Accurate genotyping of three major respiratory bacterial pathogens with ONT R10.4.1 long-read sequencing

Nora Zidane, Carla Rodrigues, Valerie Bouchez, et al.

Genome Res. published online June 2, 2025
Access the most recent version at doi:10.1101/gr.279829.124

Accurate genotyping of three major respiratory bacterial pathogens with ONT R10.4.1 long-read sequencing

Nora Zidane[1], Carla Rodrigues[1,3], Valérie Bouchez[1,3], Martin Rethoret-Pasty[1], Virginie Passet[1,2], Sylvain Brisse[1,2,3] and Chiara Crestani[1,*]

[1]Institut Pasteur, Université Paris Cité, Biodiversity and Epidemiology of Bacterial Pathogens, Paris, France

[2]Institut Pasteur, National Reference Center for Corynebacteria of the *diphtheriae* species complex

[3]Institut Pasteur, National Reference Center for Whooping cough and other *Bordetella* infections

*Correspondence:

Chiara Crestani: Institut Pasteur, Biodiversity and Epidemiology of Bacterial Pathogens, 25-28 rue du Docteur Roux, F-75724, Paris, France; Phone: +33 1 45 68 80 05; E-mail: chiara.crestani@pasteur.fr.

Abstract

High-throughput massive parallel sequencing has significantly improved bacterial pathogen genomics, diagnostics, and epidemiology. Despite its high accuracy, short-read sequencing struggles with complete genome reconstruction and assembly of extrachromosomal elements such as plasmids. Long-read sequencing with Oxford Nanopore Technologies (ONT) presents an alternative that offers benefits including real-time sequencing and cost-efficiency, particularly useful in resource-limited settings. However, the historically higher error rates of ONT data have so far limited its application in high-precision genomic typing. The recent release of ONT's R10.4.1 chemistry, with significantly improved raw read accuracy (Q20+), offers a potential solution to this problem.

The aim of this study was to evaluate the performance of ONT's latest chemistry for bacterial genomic typing against the gold standard Illumina technology, focusing on three respiratory pathogens of public health importance, *Klebsiella pneumoniae*, *Bordetella pertussis*, and *Corynebacterium diphtheriae*, and their related species. Using the Rapid Barcoding Kit V14, we generated and analyzed genome assemblies with different basecalling models, at different simulated depths of coverage. ONT assemblies were compared to the Illumina reference for completeness and core genome multilocus sequence typing (cgMLST) accuracy (number of allelic mismatches).

Our results show that genomes obtained from raw ONT data basecalled with Dorado SUP v0.9.0, assembled with Flye, and with a minimum coverage depth of 35×, optimized accuracy for all bacterial species tested. Error rates were consistently below 0.5% for each cgMLST scheme, indicating that ONT R10.4.1 data is suitable for high-resolution genomic typing applied to outbreak investigations and public health surveillance.

Keywords: *Klebsiella pneumoniae, Bordetella pertussis, Corynebacterium diphtheriae*, long-read sequencing, genomic typing, Oxford Nanopore sequencing

Introduction

Whole genome sequencing has revolutionized the study of bacterial pathogens, emerging as crucial tool for molecular diagnostics and epidemiology, and as cornerstone of public health and clinical microbiology (Bagger et al., 2024; Doll et al., 2024; Revez et al., 2017). Over the

past two decades, short-read sequencing technologies have dominated the research field and the market for molecular diagnostics and public health surveillance due to their high throughput and low error rates (Fox et al., 2014; Pfeiffer et al., 2018). This has provided scientists worldwide with high-resolution data for bacterial strain subtyping, which is indispensable for accurate and reliable public health surveillance. Today, bacterial isolate differentiation and outbreak investigation are mainly carried out using Single-Nucleotide-Polymorphism (SNP) analysis and gene-by-gene methods, including core genome multilocus sequence typing (cgMLST) schemes. These schemes are available on curated databases like BIGSdb-Pasteur, PubMLST (Jolley et al., 2018; Jolley & Maiden, 2010), and EnteroBase (Zhou et al., 2020), facilitating standardized genomic typing, surveillance, and outbreak investigation of key pathogens (e.g., *Listeria monocytogenes*, and *Salmonella enterica*) from short-read assemblies.

However, reconstructing a complete bacterial genome *de novo* from short-read data is rarely possible due to complex, repetitive genomic regions such as insertion sequences (IS) and other repetitive elements (Ring et al., 2018). Short reads also struggle to reconstruct extra-chromosomal elements, such as plasmids (Arredondo-Alonso et al., 2017), making it difficult to map specific genes, like antimicrobial resistance (AMR) genes, to either the chromosome or mobile genetic elements. Additionally, short-read sequencing technologies like Illumina sequencing by synthesis remain relatively expensive, both in terms of price *per* genome and acquisition cost of these sequencing platforms. These factors, along with limited portability, hinder their use in small labs and low- and middle-income countries (LMICs).

Oxford Nanopore Technologies (ONT) sequencing overcomes several of these issues, including portability (e.g., MinION device), cost-efficiency (especially when multiplexing; Sanderson et al., 2024), and the ability to circularize chromosomes and plasmids due to its long-read nature (Lerminiaux et al., 2024). It also provides options for real-time sequencing and rapid library preparation, essential for quick outbreak responses (Wagner et al., 2023). Early work on ONT data showed promising results with regards to typing and outbreak investigation (Liou et al., 2020; Quick et al., 2015); however, higher error rates of early ONT

chemistries (e.g. R9) compared to Illumina data (Dohm et al., 2020; Jain et al., 2017) have so far limited its use in surveillance, as these may strongly impact gene-by-gene approaches like cgMLST. In fact, spurious SNPs introduced by sequencing errors can create artificial alleles, increasing allelic distances between isolates. As a result, most public bacterial genotyping databases, such as BIGSdb-Pasteur, do not currently accept ONT-only data.

Recent studies have attempted to benchmark ONT sequencing for bacterial genomic typing, showing variable but promising results depending on laboratory and bioinformatics factors (Lerminiaux et al., 2024; Sanderson et al., 2023, 2024; Soto-Serrano et al., 2024; Wagner et al., 2023), as well as on the pathogen (Linde et al., 2023; Sanderson et al., 2023). The new ONT chemistry R10.4.1 may address the high error rates of previous versions, offering a declared raw read accuracy comparable to short-read technologies (Phred Q20+, i.e. ≥99 % base accuracy).

This study aimed to compare the performance of the existing gold standard short-read sequencing technology for bacterial genotyping (Illumina) with the latest ONT chemistry for cgMLST typing of three key groups of respiratory pathogens, namely *Klebsiella pneumoniae, Corynebacterium diphtheriae* (the major agent of diphtheria), *Bordetella pertussis* (the agent of whooping cough) and their related species, using the Rapid Barcoding Kit V14.

Results

Most raw reads generated with Dorado SUP v0.9.0 have Q20+ quality scores

Raw reads produced by basecalling ONT data with the high accuracy (HAC) model of Dorado v0.9.0 had lower quality scores compared to the super accurate (SUP) model, with a median read quality between 15.9 and 17.2 (Supplemental Fig. S1). Quality scores increased significantly when basecalling raw data with the SUP algorithm, with most of the basecalled reads showing Q20+ quality scores (mean read quality of 21.4 for *Klebsiella*, of 21.0 for *Corynebacterium* and of 22.6 for *Bordetella*; Supplemental Fig. S1).

ONT long-reads improve genome completeness and circularization of chromosome and extra-chromosomal elements

Illumina assemblies comprised between 14 and 295 contigs (Supplemental Fig. S2), with the *Bordetella* genus exhibiting the highest average number of contigs (n=186). *Klebsiella* genomes had an average of 53 contigs, and *Corynebacterium* genomes averaged 27.

ONT genomes consisted of one to 19 contigs (Supplemental Fig. S2). Assemblies generated from HERRO (Haplotype-aware ERRor cOrrection)-corrected reads displayed an overall lower level of circularization (both chromosome and plasmids) and higher genome fragmentation, particularly in the Corynebacteria of the *diphtheriae* Species Complex (CdSC), compared to assemblies derived from uncorrected reads (Supplemental Fig. S3, Supplemental Fig. S4, Supplemental Fig. S5). Most of the latter demonstrated circularization of the chromosome (contigs >2 Mbp) across the three genera (Supplemental Fig. S3, Supplemental Fig. S4, Supplemental Fig. S5), with CdSC SUP data showing a slightly higher number of linearized, although mostly complete, chromosome sequences. Notably, most *Bordetella* genome assemblies were complete and circularized (Supplemental Fig. S5).

Among our dataset, none of the *Bordetella* isolates carried plasmids, and no circular elements of size <500 kbp were detected in *Bordetella* ONT genome assemblies (Supplemental Fig. S5). In CdSC, two isolates (FRC1356 and FRC1385) contained one plasmid each, which was generally assembled and circularized correctly (Supplemental Fig. S4). In isolate FRC1385, two SUP assemblies (>65× and >75×) showed the presence of 1-2 larger plasmids (51 kbp and 68 kbp, respectively) compared to HAC assemblies, which contained solely a 30 kbp plasmid (Supplemental Fig. S6); these larger plasmids appear to comprise repeated sequences from the smaller plasmid, which is likely an assembly artifact. An additional circular element of size 11 kbp was detected uniquely in one genome assembly (>75×) obtained from SUP data of isolate FRC0466 (Supplemental Fig. S4, Supplemental Fig. S6). This plasmid shows high sequence similarity (>99%) with plasmids found in *C. diphtheriae* (e.g., FRC0402_p2, OV884290.1) and *Corynebacterium* spp. (e.g., MSK107 unnamed plasmid, CP176013.1). For the *Klebsiella pneumoniae* Species Complex (KpSC), a high concordance

in the detected circular elements was observed between assemblies from HAC and SUP data (Supplemental Fig. S3, Supplemental Fig. S6), with three exceptions. In SB132, in addition to a 96 kbp plasmid identified in most assemblies, a larger plasmid (105 kbp) was found in SUP assemblies, showing very little sequence similarity to the first. In SB5420, most HAC and SUP assemblies displayed two plasmids (48 kbp and 204 kbp), except for one SUP assembly (Supplemental Fig. S6) that contained a segment of the smaller plasmid in a circularized form. In SB11, while a high concordance was noted for a large plasmid (150 kbp), the assembly of plasmids sized 1-25 kbp showed more uncertainty, with some differences between HAC and SUP (Supplemental Fig. S6).

Basecalling with Dorado SUP v0.9.0 allows for accurate genomic typing of the three pathogens

We analyzed the number of cgMLST mismatches after allele calling, compared to Illumina assembly calls used as reference. In most cases, the average number of mismatches per isolate did not differ significantly between data basecalled with the HAC model compared to the SUP model, as it was already low in HAC-generated assemblies (Supplemental Fig. S7). However, in a few cases HAC assemblies showed a non-negligible number of allelic mismatches, whereas SUP basecalling allowed to recover almost all the correct alleles (Supplemental Fig. S7). In particular, *C. rouxii* FRC0190^{T} shows a very high number of mismatches in HAC data (>200 at high coverage depths; Supplemental Fig. S7) compared to SUP data (between 5 and 27; Supplemental Fig. S7 and Supplemental Fig. S8), with the latter being still significantly higher than any other *Corynebacterium* isolate sequenced in this study (Supplemental Fig. S8). This is very likely due to species-specific DNA modifications, such as methylation motifs that are uncommon or unique to *C. rouxii*, not being well-represented in the training datasets of the basecalling model, especially since *C. rouxii* is a rare species. For this reason, we excluded the *C. rouxii* isolate from further analyses.

Overall, allelic mismatches appear minimized in SUP assemblies at coverage depths >25× for KpSC and *Bordetella* spp. (Figure 1) and >35× for CdSC, with an error rate below 0.5% at coverages >35× across all tested cgMLST schemes.

HERRO correction shows limited benefits on bacterial genome assembly accuracy

HERRO raw read correction did not result in significant improvements over uncorrected reads for the genera *Klebsiella* (Figure1, Supplemental Fig. S7, and Supplemental Fig. S8) and *Bordetella* (for both the cgMLST_genus and cgMLST_pertussis schemes; Figure 1). In *Corynebacterium*, the effect of correction varied (Supplemental Fig. S8), generally leading to an increase in allelic mismatches in assemblies with coverages exceeding 35× (Figure 1). Considering the higher number of mismatches observed in HERRO-corrected assemblies (e.g., in CdSC) or the absence of a strong effect in improving assembly accuracy (e.g., in KpSC and in *Bordetella* spp.), we did not perform any further analyses on these assemblies.

Medaka assembly polishing improves HAC assemblies, with limited effects on accuracy

We evaluated the impact of genome assembly polishing with Medaka on the overall number of cgMLST mismatches on assemblies generated from uncorrected reads. Polishing consistently resulted in a decrease in mismatches for genome assemblies derived from HAC basecalled reads, showing significant improvements particularly for KpSC and CdSC (Supplemental Fig. S9). However, it is important to note that most often Medaka did not reduce mismatch rates to the same low level achieved by SUP assemblies (e.g., SB48, CIP100721^{T}, FRC0190^{T}, FRC4991; Supplemental Fig. S9). Additionally, the effect of polishing observed in SUP assemblies was more moderate than in HAC ones, predominantly resulting in a reduction in mismatches (n=10/24 isolates; Supplemental Fig. S9) or no effect (n=9/24 isolates), while a slight increase was noted in some cases (n=5/24 isolates).

Considering the higher accuracy of genome assemblies generated with the SUP model from uncorrected reads, and the negligible improvements of Medaka polishing on these assemblies, results reported from here onwards uniquely refer to unpolished SUP assemblies from uncorrected reads.

ONT R10.4.1 sequencing can be used for rapid outbreak investigation and public health surveillance of KpSC, CdSC, and *B. pertussis*

Single-linkage classification and minimum-spanning trees (MSTrees) with species-specific allelic mismatch thresholds may be used in public health for surveillance of multiple pathogens and for outbreak investigations. In addition, cgMLST-based Life Identification Numbers (LIN) codes can be used to classify KpSC genomes and to detect outbreak strains (Palma et al., 2024). Here, we aimed to investigate the accuracy of ONT assemblies from uncorrected SUP raw reads for these applications.

***Klebsiella pneumoniae* Species Complex**

cgMLST profiles of ONT-assembled genomes had a maximum of two allelic mismatches compared to the Illumina reference (Supplemental Fig. S8). On the MSTree, all KpSC profiles are part of the central genotype with the Illumina assembly (Supplemental Fig. S10). These profiles have identical alleles to the reference, and they are either complete, or they are missing one or more loci due to alleles that were not tagged because of spurious SNPs. In the latter case, they still cluster with the Illumina genotype in the MSTree because missing data is handled dynamically by GrapeTree, reducing the total number of loci considered in the pairwise distance calculation (i.e., GrapeTree computes the shortest possible connections between nodes to minimize the overall length of the tree). If these ONT assemblies had been scanned for new alleles, the currently incomplete profiles could appear as more distanced from the Illumina reference, which is why we do not recommend defining new alleles on ONT data at present.

LIN codes identical to that of the Illumina reference were detected for all cgMLST profiles of genomes generated from SUP uncorrected reads from the lowest coverage (>25×; Supplemental Table S1), with one exception: most ONT assemblies of SB30 had a LIN code that differed from the reference but that was identical among them. The difference was detected in bin number nine (second to last), which corresponds to a maximum of two allelic mismatches, and it was due to the presence in the ONT assemblies of an existing allele of a locus that was missing in the Illumina reference (allele 32; locus KP1_4024_S).

Corynebacteria of the *diphtheriae* Species Complex

As in KpSC, all cgMLST profiles belonging to *C. diphtheriae* and *C. ulcerans* were part of the central genotype with the Illumina reference on MSTrees generated with GrapeTree (Supplemental Fig. S11).

***Bordetella pertussis* and other *Bordetella* species**

For the genus *Bordetella* (using the cgMLST_genus scheme), the central GrapeTree genotype included the Illumina reference and most ONT genomes for *B. holmesii* (Figure 2) and *B. bronchiseptica* II. For the remaining isolates, including *B. pertussis, B. parapertussis and B. bronchiseptica* I-4, most ONT assemblies constitute the central genotype, and they all show one to three identical allelic mismatches compared to the gold standard Illumina (Figure 2). Most of these systematic mismatches are artifacts: they are a result of two alleles of the same locus being present in the ONT assemblies (e.g., alleles “10;21” of locus BORD004759 in FR7093) and only one allele being called in the Illumina genome (e.g., allele 21 of locus BORD004759 in FR7093). For these double loci, one of the two alleles always corresponds to the Illumina one (Supplemental Table S2).

With regards to *B. pertussis,* we also investigated the performance of ONT R10.4.1 on the cgMLST scheme dedicated to this species. Similarly to KpSC and CdSC, all profiles from ONT assemblies are grouped in the central genotype with the Illumina reference (Figure 2).

Discussion

Public bacterial genome databases for bacterial strain taxonomy, such as PubMLST and BIGSdb-Pasteur, have so far not accepted genomes generated with previously existing ONT sequencing chemistries (e.g., R9.4.1) due to higher error rates compared to Illumina. This study evaluates the use of ONT R10.4.1 chemistry with the Rapid Barcoding Kit V14 for fast, high-resolution genomic typing of three respiratory pathogens (*K. pneumoniae*, *C. diphtheriae*, and *B. pertussis* and related species) curated on the BIGSdb-Pasteur database. The Rapid Barcoding kit was chosen for its cost-effectiveness, simplicity, and minimal laboratory requirements, making it ideal for low-resource and emergency settings (e.g., mobile diagnostic and sequencing laboratories). Despite earlier versions having higher error rates compared to the Native Barcoding Kit, recent studies suggest that assemblies generated with the Rapid Barcoding kit V14 are comparable to Illumina data (Sanderson et al., 2024).

In our study, most raw reads generated with Dorado SUP v0.9.0 showed a high accuracy (Q20+ quality scores, Supplemental Fig. S1), which is consistent with previous work that also assessed performance of the Rapid Barcoding Kit V14 (Hall et al., 2024). HAC-generated reads had lower quality scores, which likely influenced the number of cgMLST mismatches observed in HAC assemblies (Supplemental Fig. S1, Supplemental Fig. S7).

With regards to genome completeness, ONT long-reads reduced the overall number of contigs across the three genera (Supplemental Fig. S2) and most often led to circularization of the chromosome and extra-chromosomal elements of genome assemblies from uncorrected reads. This is particularly remarkable for the genus *Bordetella*, in which high genome fragmentation due to the carriage of multiple IS copies is often observed when assembling short reads *de novo* (Ring et al., 2018). Regarding plasmids, it has been demonstrated how reconstructing or predicting plasmid sequences from short-read data poses challenges, especially for large plasmids with repeated sequences (Arredondo-Alonso et al., 2017), with long-read sequencing offering a solution to this bioinformatic issue. The reconstruction of extra-chromosomal elements is particularly beneficial for bacteria harboring AMR or virulence

plasmids, such as KpSC species. Underrepresentation of small plasmids can be an issue when assembling long-read data with certain long-read assemblers like Flye (Johnson et al., 2023). In our dataset, we observed a high concordance in reconstruction with minimal plasmid loss across our assembly sets, likely attributed to the choice of the Rapid Barcoding Kit over the Ligation Kit, which is known to cause underrepresentation of small plasmids in genome assemblies (Wick et al., 2021).

Our data show that ONT genome assemblies of the three pathogens tested can be used for genomic strain typing if generated with the following workflow: library preparation and sequencing with the Rapid Barcoding Kit V14 on R10.4.1 flow cells, Dorado basecalling (v0.9.0 or above, and the latest SUP model available), Flye assembly (v2.9.5 or higher), and a minimum coverage of 35×. Other basecalling models, such as HAC, have proven to be less accurate in both our study and previous ones (Lerminiaux et al., 2024). If computational resources are not available, genomes generated with HAC models could be polished using Medaka, which can increase assembly accuracy (Arredondo-Alonso et al., 2017; Foster-Nyarko et al., 2023; Sanderson et al., 2023); however, this process rarely results in improvements sufficient to match the quality of SUP assemblies (Supplemental Fig. S9). Thus, SUP models should be prioritized whenever possible. The higher accuracy observed in data obtained with the dna_r10.4.1_e8.2_400bps_sup@v5.0.0 model is likely a result of the overall higher quality of the raw reads basecalled with this algorithm (Supplemental Fig. S1). We anticipate that new and improved versions of Dorado, together with the latest chemistry transition that was silently released by Nanopore during the first quarter of 2024 (new motor protein E8.2.1), should lead to yet less allelic mismatches.

In addition, although preliminary analyses from other authors suggested promising performances of HERRO on microbial genomes (Ryan Wick's blog, 2024a; Ryan Wick's blog, 2024b), our results did not align with these findings, with HERRO correction having either a neutral or detrimental impact on the total number of allelic cgMLST mismatches, as illustrated in Figure 1. This could be because, while HERRO's model generalizes well to various organisms, it was primarily trained on human genome data. Some bacterial species might

have unique genomic features that the model hasn't been optimized for, hence resulting in higher error rates (e.g., as observed in the CdSC). In addition, the increased number of mismatches observed could also be explained by a higher genome fragmentation (Supplemental Fig. S3, Supplemental Fig. S4, Supplemental Fig. S5) and by the negative effect of HERRO correction on the average assembly coverage depth (Supplemental Fig. S12). In fact, HERRO discards all raw reads shorter than 10 kbp, which represented most of our data (Supplemental Fig. S1), resulting in an overall reduction of genome assembly coverage.

Though we currently advise to use ONT-generated genomes only for tagging existing cgMLST alleles and not for defining new ones due to possible spurious SNPs, here we show how this method still offers sufficient resolution for outbreak investigations and classifications (e.g., single-linkage clustering, MSTrees, and LIN code classification). In previous work on the KpSC (Hennart et al., 2022), we observed that profiles of isolates involved in reported outbreaks generally differed by only one or no allelic mismatch, with a maximum of five, and their LIN codes were either identical or only differed in the last three bins. Here, our data shows how cgMLST profiles defined on ONT R10.4.1 genomes with >25× coverage can now be used for outbreak investigation of KpSC, as they perform similarly to Illumina-generated profiles with regards to allelic distances on MSTrees (GrapeTree) and to LIN codes. In CdSC, a threshold of 25 allelic mismatches for single linkage groups was identified in previous work (Crestani et al., 2024; Guglielmini et al., 2021) as the maximum observed for known clusters of infection, and hence to define genetic clusters in both *C. diphtheriae* and *C. ulcerans*. Based on our analyses, ONT cgMLST profiles defined by tagging existing alleles perform equally to Illumina data when classifying isolates into genetic clusters. For the genus *Bordetella*, MSTrees generated from the cgMLST_genus scheme in most cases generated a central ONT genotype, which differed from the Illumina assembly by 1 to 3 allelic mismatches. This atypical observation stems from the fact of two distinct copies/alleles of the same locus are resolved in ONT assemblies, a direct consequence of ONT genomes being more complete than Illumina assemblies, as described above. In contrast, Illumina reads often collapse these two

gene copies into a single contig, producing a single consensus allele. Therefore, when investigating outbreaks of *Bordetella* with the cgMLST_genus scheme, it is important to keep in mind that allelic distances between isolates could be overestimated when including both ONT and Illumina data (Figure 2). When investigating epidemics of *B. pertussis,* it is recommended to use the cgMLST_pertussis scheme (current threshold of 3-4 allelic mismatches).

The ability to perform precise genotyping with a low-cost, portable sequencing technology, such as ONT, represents a significant advance. It is also highly timely, considering the current epidemiological situation with i) rising cases of whooping cough in multiple world regions in 2024 (Fu et al., 2023; Pan American Health Organization, 2024; Rodrigues et al., 2024; ECDC, 2024), ii) one of the largest diphtheria outbreaks of recent times in West Africa (Balakrishnan, 2024; Samarasekera, 2024; WHO, 2024) and diphtheria resurgence in Europe (Hoefer et al., 2023) and iii) the rising importance of multidrug-resistant infections caused by *K. pneumoniae* (Antimicrobial Resistance Collaborators, 2022; WHO, 2024). With ongoing technological advancements, and pending efficient procurement solutions in LMICs, ONT could soon play a crucial role in global pathogen surveillance and outbreak response, including in low-resource settings. In line with this potential, BIGSdb-Pasteur now accepts ONT assemblies form R10 chemistry for tagging known alleles, thereby facilitating broader utilization of ONT-derived data.

Methods

Bacterial isolates included in the study

Twenty-four isolates from 12 bacterial species were included in this study (Supplemental Table S1). Isolates belonged to the genera *Klebsiella* (in particular, to the KpSC; genome sizes 4.7-6.3 Mbp), *Corynebacterium* (in particular, to the CdSC; genome sizes 2.2-2.9 Mbp), and *Bordetella (B. pertussis, B. parapertussis, B. holmesii* and *B. bronchiseptica;* genome sizes 3.3-5.6 Mbp). Isolates are either reference or type strains, or they were selected among the

bacterial collections of our laboratory: i) the KpSC collection, ii) the French National Reference Centre (NRC) for CdSC collection, and iii) the French NRC for Whooping cough and other *Bordetella* infections collection.

Isolate growth and DNA extraction

KpSC and CdSC were plated on Tryptic Soy Agar (TSA) and grown at 37 °C for 24h, and at 35–37 °C for 24-48 hours, respectively. *Bordetella* spp. isolates were grown at 36°C for 24 to 72 hours on Bordet-Gengou agar (Becton Dickinson, Le Pont de Claix, France) supplemented with 15% defibrinated horse blood (BioMérieux, Marcy l'Étoile, France) and subcultured in the same medium for 24 hours in standardized conditions, as previously described (Bouchez et al., 2018).

DNA extraction was performed on a Maxwell RSC Instrument (Promega, Madison, USA) with the Maxwell RSC Blood DNA Kit (Promega, Madison, Wisconsin, USA) following the manufacturer's instructions.

Library preparation and whole genome sequencing

Libraries for short-read sequencing were prepared and sequenced at the Mutualized Platform for Microbiology (P2M, Institut Pasteur) using the Nextera XT DNA library preparation kit (Illumina, San Diego, USA) on a NextSeq 500 or NextSeq 2000 apparatuses (Illumina, San Diego, USA) with a 2×150 nt paired-end protocol.

Libraries for long-read sequencing were prepared with the Rapid Barcoding Kit V14 (SQK-RBK114.24; Oxford Nanopore Technologies, Oxford, UK), and sequenced on three R10.4.1 flow cells (FLO-MIN114, one per pathogen) on a GridION machine for 72 hours. The minimal fragment length was set at 200 bp on the GridION software, v23.11.7. All libraries included a negative control barcode prepared with nuclease-free water.

Long read basecalling and data processing

We tested different combinations of basecalling models and coverage to establish the best workflow possible (Figure 3). All combinations used to generate the final assemblies can be found in Table 1, and all bioinformatic commands can be found in Supplemental Methods.

The latest version of Dorado (https://github.com/nanoporetech/dorado) available to date, v0.9.0, was tested using two basecalling models: HAC and SUP (Table 1). Only raw reads with a quality score of 10 or more were kept after basecalling. Dorado was also used for demultiplexing and trimming of barcodes and adapters. Data was then converted from BAM to FASTQ with SAMtools v1.21 (Danecek et al., 2021). Raw read quality was assessed with NanoStat v1.6.0 (De Coster et al., 2018), and data was plotted with Python seaborn v0.13.2 (Waskom, 2021).

Long-reads were subsampled with Rasusa v0.8.0 (Hall, 2022) to simulate different depths of coverage (from 30× to 90×, based on the maximum coverage possible per isolate). To simulate an actual sequencing run, a cumulative subsampling strategy was used. Starting with the original file containing all raw reads from an isolate, a random subset of reads was drawn to simulate 30× coverage. These selected reads were removed from the original file using the fqextract function of fqtools v1.2 (Droop, 2016). From the remaining reads, another random subset was drawn to simulate 10× coverage. This 10× subset was then combined with the initial 30× set to create a file simulating 40× coverage. This process was repeated iteratively: after subtracting the previously used reads, new subsets were drawn and combined until the maximum possible coverage for each isolate was reached. This approach ensured that there was no duplication of reads: if independently drawn subsets for 10× coverage were repeatedly added to the 30× set, as in typical random sampling using Rasusa, it could result in some reads appearing multiple times in files simulating higher coverages (e.g. 40× or more).

We also wanted to test the effect of HERRO v1 (Stanojevic et al., 2024), a deep-learning tool designed for error correction of Nanopore R10.4.1 (https://github.com/lbcb-sci/herro), on the accuracy of the final assemblies, as this tool showed promising results in bacterial genomes (Ryan Wick's blog, 2024a; Ryan Wick's blog, 2024b). To this end, each subsampled read set was corrected with HERRO through its integrated version within Dorado.

Table 1. Different combinations of data processing used in this work, which generated long-read genome assemblies with different depths of coverage.

BASECALLER	MODEL	SUBSAMPLING	HERRO	ASSEMBLY	MEDAKA
DORADO V0.9.0	dna_r10.4.1_e8.2_400bps_hac@v5.0.0	30× 40× 50× 60×	N	Flye v2.9.5	r1041_e82_400bps_hac_v5.0.0
			Y		NA
	dna_r10.4.1_e8.2_400bps_sup@v5.0.0	70× 80× 90×	N		r1041_e82_400bps_sup_v5.0.0
			Y		NA

De novo assembly

Short-read sequence data was assembled with fq2dna v21.06 (https://gitlab.pasteur.fr/GIPhy/fq2dna). Subsampled long-read data sets were assembled with Flye v2.9.5 (Kolmogorov et al., 2019). We did not compare results from multiple assembly algorithms, as Flye has already been shown to lead to more complete and accurate genome assemblies compared to other tools (e.g., Unicycler, Raven; Lerminiaux et al., 2024). Low variability in average genome coverage was observed in the final assemblies generated from uncorrected reads (Supplemental Fig. S12). The coverage values generally fell within ±5 of the target coverage (e.g. when subsampling aimed for 30× coverage, most assemblies had a coverage between 25× and 35×). For this reason, the coverage is reported in a “greater than” format (e.g., for data subsampled at 30×, results are given as >25×). In contrast, assemblies generated from HERRO-corrected reads showed higher variability and much lower average genome coverage (Supplemental Fig. S12). This increased variability is a consequence of performing read correction on the subsampled data with HERRO (see Results section), which reflects a real-world sequencing scenario.

To test the performance of genome polishing with Medaka v2.0.1 (https://github.com/nanoporetech/medaka), assemblies generated from uncorrected reads and basecalled with either the HAC or SUP models underwent one polishing round (Table 1).

Klebsiella spp., _Corynebacterium_ spp. and _Bordetella_ spp. genomic typing

Genome assemblies generated with both short and long-read sequencing were uploaded to BIGSdb-Pasteur (https://bigsdb.pasteur.fr/) in their respective species databases. We defined cgMLST alleles on reference Illumina assemblies with the BIGSdb software (Jolley & Maiden, 2010), and subsequently tagged long-read assemblies for these alleles. The cgMLST schemes used for genotyping were: i) for KpSC isolates, the scgMLST629_S scheme (including 629 loci; Hennart et al., 2022); ii) for *C. diphtheriae* and *C. rouxii,* the *C. diphtheriae* cgMLST scheme (1,305 loci; Guglielmini et al., 2021), and for *C. ulcerans* the cgMLST_ulcerans scheme (1,628 loci; Crestani et al., 2024); iii) for *Bordetella* species (*B. pertussis*, *B. parapertussis*, *B. bronchiseptica*, and *B. holmesii),* the cgMLST_genus scheme (1,415 loci; Bridel et al., 2022). Additionally, we used the cgMLST_pertussis scheme (2,038 loci; Bouchez et al., 2018) on *B. pertussis* genomes.

MSTrees based on cgMLST profiles of genome assemblies generated from SUP data were constructed with GrapeTree (Zhou et al., 2018) for each genus.

In addition, LIN codes using a gene-by-gene approach (Hennart et al., 2022; Palma et al., 2024) were assigned to KpSC assemblies.

Allelic mismatch analysis and data visualization

The number of allelic mismatches between cgMLST profiles of Illumina vs ONT assemblies were computed with Python pandas v1.4.3 (The pandas development team, 2020). Missing loci from short-read reference genomes were not considered for the analysis. The type of mismatches obtained by this comparison method include: i) spurious SNPs matching by chance alleles already existing in the database; ii) spurious SNPs generating new artificial alleles (as no new alleles were defined on long-read assemblies, these would appear as

missing data in the ONT profile). The script used to compare cgMLST profiles is available in Supplemental Methods and at GitHub (https://github.com/chcrestani/Comparison_cgMLST_profiles/).

All graphs were generated with Python seaborn v0.13.2 (Waskom, 2021).

Data access

The short-read data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA1166325. The raw Nanopore data in POD5 format have been submitted to European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/home) under accession number PRJEB89064. Genome assemblies (Illumina and ONT) can be downloaded from BIGSdb-Pasteur under the "projects" section of the Isolates & genomes database (https://bigsdb.pasteur.fr/klebsiella/, Project ID 175; https://bigsdb.pasteur.fr/diphtheria/, Project ID 59; https://bigsdb.pasteur.fr/bordetella/, Project ID 82). The script for cgMLST allelic mismatch calculations is available as Supplemental Code and at GitHub (https://github.com/chcrestani/Comparison_cgMLST_profiles).

Competing interest statement

The authors declare no conflict of interest.

Acknowledgements

The National Reference Center for Corynebacteria of the *diphtheriae* complex and the National Reference Center for Whooping cough and other *Bordetella* infections are supported financially by Institut Pasteur and Santé Publique France (Public Health France). This work was supported financially by the French Government's Investissement d'Avenir grant Laboratoire d'Excellence Integrative Biology of Emerging Infectious Diseases (ANR-10-LABX-

62-IBEID) and by the Bill & Melinda Gates Foundation funded Project “An integrated platform for *Klebsiella pneumoniae* genomic surveillance” (Funder Project reference INV-025280).

Authors’ contributions are as follows. Conceptualization, Methodology and Visualization: CC; Data Curation, Validation, Formal Analysis: CC, CR, VB, MRP; Experimental work: NZ, VP; Writing original draft: CC; Writing review & editing: CC, CR, VB, SB; Resources and Funding Acquisition: SB.

We thank the Biomics Platform at Institut Pasteur for sharing their GridION machine, and in particular Chloé Baum for her support. We also thank the Mutualized Platform for Microbiology (P2M) for sequencing isolates using Illumina technology. This work used the computational and storage services provided by the IT Department at Institut Pasteur. We would also like to thank Alexis Criscuolo for his support in bioinformatics developments and analyses.

References

Antimicrobial Resistance Collaborators. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet*, **399**, 629–655. doi: 10.1016/S0140-6736(21)02724-0

Arredondo-Alonso, S., Willems, R. J., van Schaik, W., Schürch, A. C. (2017). On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom*, **3**: e000128. doi: 10.1099/mgen.0.000128

Bagger, F. O., Borgwardt, L., Jespersen, A. S., Hansen, A. R., Bertelsen, B., Kodama, M., Nielsen, F. C. (2024). Whole genome sequencing in clinical practice. In *BMC Med Genomics,* **17**: 39. doi: 10.1186/s12920-024-01795-w

Balakrishnan, V. S. (2024). Diphtheria outbreak in Nigeria. *Lancet Microbe*, **5**: e11. doi: 10.1016/S2666-5247(23)00330-0

Bouchez, V., Guglielmini, J., Dazas, M., Landier, A., Toubiana, J., Guillot, S., Criscuolo, A., Brisse, S. (2018). Genomic sequencing of *Bordetella pertussis* for epidemiology and

global surveillance of whooping cough. *Emerg Infect Dis*, **24**: 988–994. doi: 10.3201/eid2406.171464

Bridel, S., Bouchez, V., Brancotte, B., Hauck, S., Armatys, N., Landier, A., Mühle, E., Guillot, S., Toubiana, J., Maiden, M. C. J. et al. (2022). A comprehensive resource for *Bordetella* genomic epidemiology and biodiversity studies. *Nat Commun*, **13**: 3807. doi: 10.1038/s41467-022-31517-8

Crestani, C., Passet, V., Rethoret-Pasty, M., Criscuolo, A., Zidane, N., Brémont, S., Badell, E., Brisse, S. (2024). Genomic epidemiology and microevolution of the zoonotic pathogen *Corynebacterium ulcerans*. *bioRxiv,* 2024.08.22.609154. doi: 10.1101/2024.08.22.609154

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, **10**: giab008. doi: 10.1093/gigascience/giab008

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, **34**: 2666–2669. doi: 10.1093/bioinformatics/bty149

Dohm, J. C., Peters, P., Stralis-Pavese, N., Himmelbauer, H. (2020). Benchmarking of long-read correction methods. *NAR Genom Bioinform*, **2**: lqaa037. doi: 10.1093/nargab/lqaa037

Doll, M., Bryson, A. L., Palmore, T. N. (2024). Whole genome sequencing applications in hospital epidemiology and infection prevention. *Curr Infect Dis Rep*, **26**: 115-121. doi: 10.1007/s11908-024-00836-w

Droop, A. P (2016). fqtools: an efficient software suite for modern FASTQ file manipulation. *Bioinformatics*, **32**:1883-1884. doi: 10.1093/bioinformatics/btw088.

Foster-Nyarko, E., Cottingham, H., Wick, R. R., Judd, L. M., Lam, M. M. C., Wyres, K. L., Stanton, T. D., Tsang, K. K., David, S., Aanensen, D. M. et al. (2023). Nanopore-only assemblies for genomic surveillance of the global priority drug-resistant pathogen, *Klebsiella pneumoniae*. *Microb Genom,* **9**: mgen000936. doi: 10.1099/mgen.0.000936

530 Fox, E. J., Reid-Bayliss, K. S., Emond, M. J., Loeb, L. A. (2014). Accuracy of next generation
531 sequencing platforms. *Next Gener Seq Appl,* **1**:1000106. doi: 10.4172/jngsa.1000106
532 Fu, P., Zhou, J., Meng, J., Liu, Z., Nijiati, Y., He, L., Li, C., Chen, S., Wang, A., Yan, G., Lu,
533 G., Zhou, L., Zhai, X., Wang, C. (2023). Emergence and spread of MT28 ptxP3 allele
534 macrolide-resistant *Bordetella pertussis* from 2021 to 2022 in China. *Int J Infect Dis*,
535 **128**: 205–211. doi: 10.1016/j.ijid.2023.01.005
536 Guglielmini, J., Hennart, M., Badell, E., Toubiana, J., Criscuolo, A., Brisse, S. (2021).
537 Genomic epidemiology and strain taxonomy of *Corynebacterium diphtheriae*. *J Clin*
538 *Microbiol*, **59**: e0158121. doi: 10.1128/JCM.01581-21
539 Hall, M. B. (2022). Rasusa: Randomly subsample sequencing reads to a specified coverage.
540 *J Open Source Soft*, **7**: 3941. doi: 10.21105/joss.03941
541 Hall, M. B., Wick, R. R., Judd, L. M., T Nguyen, A. N., Steinig, E. J., Xie, O., Davies, M. R.,
542 Seemann, T., Stinear, P., M Coin, L. J. (2024). Benchmarking reveals superiority of
543 deep learning variant callers on bacterial nanopore sequence data. *eLife,* 13:RP98300.
544 Doi: 10.7554/eLife.98300.2
545 Hennart, M., Guglielmini, J., Bridel, S., Maiden, M. C. J., Jolley, K. A., Criscuolo, A., Brisse,
546 S. (2022). A dual barcoding approach to bacterial strain nomenclature: genomic
547 taxonomy of *Klebsiella pneumoniae* strains. *Mol Biol Evol*, **39**: msac135. doi:
548 10.1093/molbev/msac135
549 Hoefer, A., Seth-Smith, H., Palma, F., Schindler, S., Freschi, L., Dangel, A., Berger, A.,
550 D'Aeth, J., Indra, A., Fry, N. et al. (2023). Phenotypic and genomic analysis of a large-
551 scale *Corynebacterium diphtheriae* outbreak among migrant populations in Europe.
552 *medRxiv,* 2023.11.10.23297228. doi: 10.1101/2023.11.10.23297228
553 Jain, M., Tyson, J. R., Loose, M., Ip, C. L. C., Eccles, D. A., O'Grady, J., Malla, S., Leggett,
554 R. M., Wallerman, O., Jansen et al. (2017). MinION analysis and reference consortium:
555 phase 2 data release and analysis of R9.0 chemistry. *F1000Res,* **6**: 760. doi:
556 10.12688/f1000research.11354.1

557 Johnson, J., Soehnlen, M., Blankenship, H. M. (2023). Long read genome assemblers
558 struggle with small plasmids. *Microb Genom*, **9**: mgen001024. doi:
559 10.1099/mgen.0.001024
560 Jolley, K. A., Bray, J. E., Maiden, M. C. J. (2018). Open-access bacterial population
561 genomics: BIGSdb software, the PubMLST.org website and their applications.
562 *Wellcome Open Res*, **3***: 124*. doi: 10.12688/wellcomeopenres.14826.1
563 Jolley, K. A., Maiden, M. C. J. (2010). BIGSdb: Scalable analysis of bacterial genome
564 variation at the population level. *BMC Bioinformatics*, **11***: 595*. doi: 10.1186/1471-2105-
565 11-595
566 Kolmogorov, M., Yuan, J., Lin, Y., Pevzner, P. A. (2019). Assembly of long, error-prone
567 reads using repeat graphs. *Nat Biotechnol*, **37**: 540–546. doi: 10.1038/s41587-019-
568 0072-8
569 Lerminiaux, N., Fakharuddin, K., Mulvey, M. R., Mataseje, L. (2024). Do we still need
570 Illumina sequencing data? Evaluating Oxford Nanopore Technologies R10.4.1 flow
571 cells and the Rapid v14 library prep kit for Gram negative bacteria whole genome
572 assemblies. *Can J Microbiol*, **70**:178-189. doi: 10.1139/cjm-2023-0175.
573 Linde, J., Brangsch, H., Hölzer, M., Thomas, C., Elschner, M. C., Melzer, F., Tomaso, H.
574 (2023). Comparison of Illumina and Oxford Nanopore Technology for genome analysis
575 of *Francisella tularensis, Bacillus anthracis,* and *Brucella suis*. *BMC Genomics*, **24**: 258.
576 doi: 10.1186/s12864-023-09343-z
577 Liou, C.H., Wu, H.C., Liao, Y.C., Yang Lauderdale, T.L., Huang, I.W., Chen, F.J. (2020).
578 nanoMLST: accurate multilocus sequence typing using Oxford Nanopore Technologies
579 MinION with a dual-barcode approach to multiplex large numbers of samples. *Microb*
580 *Genom*, **6**:e000336. doi: 10.1099/mgen.0.000336
581 Palma, F., Hennart, M., Jolley, K. A., Crestani, C., Wyres, K. L., Bridel, S., Yeats, C. A.,
582 Brancotte, B., Raffestin, B., David, S. et al. (2024). Bacterial strain nomenclature in the
583 genomic era: Life Identification Numbers using a gene-by-gene approach. *bioRxiv,*
584 2024.03.11.584534. doi: 10.1101/2024.03.11.584534

585 Pan American Health Organization. (2024). Epidemiological alert Pertussis (whooping
586 cough) in the Region of the Americas, 22 July 2024.
587 Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., Mayer, G. (2018).
588 Systematic evaluation of error rates and causes in short samples in next-generation
589 sequencing. *Sci Rep*, **8**: 10950. doi: 10.1038/s41598-018-29325-6
590 Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., Nair, S., Nea,l K., Nye,
591 K., Peters, T., et al. (2015). Rapid draft sequencing and real-time nanopore sequencing
592 in a hospital outbreak of *Salmonella. Genome Bio,* **16**:114. doi: 10.1186/s13059-015-
593 0677-2
594 Revez, J., Espinosa, L., Albiger, B., Leitmeyer, K. C., Struelens, M. J. (2017). Survey on the
595 use of whole-genome sequencing for infectious diseases surveillance: rapid expansion
596 of European national capacities, 2015–2016. *Front Public Health*, **5**: 347. doi:
597 10.3389/fpubh.2017.00347
598 Ring, N., Abrahams, J. S., Jain, M., Olsen, H., Preston, A., Bagby, S. (2018). Resolving the
599 complex *Bordetella pertussis* genome using barcoded nanopore sequencing. *Microb*
600 *Genom*, **4**: e000234. doi: 10.1099/mgen.0.000234
601 Rodrigues, C., Bouchez, V., Soares, A., Trombert-Paolantoni, S., Aït El Belghiti, F., Cohen,
602 J. F., Armatys, N., Landier, A., Blanchot, T., Hervo, M. et al. (2024). Resurgence of
603 *Bordetella pertussis*, including one macrolide-resistant isolate, France, 2024. *Euro*
604 *Surveill,* **29**: 2400459. doi: 10.2807/1560-7917.ES.2024.29.31.2400459
605 Samarasekera, U. (2024). Diphtheria outbreak in west Africa. *Lancet Infect Diss*, **24**: e87.
606 Sanderson, N. D., Hopkins, K. M. V., Colpus, M., Parker, M., Lipworth, S., Crook, D.,
607 Stoesser, N. (2024). Evaluation of the accuracy of bacterial genome reconstruction with
608 Oxford Nanopore R10.4.1 long-read-only sequencing. *Microb Genom*, **10**: 001246. doi:
609 10.1099/mgen.0.001246
610 Sanderson, N. D., Kapel, N., Rodger, G., Webster, H., Lipworth, S., Street, T. L., Peto, T.,
611 Crook, D., Stoesser, N. (2023). Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford

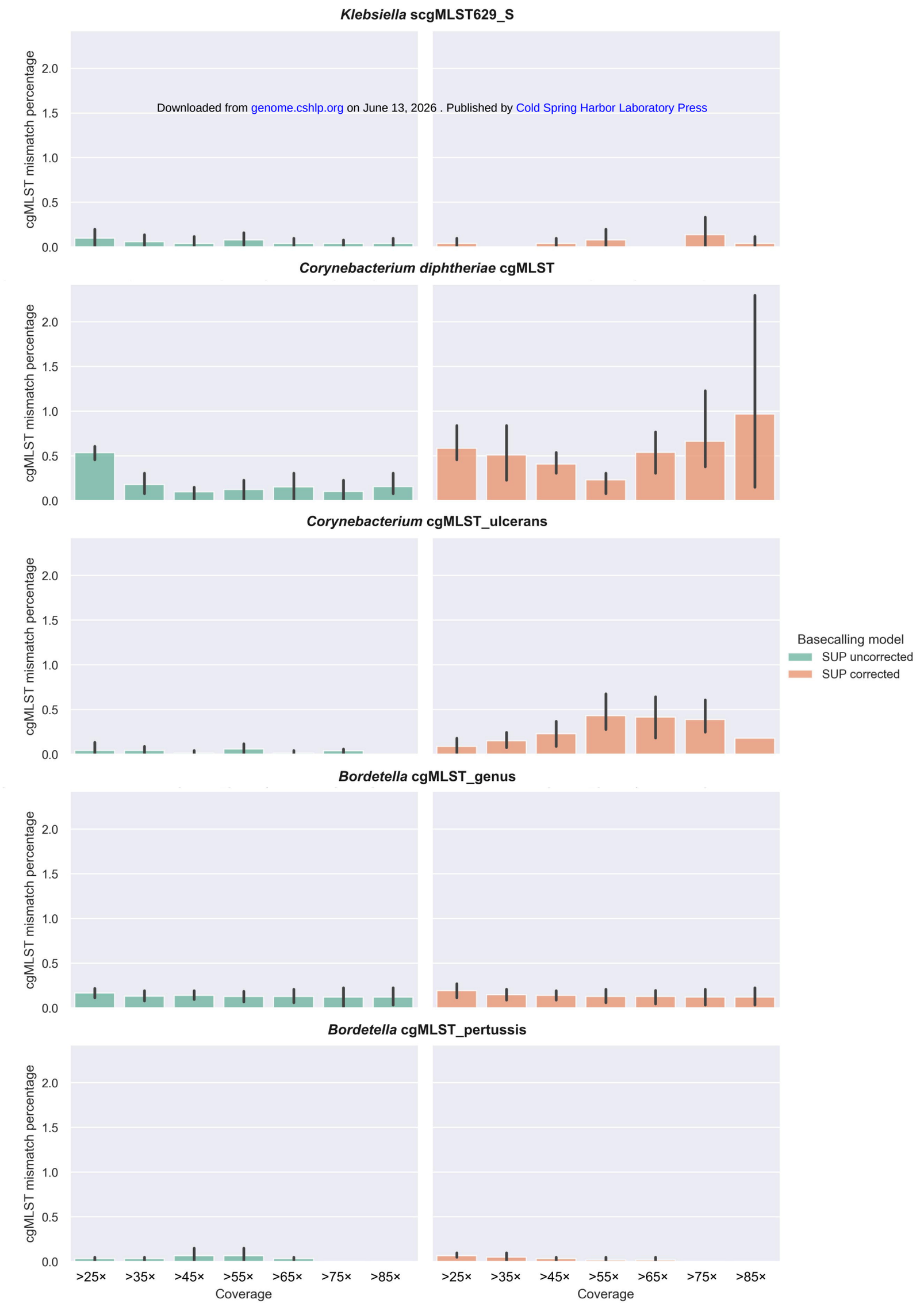
612 Nanopore flowcells and chemistries in bacterial genome reconstruction. *Microb Genom*,
613 **9:** mgen000910. doi: 10.1099/mgen.0.000910
614 Soto-Serrano, A., Li, W., Panah, F. M., Hui, Y., Atienza, P., Fomenkov, A., Roberts, R. J.,
615 Deptula, P., Krych, L. (2024). Matching excellence: ONT's rise to parity with PacBio in
616 genome reconstruction of non-model bacterium with high GC content. *bioRxiv,*
617 2024.02.26.582104. doi: 10.1101/2024.02.26.582104
618 Stanojević, D., Lin, D., Nurk, S., Florez de Sessions, P., Šikić, M. (2024). Telomere-to-
619 telomere phased genome assembly using HERRO-corrected simplex nanopore reads.
620 *bioRxiv,* 2024.05.18.594796. doi: 10.1101/2024.05.18.594796
621 The pandas development team. (2020). pandas-dev/pandas: Pandas.
622 Wagner, G. E., Dabernig-Heinz, J., Lipp, M., Cabal, A., Simantzik, J., Kohl, M., Scheiber, M.,
623 Lichtenegger, S., Ehricht, R., Leitner, E. et al. (2023). Real-Time nanopore Q20+
624 sequencing enables extremely fast and accurate core genome MLST typing and
625 democratizes access to high-resolution bacterial pathogen surveillance. *J Clin*
626 *Microbiol,* **61**: e0163122. doi: 10.1128/jcm.01631-22
627 Waskom, M. L. (2021). seaborn: statistical data visualization. *J Open Source Soft*, **6**: 3021.
628 doi: 10.21105/joss.03021
629 Wick, R. R., Judd, L. M., Wyres, K. L., Holt, K. E. (2021). Recovery of small plasmid
630 sequences via Oxford Nanopore sequencing. *Microb Genom*, **7**: 000631. doi:
631 10.1099/mgen.0.000631
632 Zhou, Z., Alikhan, N. F., Mohamed, K., Yulei, F., the Agama Study Group, Achtman, M.
633 (2020). The EnteroBase user's guide, with case studies on *Salmonella* transmissions,
634 *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res*, **30**:
635 138–152. doi: 10.1101/gr.251678.119
636 Zhou, Z., Alikhan, N. F., Sergeant, M. J., Luhmann, N., Vaz, C., Francisco, A. P., Carriço, J.
637 A., Achtman, M. (2018). Grapetree: visualization of core genomic relationships among
638 100,000 bacterial pathogens. *Genome Res*, **28**: 1395–1404. doi:
639 10.1101/gr.232397.117

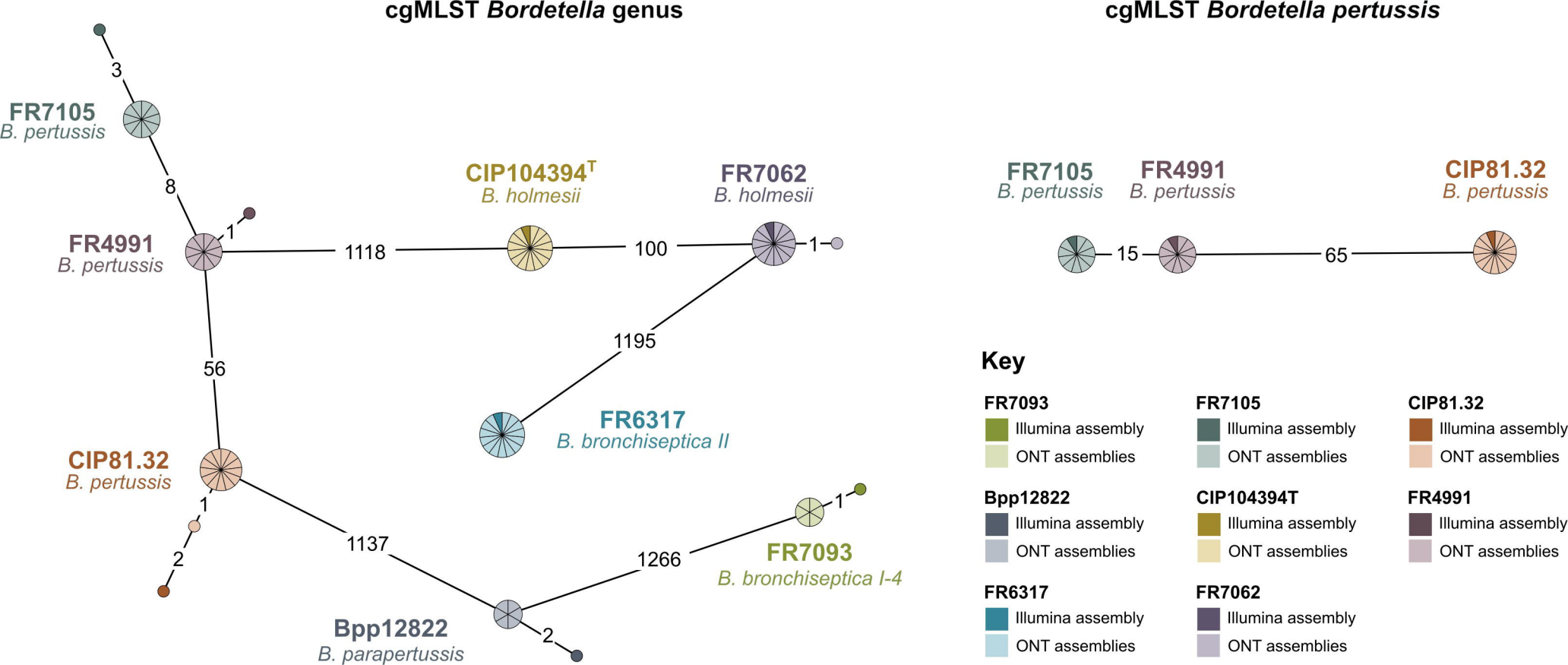
Figure legends

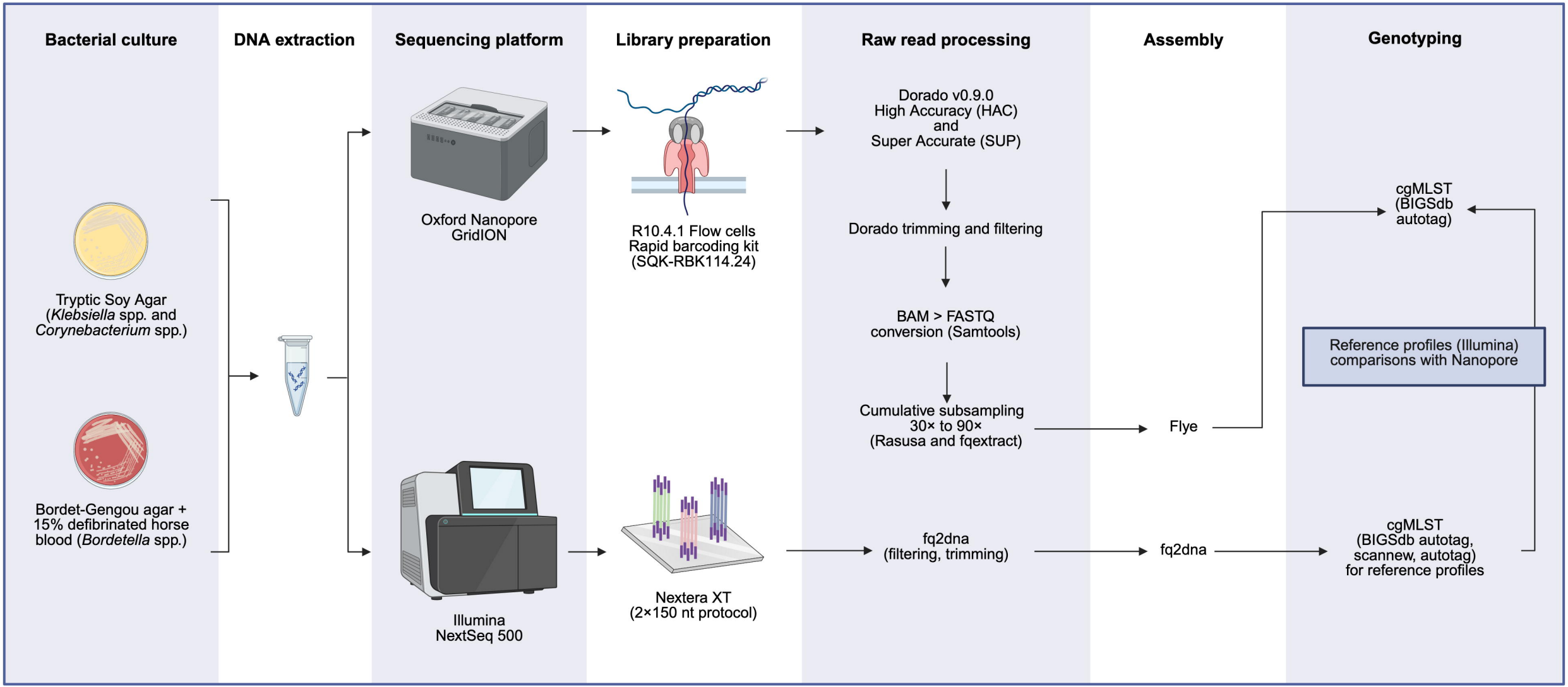
Figure 1. Bar plots showing the percentages of cgMLST allelic mismatches between short-read assemblies generated with Illumina and long-read assemblies generated with Nanopore R10.4.1 sequencing from data basecalled with Dorado SUP v0.9.0. Mismatches are shown at different simulated depths of coverage, which were obtained with a cumulative subsampling strategy. Data on the left represents SUP assemblies without raw read correction (shown in green), whereas data on the right shows mismatches for assemblies generated from HERRO-corrected reads (orange). Results for the *C. rouxii* isolate FRC0190^{T} were excluded from this figure (see Results section).

Figure 2. Minimum Spanning Trees of *Bordetella pertussis* and other *Bordetella* species investigated in this work (computed with GrapeTree). Core genome multilocus sequence typing (cgMLST) profiles used for pairwise comparisons in these trees were generated from i) Illumina genome assemblies (dark triangles); ii) Oxford Nanopore Technology (ONT) genome assemblies (including all assemblies from different simulated coverage depths), whose raw data was basecalled with Dorado SUP v0.9.0 (lighter colors). The tree on the left was generated from cgMLST profiles of the cgMLST_genus scheme, whereas the tree on the right from cgMLST profiles of the cgMLST_pertussis scheme (uniquely applied to *B. pertussis* genomes). Edges numbers indicate allelic distances between entries (edge length: log-scale).

Figure 3. Graphical summary showing the experimental workflow followed in this study.




cgMLST Bordetella genus
FR7105
B. pertussis
3
8
FR4991
B. pertussis
1
1118
CIP104394T
B. holmesii
100
FR7062
B. holmesii
1
1195
56
FR6317
B. bronchiseptica II
CIP81.32
B. pertussis
1
2
1137
1266
1
FR7093
B. bronchiseptica I-4
Bpp12822
B. parapertussis
2
cgMLST Bordetella pertussis
FR7105
B. pertussis
FR4991
B. pertussis
CIP81.32
B. pertussis
15
65
Key
FR7093
Illumina assembly
ONT assemblies
FR7105
Illumina assembly
ONT assemblies
CIP81.32
Illumina assembly
ONT assemblies
Bpp12822
Illumina assembly
ONT assemblies
CIP104394T
Illumina assembly
ONT assemblies
FR4991
Illumina assembly
ONT assemblies
FR6317
Illumina assembly
ONT assemblies
FR7062
Illumina assembly
ONT assemblies


Bacterial culture
DNA extraction
Sequencing platform
Library preparation
Raw read processing
Assembly
Genotyping
Tryptic Soy Agar
(Klebsiella spp. and
Corynebacterium spp.)
Bordet-Gengou agar +
15% defibrinated horse
blood (Bordetella spp.)
Oxford Nanopore
GridION
R10.4.1 Flow cells
Rapid barcoding kit
(SQK-RBK114.24)
Dorado v0.9.0
High Accuracy (HAC)
and
Super Accurate (SUP)
Dorado trimming and filtering
BAM > FASTQ
conversion (Samtools)
Cumulative subsampling
30× to 90×
(Rasusa and fqextract)
Flye
cgMLST
(BIGSdb
autotag)
Reference profiles (Illumina)
comparisons with Nanopore
Illumina
NextSeq 500
Nextera XT
(2×150 nt protocol)
fq2dna
(filtering, trimming)
fq2dna
cgMLST
(BIGSdb autotag,
scannew, autotag)
for reference profiles