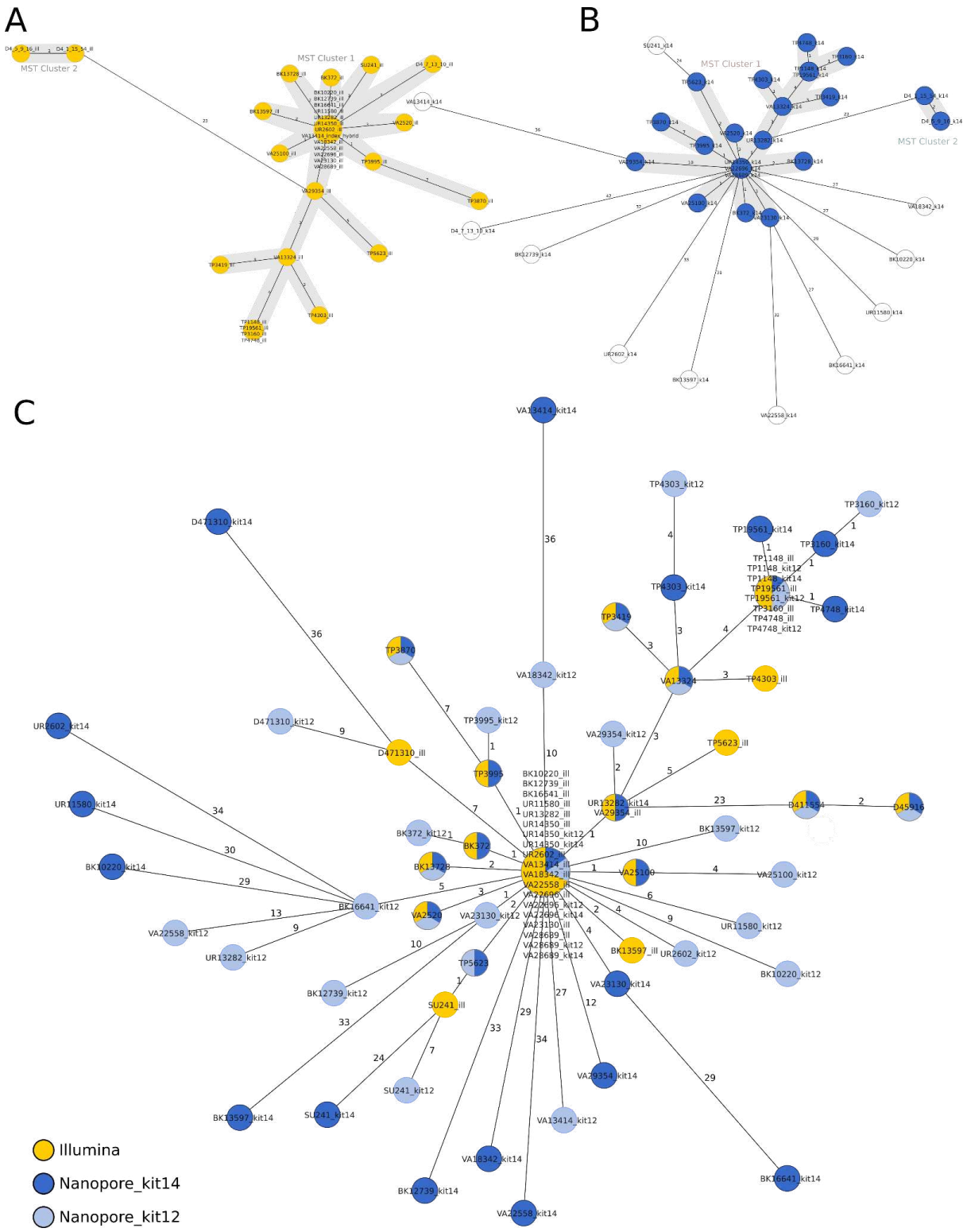
[25] S. Jünemann *et al.*, “Updating benchtop sequencing performance comparison,” *Nat. Biotechnol.*, vol. 31, no. 4, pp. 294–296, Apr. 2013, doi: 10.1038/nbt.2522.

[26] D. A. R. Eaton, “Toytree: A minimalist tree visualization and manipulation library for Python,” *Methods Ecol. Evol.*, vol. 11, no. 1, pp. 187–191, Jan. 2020, doi: 10.1111/2041-210X.13313.

[27] S. Argimón *et al.*, “Microreact: visualizing and sharing data for genomic epidemiology and phylogeography,” *Microb. Genomics*, vol. 2, no. 11, p. e000093, Nov. 2016, doi: 10.1099/mgen.0.000093.

[28] H. Wickham, *ggplot2*. in Use R! Cham: Springer International Publishing, 2016. doi: 10.1007/978-3-319-24277-4.

Supplementary Data

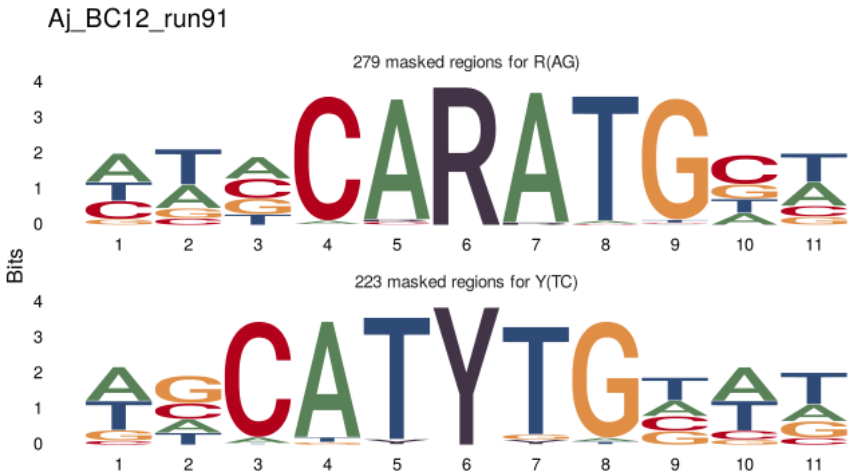
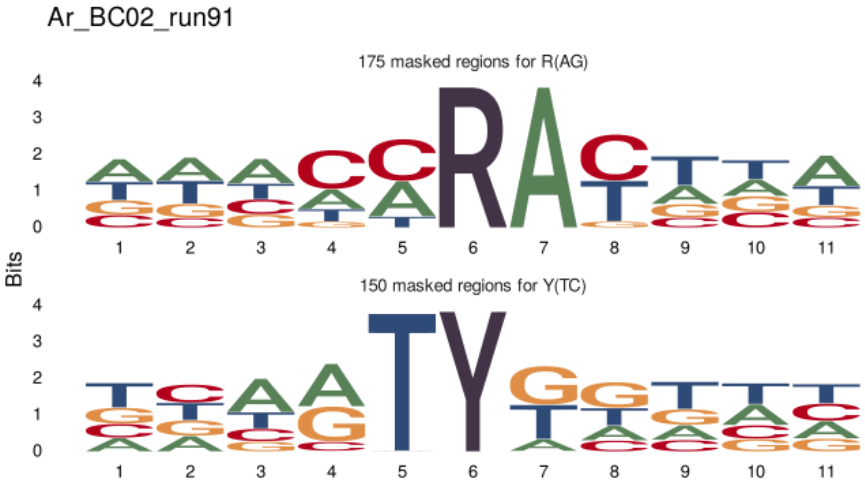


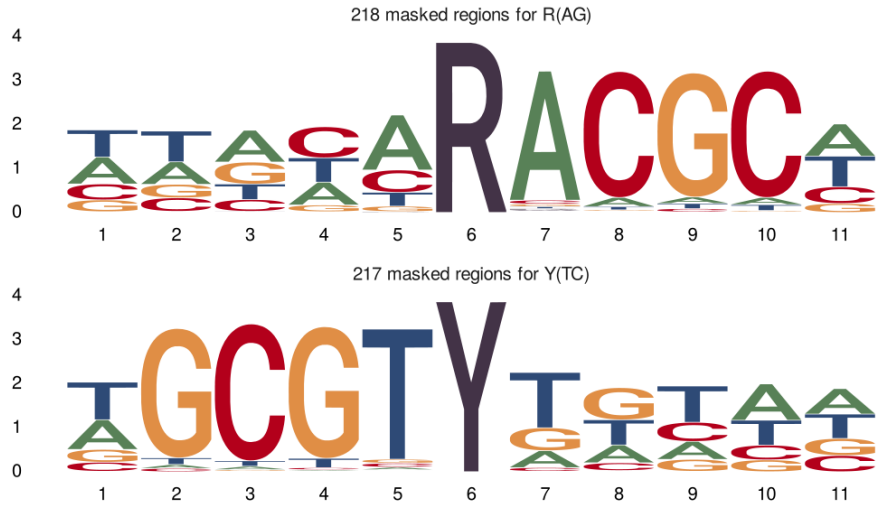
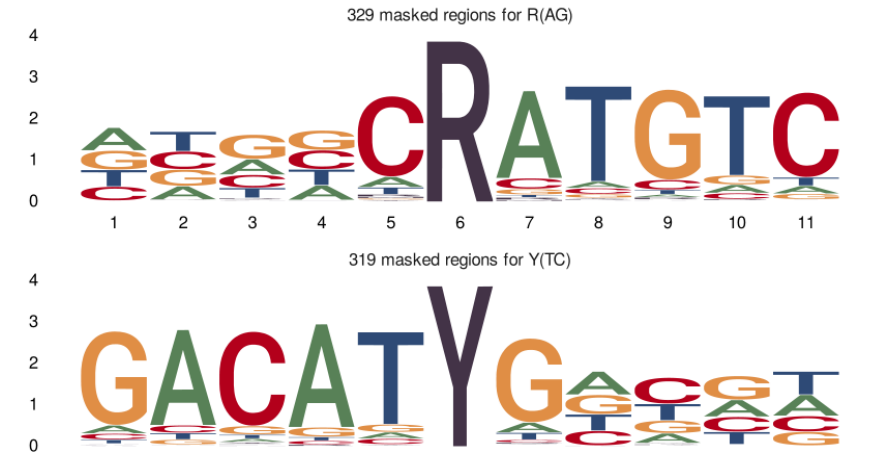
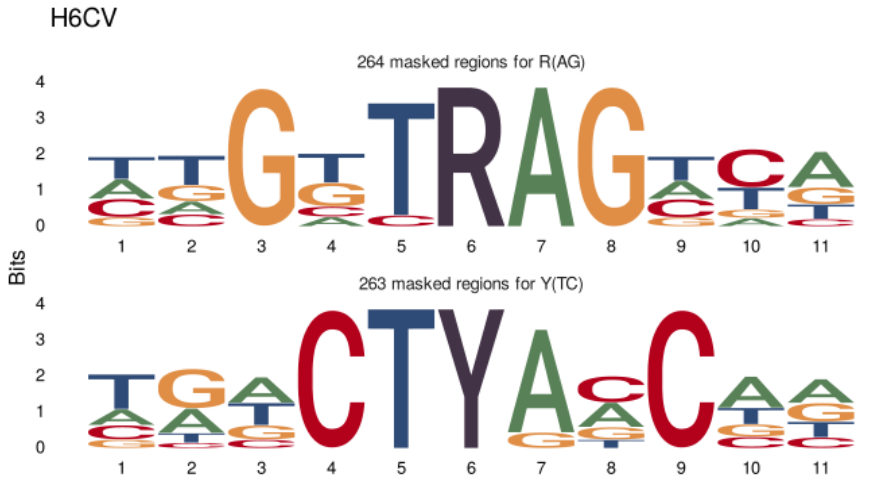
Supplementary Figure 1: A: Minimum spanning tree based 2365 loci of each 33 *K. pneumoniae* outbreak samples sequenced with Illumina **B:** Minimum spanning tree based

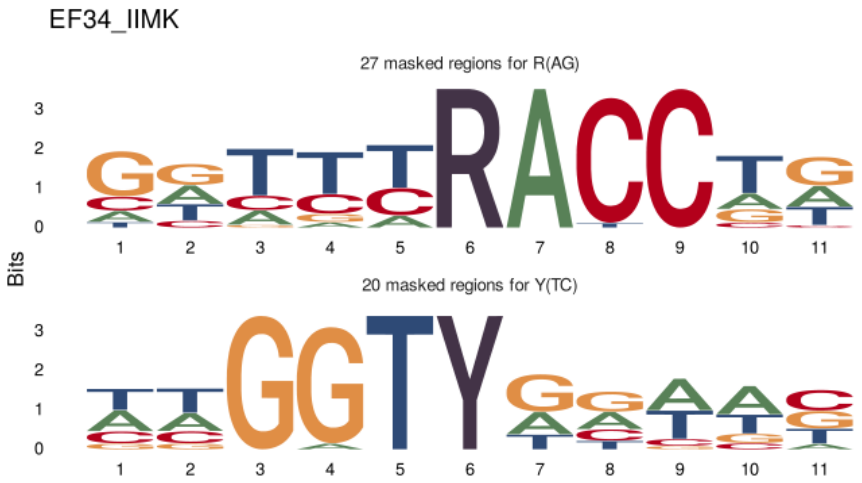
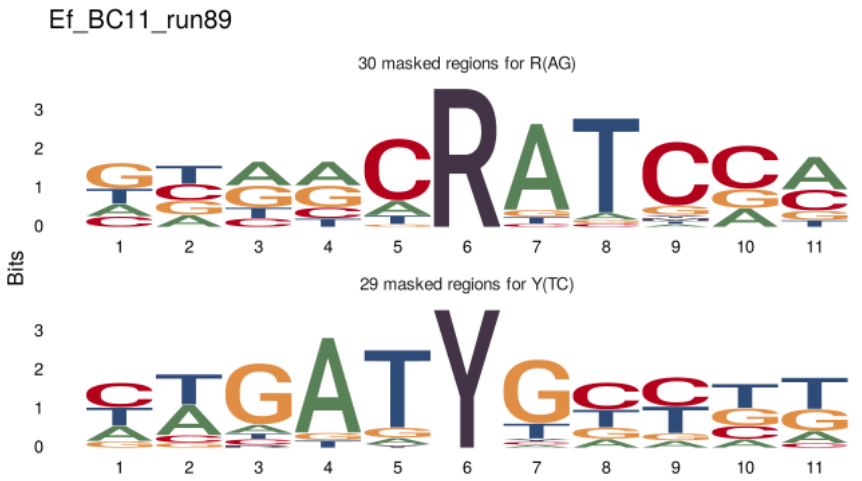
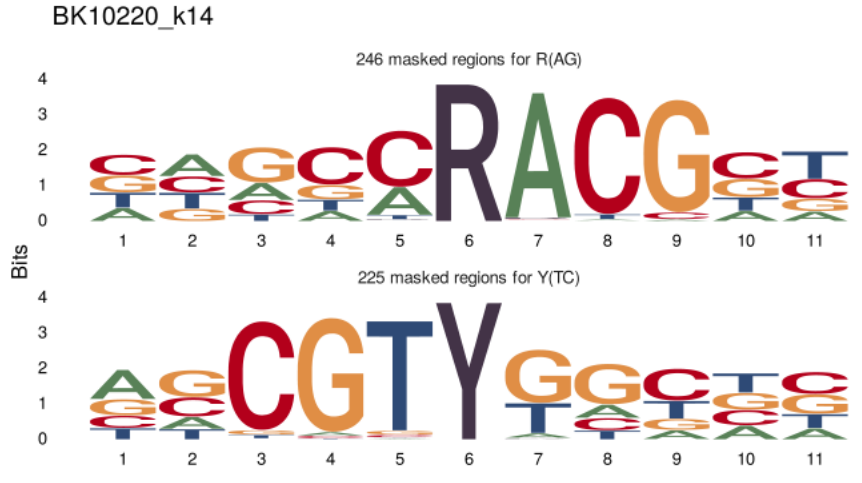
2365 loci of each 33 *K. pneumoniae* outbreak samples sequenced with Nanopore Kit 14. Outlier samples not in the outbreak cluster and showing significant differences from the Illumina data are shown in white.

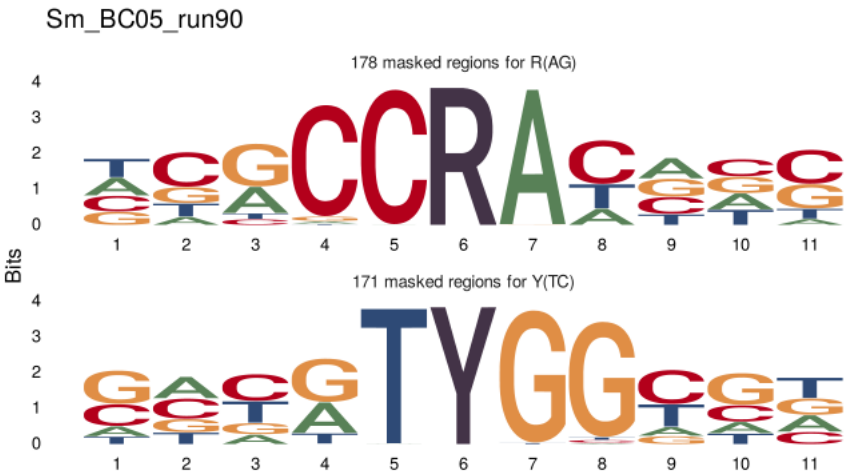
C: Minimum spanning trees of each 33 *K. pneumoniae* outbreak samples based on 2365 loci to compare the allelic variations between Illumina genomes to Nanopore SQK-NBD114.24 (kit14) and SQK-NBD114.24 (kit12). Nodes (samples) are connected by lines depicting the distance by numbers of allelic differences. Loci are considered different if one or more bases change between the samples. Loci without allelic differences are described as being the same.

Supplementary Table 1: Sequence logos of observed sequence pattern around the ambiguous bases R and Y on the chromosomal contig for different species based on one sample.

Species	Motif type
Acinetobacter junii	
Acinetobacter radioresistens	

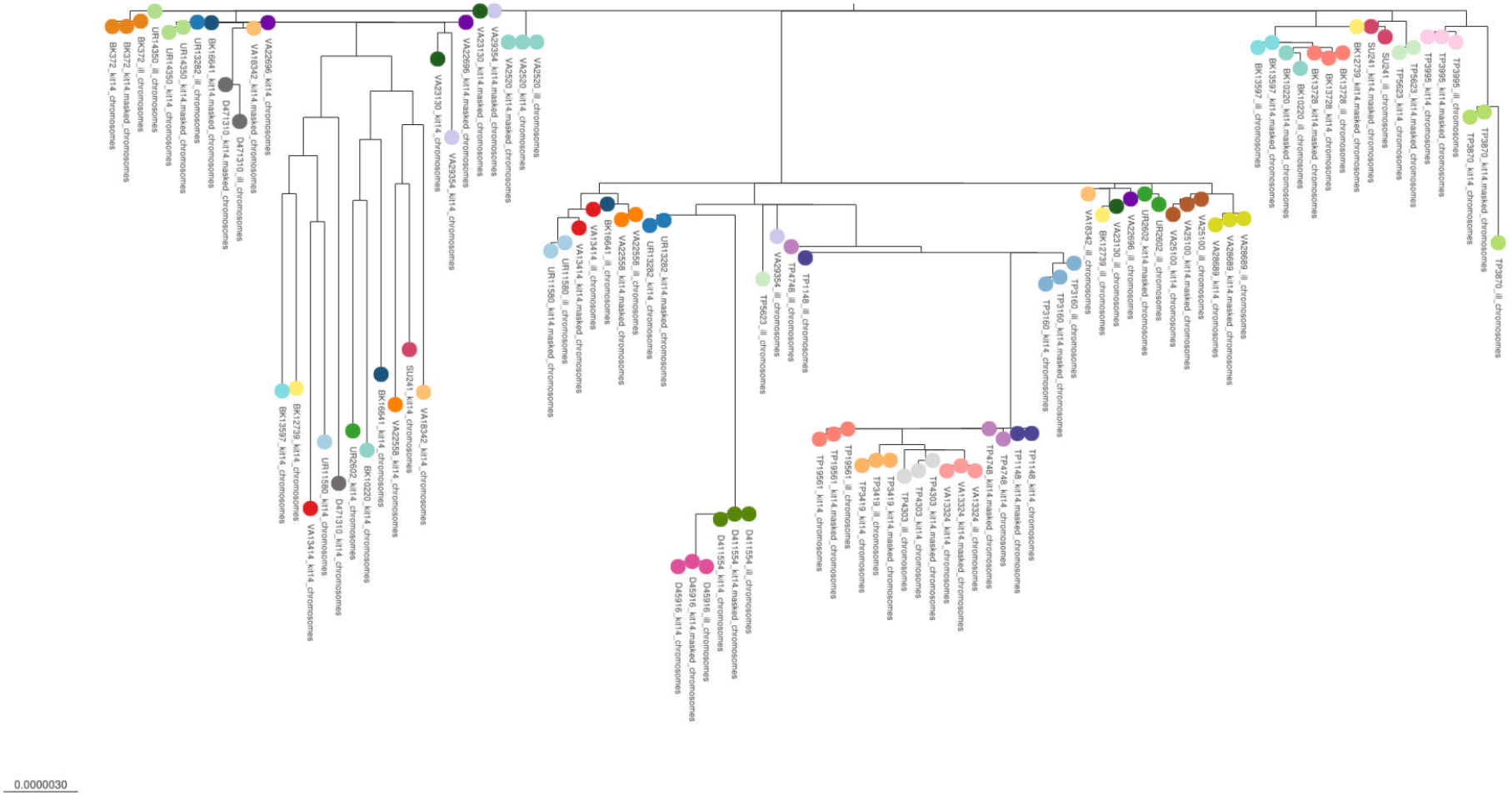
Chryseobacterium gleum	
Citrobacter freundii	
Enterococcus faecalis	

Enterococcus faecium	
Escherichia flexneri	
Klebsiella pneumonia	

Serratia marcescens	

Supplementary Table 2: Ambiguous Position within the chromosome prepared with SQK-NBD114.24 compared to SQK-RPB114.24.

	without PCR (SQK-NBD114.24)		with PCR (SQK-RPB114.24)	
	R (A or G)	Y (T or C)	R (A or G)	Y (T or C)
UR2602	260	241	0	0
VA13414	257	234	1	0
BK12739	244	243	0	0
VA18342	247	230	0	0
VA23130	111	99	3	3
BK13728	53	55	0	1
UR14350	14	5	0	0
D411554	11	12	3	3



Supplementary Figure 2: Phylogenetic tree to figure the genetic distances between 33 *K. pneumoniae* outbreak samples (coloured nodes), prepared with Illumina (ill) and Nanopore SQK-NBD114.24 (kit14) compared to the masked Kit 14 assemblies (masked).