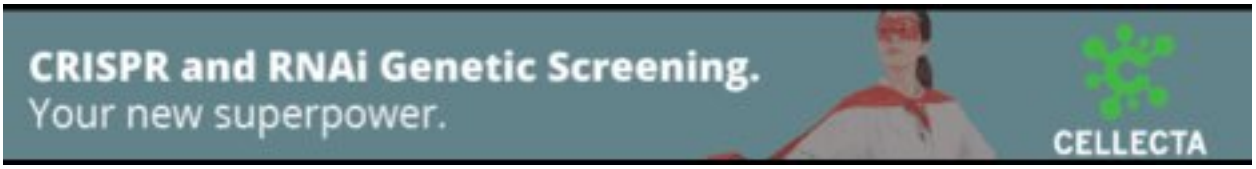
Accurate genotyping of three major respiratory bacterial pathogens with ONT R10.4.1 long-read sequencing

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

Genome Res. 2025 35: 1758-1766 originally 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,[1,2] Valérie Bouchez,[1,2] Martin Rethoret-Pasty,[1] Virginie Passet,[1,3] Sylvain Brisse,[1,2,3] and Chiara Crestani[1]

[1]*Institut Pasteur, Université Paris Cité, Biodiversity and Epidemiology of Bacterial Pathogens, 75015 Paris, France;* [2]*Institut Pasteur, National Reference Center for Whooping Cough and Other Bordetella Infections, 75015 Paris, France;* [3]*Institut Pasteur, National Reference Center for Corynebacteria of the Diphtheriae Species Complex, 75015 Paris, France*

High-throughput massive parallel sequencing has significantly improved bacterial pathogen genomics, diagnostics, and epidemiology. Despite its high accuracy, short-read sequencing struggles with the 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 is 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 generate and analyze genome assemblies with different basecalling models, at different simulated depths of coverage. ONT assemblies are 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 are consistently <0.5% for each cgMLST scheme, indicating that ONT R10.4.1 data are suitable for high-resolution genomic typing applied to outbreak investigations and public health surveillance.

[Supplemental material is available for this article.]

Whole-genome sequencing has revolutionized the study of bacterial pathogens, emerging as a crucial tool for molecular diagnostics and epidemiology and as a cornerstone of public health and clinical microbiology (Revez et al. 2017; Bagger et al. 2024; Doll et al. 2024). Over the past two decades, short-read sequencing technologies have dominated the research field and the market for molecular diagnostics and public health surveillance owing 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 and Maiden 2010; Jolley et al. 2018), 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 owing to complex, repetitive genomic regions such as insertion sequences (ISs) and other repetitive elements (Ring et al. 2018). Short reads also struggle to reconstruct extrachromosomal 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, in terms of both price per genome and acquisition cost of these sequencing platforms. These factors, along with limited portability, hinder their use in small laboratories 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 owing 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 regard to typing and outbreak investigation (Quick et al. 2015; Liou et al. 2020);

Corresponding author: chiara.crestani@pasteur.fr
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279829.124.

35:1758–1766 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/25; www.genome.org

however, higher error rates of early ONT chemistries (e.g., R9) compared with Illumina data (Jain et al. 2017; Dohm et al. 2020) 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 bioinformatic factors (Sanderson et al. 2023, 2024; Wagner et al. 2023; Lerminiaux et al. 2024; Soto-Serrano et al. 2024), 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), and *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 with 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 extrachromosomal 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 haplotype-aware error correction (HERRO)–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 with assemblies derived from uncorrected reads (Supplemental Figs. S3–S5). Most of the latter demonstrated circularization of the chromosome (contigs >2 Mbp) across the three genera (Supplemental Figs. S3–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 data set, 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 one to two larger plasmids (51 kbp and 68 kbp, respectively) compared with 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 Figs. S4, 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 *K. pneumoniae* species complex (KpSC), a high concordance in the detected circular elements was observed between assemblies from HAC and SUP data (Supplemental Figs. S3, 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, although 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 with 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 with the SUP model, as it was already low in HAC-generated assemblies (Supplemental Fig. S7). However, in a few cases HAC assemblies showed a nonnegligible number of allelic mismatches, whereas SUP basecalling allowed recovery of almost all the correct alleles (Supplemental Fig. S7). In particular, *Corynebacterium rouxii* FRC0190^T shows a very high number of mismatches in HAC data (more than 200 at high coverage depths) (Supplemental Fig. S7) compared with SUP data (between five and 27) (Supplemental Figs. S7, S8), with the latter being still significantly higher than any other *Corynebacterium* isolate sequenced in this study (Supplemental Fig. S8). This is very likely because of species-specific DNA modifications, such as methylation motifs that are uncommon or unique to *C. rouxii*, not being well represented in the training data sets of the basecalling model, especially because *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. (Fig. 1) and >35× for CdSC, with an error rate <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* (Fig. 1; Supplemental Figs. S7, S8) and *Bordetella* (for both the

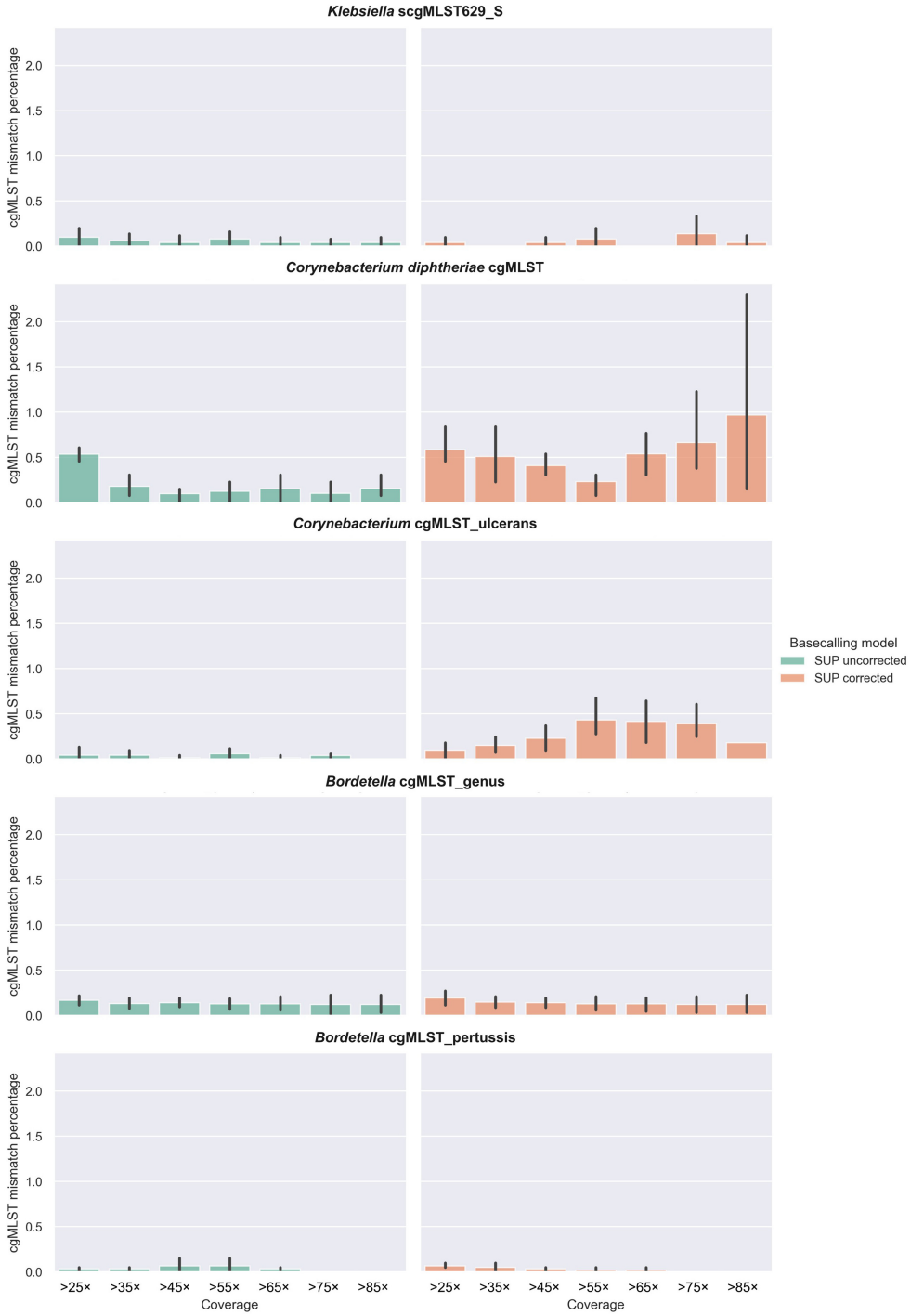


Figure 1. Bar plots showing the percentages of cgMLST allelic mismatches between short-read assemblies generated with Illumina and long-read assemblies generated with Nanopore R10.4.1 sequencing from data basecalled with Dorado SUP v0.9.0. Mismatches are shown at different simulated depths of coverage, which were obtained with a cumulative subsampling strategy. Data on the *left* represent SUP assemblies without raw read correction (shown in green), whereas data on the *right* show 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).

cgMLST_genus and cgMLST_pertussis schemes) (Fig. 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× (Fig. 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), although 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, as well as the negligible improvements of Medaka polishing on these assemblies, results reported from here onward 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 number (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.

K. pneumoniae *species complex*

cgMLST profiles of ONT-assembled genomes had a maximum of two allelic mismatches compared with 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 either are complete or are missing one or more loci owing 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 are 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 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 because of 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 *Corynebacterium ulcerans* were part of the central genotype with the Illumina reference on MSTrees generated with GrapeTree (Supplemental Fig. S11).

B. 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 *Bordetella holmesii* (Fig. 2) and *Bordetella bronchiseptica* II. For the remaining isolates, including *B. pertussis*, *Bordetella 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 (Fig. 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 regard 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 (Fig. 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) owing to higher error rates compared with 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 with 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 Figs. S1, S7).

With regard 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 extrachromosomal elements of genome assemblies from uncorrected reads. This is particularly remarkable for the genus *Bordetella*, in which high genome fragmentation

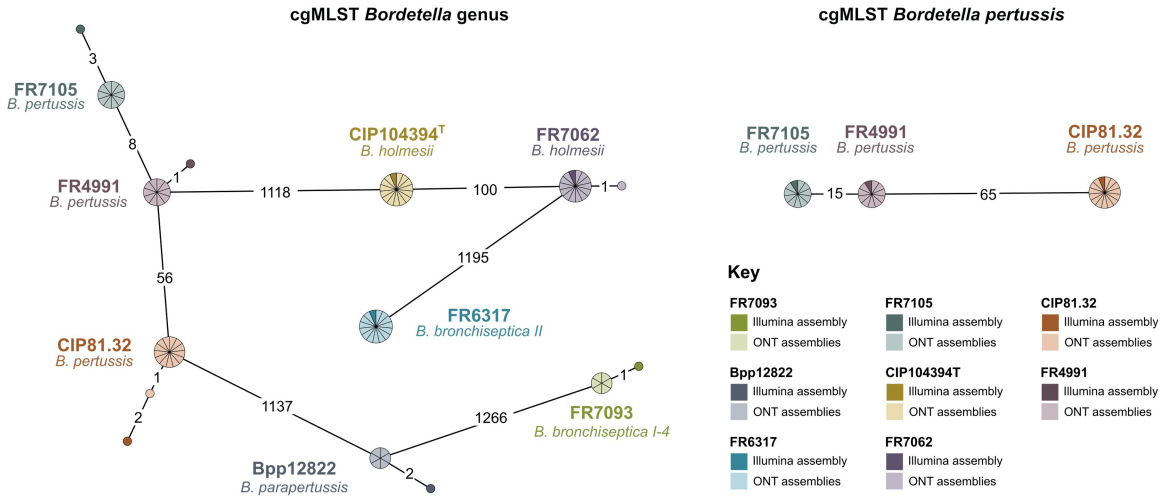


Figure 2. Minimum spanning trees of *Bordetella pertussis* and other *Bordetella* species investigated in this work (computed with GrapeTree). Core genome multilocus sequence typing (cgMLST) profiles used for pairwise comparisons in these trees were generated from (1) Illumina genome assemblies (dark triangles) and (2) Oxford Nanopore Technology (ONT) genome assemblies (including all assemblies from different simulated coverage depths), whose raw data were 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).

owing 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 extrachromosomal 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 data set, 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 less allelic mismatches.

In addition, although preliminary analyses from other authors suggested promising performances of HERRO on microbial genomes (Wick 2024a,b), 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, although 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 has not 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 Figs. S3–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 <10 kbp, which represented most of our data (Supplemental Fig. S1), resulting in an overall reduction of genome assembly coverage.

Although we currently advise to use ONT-generated genomes only for tagging existing cgMLST alleles and not for defining new ones owing 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 either were identical or only differed in the last three bins. Here, our data show 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 regard 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 (Guglielmini et al. 2021; Crestani et al. 2025) 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 one to three allelic mismatches. This atypical observation stems from the fact that 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 (Fig. 2). When investigating epidemics of *B. pertussis*, it is recommended to use the cgMLST_pertussis scheme (current threshold of three to four 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 (1) rising cases of whooping cough in multiple world regions in 2024 (Fu et al. 2023; European Centre for Disease Prevention and Control 2024; Pan American Health Organization 2024; Rodrigues et al. 2024), (2) one of the largest diphtheria outbreaks of recent times in West Africa (Balakrishnan 2024; Samarasekera 2024; World Health Organization 2024) and diphtheria resurgence in Europe (Hoefer et al. 2025), and (3) the rising importance of multidrug-resistant infections caused by *K. pneumoniae* (Antimicrobial Resistance Collaborators 2022; World Health Organization 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: (1) the KpSC collection, (2) the French national reference center (NRC) for CdSC collection, and (3) the collection of the French NRC for whooping cough and other *Bordetella* infections.

Isolate growth and DNA extraction

KpSC and CdSC were plated on tryptic soy agar (TSA) and grown for 24 h at 37°C and for 24–48 h at 35°C–37°C, respectively. *Bordetella* spp. isolates were grown for 24 to 72 h at 36°C on Bordet–Gengou agar (Becton Dickinson), supplemented with 15% defibrinated horse blood (BioMérieux), and subcultured in the same medium for 24 h in standardized conditions, as previously described (Bouchez et al. 2018).

DNA extraction was performed on a Maxwell RSC instrument (Promega) with the Maxwell RSC Blood DNA Kit (Promega) 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) on NextSeq 500 or NextSeq 2000 apparatuses (Illumina) with a 2 × 150 nt paired-end protocol.

Libraries for long-read sequencing were prepared with the Rapid Barcoding Kit V14 (SQK-RBK114.24; ONT) and sequenced on three R10.4.1 flow cells (FLO-MIN114, one per pathogen) on a GridION machine for 72 h. 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 (Fig. 3). All combinations used to generate the final assemblies can be found in Table 1, and all bioinformatic commands can be found in the 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 were 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 were 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×).

We also wanted to test the effect of HERRO v1 (Stanojević 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 (Wick 2024a,b). To this end, each subsampled read set was corrected with HERRO through its integrated version within Dorado.

De novo assembly

Short-read sequence data were assembled with fq2dna v21.06 (https://gitlab.pasteur.fr/GIPhy/fq2dna). Subsampled long-read

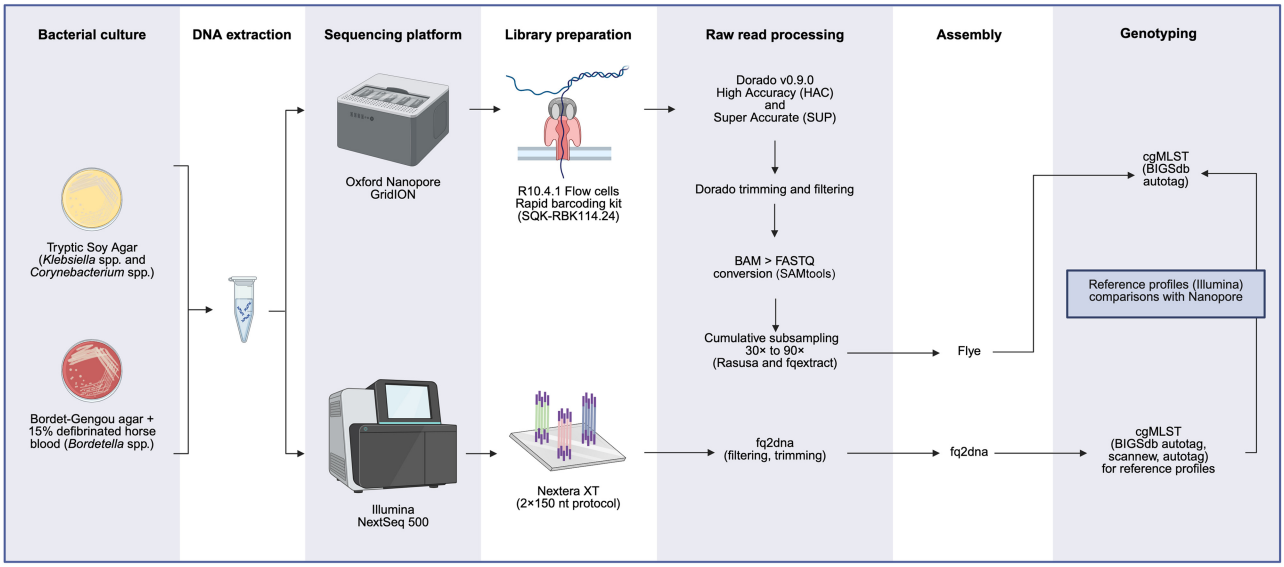


Figure 3. Graphical summary showing the experimental workflow followed in this study.

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 and Maiden 2010) and subsequently tagged long-read assemblies for these alleles. The cgMLST schemes used for genotyping were as follows: (1) for KpSC isolates, the scgMLST629_S scheme (including 629 loci) (Hennart et al. 2022); (2) for *C. diphtheriae* and *C. rouxii*, the *C. diphtheriae* cgMLST scheme (1305 loci) (Guglielmini et al. 2021), and for *C. ulcerans*, the cgMLST_ulcerans scheme (1628 loci) (Crestani et al. 2025); and (3) for *Bordetella* species (*B. pertussis*, *B. parapertussis*, *B. bronchiseptica*, and *B. holmesii*), the cgMLST_genus scheme (1415 loci) (Bridel et al. 2022). Additionally, we used the cgMLST_pertussis scheme (2038 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 versus ONT assemblies were computed with Python pandas v1.4.3 (https://github.com/pandas-dev/pandas). Missing loci from short-read reference genomes were not considered for the analysis. The type of mismatches obtained by this comparison method include (1) spurious SNPs matching by chance alleles

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×	N	Flye v2.9.5	r1041_e82_400bps_hac_v5.0.0
		40×	Y		NA
	dna_r10.4.1_e8.2_400bps_sup@v5.0.0	50×	N		r1041_e82_400bps_sup_v5.0.0
		60×	Y		NA
		70×			
		80×			
		90×			

already existing in the database and (2) 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 the 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 Illumina sequencing 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 ONT 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 and 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 competing interests.

Acknowledgments

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 also thank Alexis Criscuolo for his support in bioinformatics developments and analyses. 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 and Melinda Gates Foundation funded project "An integrated platform for *K. pneumoniae* genomic surveillance" (funder project reference INV-025280).

Author contributions: Conceptualization, methodology, and visualization were by C.C. Data curation, validation, and formal analysis were by C.C., C.R., V.B., and M.R.-P. Experimental work was by N.Z. and V.P. Writing of the original draft was by C.C. Reviewing and editing were by C.C., C.R., V.B., and S.B. Resources and funding acquisition was by S.B.

References

Antimicrobial Resistance Collaborators, Murray CJL, Ikuta KS, Sharara F, Swetschinski L, Robles Aguilar G, Gray A, Han C, Bisignano C, Rao P, et al. 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 RJ, van Schaik W, Schürch AC. 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 FO, Borgwardt L, Jespersen AS, Hansen AR, Bertelsen B, Kodama M, Nielsen FC. 2024. Whole genome sequencing in clinical practice. *BMC Med Genomics* **17:** 39. doi:10.1186/s12920-024-01795-w

Balakrishnan VS. 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 MCJ, 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, Zidane N, Brémont S, Badell E, Criscuolo A, Brisse S. 2025. Microevolution and genomic epidemiology of the diphtheria-causing zoonotic pathogen *Corynebacterium ulcerans*. *Nat Commun* **16:** 4843. doi:10.1038/s41467-025-60065-0

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10:** giab008. doi:10.1093/gigascience/giab008

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34:** 2666–2669. doi:10.1093/bioinformatics/bty149

Dohm JC, 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 AL, Palmore TN. 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 AP. 2016. fqtools: an efficient software suite for modern FASTQ file manipulation. *Bioinformatics* **32:** 1883–1884. doi:10.1093/bioinformatics/btw088

European Centre for Disease Prevention and Control. 2024. Increase of pertussis cases in the EU/EEA, 8 May 2024. ECDC, Stockholm.

Foster-Nyarko E, Cottingham H, Wick RR, Judd LM, Lam MMC, Wyres KL, Stanton TD, Tsang KK, David S, Aanensen DM, 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

Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA. 2014. Accuracy of next generation sequencing platforms. *Next Gener Seq Appl* **1:** 1000106. doi:10.4172/jngsa.1000106

Fu P, Zhou J, Meng J, Liu Z, Nijiati Y, He L, Li C, Chen S, Wang A, Yan G, et al. 2023. Emergence and spread of MT28 ptxP3 allele macrolide-resistant *Bordetella pertussis* from 2021 to 2022 in China. *Int J Infect Dis* **128:** 205–211. doi:10.1016/j.ijid.2023.01.005

Guglielmini J, Hennart M, Badell E, Toubiana J, Criscuolo A, Brisse S. 2021. Genomic epidemiology and strain taxonomy of *Corynebacterium diphtheriae*. *J Clin Microbiol* **59:** e0158121. doi:10.1128/JCM.01581-21

Hall MB. 2022. Rasusa: randomly subsample sequencing reads to a specified coverage. *J Open Source Soft* **7:** 3941. doi:10.21105/joss.03941

Hall MB, Wick RR, Judd LM, Nguyen AN, Steinig EJ, Xie O, Davies M, Seemann T, Stinear TP, Coin L. 2024. Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data. *eLife* **13:** RP98300. doi:10.7554/eLife.98300.2

Hennart M, Guglielmini J, Bridel S, Maiden MCJ, Jolley KA, Criscuolo A, Brisse S. 2022. A dual barcoding approach to bacterial strain nomenclature: genomic taxonomy of *Klebsiella pneumoniae* strains. *Mol Biol Evol* **39:** msac135. doi:10.1093/molbev/msac135

Hoefer A, Seth-Smith H, Palma F, Schindler S, Freschi L, Dangel A, Berger A, D'Aeth J, Cordery R, Delgado-Rodriguez E, et al. 2025. *Corynebacterium diphtheriae* outbreak in migrant populations in Europe. *N Engl J Med* **392:** 2334–2345. doi:10.1056/NEJMoa2311981

Jain M, Tyson JR, Loose M, Ip CLC, Eccles DA, O'Grady J, Malla S, Leggett RM, Wallerman O, Jansen HJ, et al. 2017. MinION analysis and reference consortium: phase 2 data release and analysis of R9.0 chemistry. *F1000Res* **6:** 760. doi:10.12688/f1000research.11354.1

Johnson J, Soehnlen M, Blankenship HM. 2023. Long read genome assemblers struggle with small plasmids. *Microb Genom* **9:** mgen001024. doi:10.1099/mgen.0.001024

Jolley KA, Maiden MCJ. 2010. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11:** 595. doi:10.1186/1471-2105-11-595

Jolley KA, Bray JE, Maiden MCJ. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* **3:** 124. doi:10.12688/wellcomeopenres.14826.1

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37:** 540–546. doi:10.1038/s41587-019-0072-8

Lerminiaux N, Fakharuddin K, Mulvey MR, Mataseje L. 2024. Do we still need Illumina sequencing data? Evaluating Oxford Nanopore Technologies R10.4.1 flow cells and the Rapid v14 library prep kit for Gram negative bacteria whole genome assemblies. *Can J Microbiol* **70:** 178–189. doi:10.1139/cjm-2023-0175

Linde J, Brangsch H, Hölzer M, Thomas C, Elschner MC, Melzer F, Tomaso H. 2023. Comparison of Illumina and Oxford Nanopore Technology for genome analysis of *Francisella tularensis, Bacillus anthracis*, and *Brucella suis*. *BMC Genomics* **24:** 258. doi:10.1186/s12864-023-09343-z

Liou CH, Wu HC, Liao YC, Yang Lauderdale TL, Huang IW, Chen FJ. 2020. nanoMLST: accurate multilocus sequence typing using Oxford Nanopore Technologies MinION with a dual-barcode approach to multiplex large numbers of samples. *Microb Genom* **6:** e000336. doi:10.1099/mgen.0.000336

Palma F, Hennart M, Jolley KA, Crestani C, Wyres KL, Bridel S, Yeats CA, Brancotte B, Raffestin B, David S, et al. 2024. Bacterial strain nomenclature in the genomic era: life identification numbers using a gene-by-gene approach. bioRxiv doi:10.1101/2024.03.11.584534

Pan American Health Organization. 2024. *Epidemiological alert pertussis (whooping cough) in the region of the Americas, 22 July 2024*. PAHO/WHO, Washington, District of Columbia.

Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, Mayer G. 2018. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci Rep* **8:** 10950. doi:10.1038/s41598-018-29325-6

Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, Nair S, Neal K, Nye K, Peters T, et al. 2015. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol* **16:** 114. doi:10.1186/s13059-015-0677-2

Revez J, Espinosa L, Albiger B, Leitmeyer KC, Struelens MJ. 2017. Survey on the use of whole-genome sequencing for infectious diseases surveillance: rapid expansion of European national capacities, 2015–2016. *Front Public Health* **5:** 347. doi:10.3389/fpubh.2017.00347

Ring N, Abrahams JS, Jain M, Olsen H, Preston A, Bagby S. 2018. Resolving the complex *Bordetella pertussis* genome using barcoded nanopore sequencing. *Microb Genom* **4:** e000234. doi:10.1099/mgen.0.000234

Rodrigues C, Bouchez V, Soares A, Trombert-Paolantoni S, Aït El Belghiti F, Cohen JF, Armatys N, Landier A, Blanchot T, Hervo M, et al. 2024. Resurgence of *Bordetella pertussis*, including one macrolide-resistant isolate, France, 2024. *Euro Surveill* **29:** 2400459. doi:10.2807/1560-7917.ES.2024.29.31.2400459

Samarasekera U. 2024. Diphtheria outbreak in West Africa. *Lancet Infect Diss* **24:** e87. doi:10.1016/S1473-3099(24)00026-4

Sanderson ND, Kapel N, Rodger G, Webster H, Lipworth S, Street TL, Peto T, Crook D, Stoesser N. 2023. Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction. *Microb Genom* **9:** mgen000910. doi:10.1099/mgen.0.000910

Sanderson ND, Hopkins KMV, Colpus M, Parker M, Lipworth S, Crook D, Stoesser N. 2024. Evaluation of the accuracy of bacterial genome reconstruction with Oxford Nanopore R10.4.1 long-read-only sequencing. *Microb Genom* **10:** 001246. doi:10.1099/mgen.0.001246

Soto-Serrano A, Li W, Panah FM, Hui Y, Atienza P, Fomenkov A, Roberts RJ, Deptula P, Krych L. 2024. Matching excellence: Oxford Nanopore Technologies' rise to parity with Pacific Biosciences in genome reconstruction of non-model bacterium with high G+C content. *Microb Genom* **10:** 001316. doi:10.1099/mgen.0.001316

Stanojević D, Lin D, Nurk S, Florez de Sessions P, Šikić M. 2024. Telomere-to-telomere phased genome assembly using HERRO-corrected simplex nanopore reads. bioRxiv doi:10.1101/2024.05.18.594796

Wagner GE, Dabernig-Heinz J, Lipp M, Cabal A, Simantzik J, Kohl M, Scheiber M, Lichtenegger S, Ehricht R, Leitner E, et al. 2023. 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* **61:** e0163122. doi:10.1128/jcm.01631-22

Waskom ML. 2021. seaborn: statistical data visualization. *J Open Source Soft* **6:** 3021. doi:10.21105/joss.03021

Wick RR. 2024a. HERRO read correction – part 1. Ryan Wick's bioinformatics blog. Available from: https://rrwick.github.io/

Wick RR. 2024b. HERRO read correction – part 2. Ryan Wick's bioinformatics blog. Available from: https://rrwick.github.io/

Wick RR, Judd LM, Wyres KL, Holt KE. 2021. Recovery of small plasmid sequences via Oxford Nanopore sequencing. *Microb Genom* **7:** 000631. doi:10.1099/mgen.0.000631

World Health Organization. 2024. Weekly Regional Diphtheria Bulletin 008: 26 May 2024. Available from: https://iris.who.int/handle/10665/376981

Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, Carriço JA, Achtman M. 2018. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res* **28:** 1395–1404. doi:10.1101/gr.232397.117

Zhou Z, Alikhan NF, Mohamed K, Yulei F, the Agama Study Group, Achtman M. 2020. The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res* **30:** 138–152. doi:10.1101/gr.251678.119

Received October 3, 2024; accepted in revised form May 28, 2025.