METHODS ARTICLE

MINTyper: an outbreak-detection method for accurate and rapid SNP typing of clonal clusters with noisy long reads

Malte B. Hallgren,[1] Søren Overballe-Petersen,[2] Ole Lund,[1] Henrik Hasman[2] and Philip T. L. C. Clausen [ID] [1],*

[1]National Food Institute, Technical University of Denmark, Kemitorvet 204, 2800 Kgs. Lyngby, Denmark and
[2]Department of Bacteria, Parasites and Fungi, Statens Serum Institut, 2300 Copenhagen, Denmark

*Correspondence address. National Food Institute, Technical University of Denmark, Kemitorvet 204, 2800 Kgs. Lyngby, Denmark; E-mail: plan@dtu.dk

Abstract

For detection of clonal outbreaks in clinical settings, we present a complete pipeline that generates a single-nucleotide polymorphisms-distance matrix from a set of sequencing reads. Importantly, the program is able to handle a separate mix of both short reads from the Illumina sequencing platforms and long reads from Oxford Nanopore Technologies' (ONT) platforms as input. MINTyper performs automated reference identification, alignment, alignment trimming, optional methylation masking, and pairwise distance calculations. With this approach, we could rapidly and accurately cluster a set of DNA sequenced isolates, with a known epidemiological relationship to confirm the clustering. Functions were built to allow for both high-accuracy methylation-aware base-called MinION reads (hac_m Q10) and fast generated lower-quality reads (fast Q8) to be used, also in combination with Illumina data. With fast Q8 reads a higher number of base pairs were excluded from the calculated distance matrix, compared with the high-accuracy methylation-aware Q10 base-calling of ONT data. Nonetheless, when using different qualities of ONT data with corresponding input parameters, the clustering of isolates were nearly identical.

Keywords: ONT; SNP; clustering; bioinformatics

Introduction

Until the 21st century the field of microbial diagnostics was dominated by non-computational methods. These methods ranged from cultivation and microscopic visualization to a wide variety of laboratory-based assay-technologies. Shared short-comings of these methods were long diagnostic times and/or relatively low precision. Often the identity of a pathogenic isolate could not be determined with greater accuracy than the sample's species or genus, and it could take several days, if not weeks, to perform additional tests [1]. The introduction of DNA sequencing in the late 1970s by Sanger, and the subsequent improvements of the concept in the form of second and third generation sequencing, has allowed for better and faster analysis of microbes at a phylogenetic level [2]. Illumina second generation sequencing technology have dominated the market for the previous 10 years, as it allows for precise and cost-effective sequencing, when a large pool of samples are sequenced together using multiplexing [3].

Due to the requirement of sample multiplexing in order to make Illumina platforms cost-effective in a clinical setting, researchers are looking for more agile sequencing alternatives with a shorter turnaround time. The ONT MinION platform offers great potential due to the low cost of the machine, low average sequencing price per run, and short turnaround time, thus allowing for much smaller pools of samples to be sequenced.

One of the most significant factors that currently is preventing third-generation long-read sequencing platforms from replacing the short-read sequencers is the increased error rate of long-read sequencing technologies [4]. For some research purposes, an increased error rate can be overcome, but especially when working with genetics and phylogeny, where the difference between defining if an isolate is part of an outbreak or not may be down to a few single-nucleotide polymorphisms (SNPs), error-prone sequences can distort the analysis.

When screening for clonal bacterial outbreaks, a widely used method is to perform a SNP-typing analysis [5]. Assuming that the SNPs are the result of random point-mutations, i.e. not a result of recombination, SNP distances between isolates can be used as measurements of relatedness.

Here, we present for the first time an automated SNP-typing method (MINTyper), that based on long-read sequencing can infer the same clonal clusters as methods based on short-read sequencing. The method was validated on a test set containing 12 *Escherichia coli* isolates separated into two subgroups: (i) Outbreak isolates of ST410 type ($n = 6$) and (ii) non-outbreak isolates of ST410 type ($n = 6$) based on their epidemiological relationship (patients travel and hospitalization data). Furthermore, MINTyper has been designed to handle a mix of short-read and long-read sequencing samples. This enables comparison of historical data from older sequencing platforms with new data from long read sequencing platforms.

Materials and methods

Complete pipeline

MINTyper is a complete pipeline to identify and cluster clonal bacterial outbreak strains, including automatic identification of a bacterial reference sequence. The pipeline is composed of five main steps: Identification of reference sequence (if one is not provided by the user), reference guided multiple alignment, trimming of alignments, SNP calling and finally clustering the sequences using Neighbor-Joining or infer phylogenetic relationships using maximum likelihood. MINTyper outputs a distance matrix in relaxed phylip format, a list of identified template candidates, the highest scoring template sequence, a vcf with identified variants in each genome, a Newick file of the phylogenetic tree generated from either Neighbor-Joining or a maximum-likelihood approach, a list describing which cluster each isolate belongs to, and a log file. Additionally, the alignment data generated by KMA is also preserved in a specified data folder. The pipeline is freely available as open-source at: https://bitbucket.org/genomicepidemiology/mintyper.git and as web-service at: https://cge.cbs.dtu.dk/services/MINTyper.

Data

The data used to test the performance of MINTyper originates from 12 *E. coli* isolates that has been sequenced by Statens Serum Institut (SSI) in Denmark. The 12 isolates had previously been studied using Illumina sequencing to perform Multi-Locus

Sequence Typing (MLST) and core-genome MLST (cgMLST) analysis [6]. It was found that all of the isolates were of the sequence type 410 (ST410) [6]. Six isolates (Ec01–Ec06) were all of the same bacterial clone (cgMLST type CT587) from patients who had been hospitalized in Denmark concurrently at the same ward and thus were part of the same outbreak. The remaining six isolates (Ec07–Ec12) were acquired from patients, who had never been hospitalized with any of the 11 other patients in this study. Rather, these six patients had previously been hospitalized outside of Denmark and after returning to Denmark for further treatment had been found to be colonized or infected with isolates belonging to six non-related cgMLST types. This knowledge of the phylogenetic relationship of the isolates will be used to benchmark the quality of the final distance matrix calculated by MINTyper.

Each bacterial isolate was sequenced using both Illumina's MiSeq sequencer and Oxford Nanopore's MinION MK1B sequencer. For Illumina sequencing, the DNA was extracted using the QIAGEN DNeasy Blood & Tissue kit and the library was prepared with the Nextera XT kit. Afterward, sequencing adaptors were removed and reads end-trimmed to quality $Q \geq 20$ using Trimmomatic v0.36 [7]. For ONT sequencing, the DNA was extracted with the Beckman Coulter GenFind v2 kit with a DynaMag-2 magnet. The library was prepared with the 1-D ligation barcoding protocol followed by sequencing with a R9.4.1 flow cell. Base-calling, demultiplexing, and conversion to fastq format from the raw fast5 reads were done using Albacore v2.3.4. Sequencing adapters were removed with Porechop v0.2.3 [8]. Finally, quality filtering to $Q \geq 8$ was done using NanoFilt [9]. For the high-accuracy Q10 ONT reads, the base-calling was performed with Guppy 3.6.0 with high-accuracy methylation-aware configuration.

Identification of reference sequence

The Center for Genomic Epidemiology provides a variety of sequence databases that are pre-indexed for use with KMA alignment and are automatically updated weekly by pulling changes from NCBI Resource Coordinators [10]. From these, a database containing a total of 20 377 complete bacterial chromosomal reference sequences, excluding plasmid sequences, was downloaded from: http://www.cbs.dtu.dk/public/CGE/databases/KmerFinder/version/20201028/bacteria.tar.gz. This database was searched using KMA v1.3.8 [11] with the "-Sparse" option, which identifies the reference in the database with the largest k-mer overlap to a set of query sequences.

Reference-guided multiple alignment

Reads were aligned to the reference using KMA v1.3.8 [11], with the preset option "-mint2" for the Illumina sequences and "-mint3" for the ONT sequences. In addition to aligning the reads, KMA produces a consensus sequence for each sample, where single positions are signified with upper-case bases at significant positions and lower-case bases at positions that did not fulfill the SNP-calling criteria. These criteria are for the "-mint2" preset option: reads are mapped unambiguously and positions have a depth of at least 10, at least 90% support, and are significantly over represented using a McNemar test with an α of 0.05. Whereas for the "-mint3" preset option these criteria are: reads are mapped unambiguously and positions have a depth of at least 10, at least 70% support, and are significantly over represented using a McNemar test with an α of 0.05. These options

have been verified in an earlier study by Forth et al. [12], revealing high quality consensus sequences from Nanopore data [12].

Trimming of alignments

Trimming multiple alignments have proven useful when determining the phylogeny between closely related sequences. Highly divergent areas are removed in order to select the conserved blocks determining the clonal relationship between the isolates [13, 14]. This ensures a phylogeny based on vertical evolution rather than horizontal evolution, where entire blocks of sequences are inherited in a single evolutionary event. In addition, this eliminates genomic regions with low sequencing quality. For outbreak investigation, it is important to ensure a high specificity over sensitivity in order to distinguish whether a strain is part of an outbreak or not. Therefore, core-genome SNPs are identified after trimming of alignments [15, 16]. CCPhylo v0.2.2 was developed to trim the alignments across all isolates to only include positions that were significant according to the SNP calling criteria. Additionally, SNPs located within a proximity of 10 bases of each other were trimmed away, to reduce the effect of hyper-variable regions leading to sub-optimal alignments and remove regions that likely originated from horizontal evolution [13, 14, 17, 18]. Additionally, DCM-methylation motifs (CCWGG) were trimmed away from the fast Q8 ONT data, which was base-called with Albacore, as these motifs constitute as much as 95% of discrepancies between Illumina and ONT data according to Grieg et al. [19]. The masking of DNA motifs provides an additional filtering option in comparison to other alignment trimmers, whereas the remaining options are either similar or identical to other tools, such as Gubbins [20].

Clustering and phylogenetic analysis

CCPhylo v0.2.2 was used to identify SNP differences between the isolates, and perform a subsequent hierarchical clustering using the Neighbor-Joining algorithm [21]. As an alternative to Neighbor-Joining, maximum-likelihood trees were generated using IQtree v2.0.3 [22] with parameters: "-seqtype DNA -seed 256" and FastTree v2.1.11 (compiled with double precision) [23] with the parameters: "-gtr –nt."

Performance evaluation and comparison

MINTyper was evaluated on the data set containing 12 *E. coli*, where Illumina and ONT sequenced data were treated as separate samples to test the cross-platform performance of MINTyper. MINTyper was compared with MASH v2.2 [24] and CSIPhylogeny v1.4 [18] to evaluate the performance of MINTyper compared with other methods. The CPU-time and peak memory were measured for MASH, CSIPhylogeny, and the individual parts of the MINTyper pipeline using GNU time v1.7. Where CSIPhylogeny represents a traditional approach of phylogenetic analysis, based on alignment with BWA-MEM [25], SNP-calling with samtools [26] and BEDTools [27], and phylogenetic inference with FastTree [23]. The robustness of the methods was further validated on three separate data sets with annotated outbreaks from three separate species, 70 *Acinetobacter baumannii* [28], 17 *Citrobacter freundii* [29], and 10 *Klebsiella pneumoniae* [28], which were all sequenced on the Illumina platform.

Results and discussion

Automated reference detection

The best matching reference for the dataset of 12 ST410 *E. coli* was automatically identified by MINTyper as "*E. coli* strain AMA1167 chromosome, complete genome", both when searching the Illumina and ONT data as a combined dataset and individually. This result was anticipated, as this reference sequence is the published complete genome from the same Danish outbreak that six of the input samples belong to Overballe-Petersen et al. [30].

Trimming of alignments

The resulting output from running the MINTyper pipeline with the 12 ST410 *E. coli* isolates was a distance matrix and a Newick file for each of the three runs: One using fast Q8 ONT data with no trimming, one using fast Q8 ONT data with a minimum proximity of 10 bases between SNPs and DCM masking, and one using hac_m Q10 ONT data along with a minimum proximity of 10 between SNPs and no DCM masking. The ΔSNP between the Illumina reads and the ONT MinION reads can be seen in Table 1.

The results of Table 1 show that 11 out of 12 isolates had completely identical consensus sequences after DCM masking and alignment trimming was used with fast base-called Q8 data. Where Ec09_ST410_CT527 had a single SNP discrepancy.

As expected, the hac_m Q10 data no longer had most of the SNPs caused by the DCM sites. However, Ec07_ST410_CT512, Ec09_ST410_CT527, Ec11_ST410_CT527, and Ec12_ST410_CT278 had one to three SNP discrepancies each.

Distance matrices after alignment trimming with the proximity: 0, 10, 20, 50, and 100 between called SNPs have been included in Supplementary File S1 for the combination of Illumina with ONT fast Q8 or ONT hac_m Q10, respectively.

Clustering and phylogenetic analysis

The three multiple alignments corresponding to the parameter setting described in Table 1 were clustered using Neighbor-Joining,

Table 1: overview of the number of SNPs differences between the consensus sequences generated by sequencing the same isolate on an Illumina platform and ONT MinION platform without trimming alignments and DCM methylation masking on fast Q8 data, alignment trimming, and DCM methylation masking on fast Q8 data and alignment trimming but not DCM methylation masking on hac_m Q10 data

Isolate name	ΔSNP Q8	ΔSNP Q8 with masking	ΔSNP Q10
Ec01_ST410_CT587	28	0	0
Ec02_ST410_CT587	28	0	0
Ec03_ST410_CT587	28	0	0
Ec04_ST410_CT587	28	0	0
Ec05_ST410_CT587	28	0	0
Ec06_ST410_CT587	28	0	0
Ec07_ST410_CT527	30	0	3
Ec08_ST410_CT611	28	0	0
Ec09_ST410_CT512	29	1	1
Ec10_ST410_CT596	28	0	0
Ec11_ST410_CT523	29	0	3
Ec12_ST410_CT278	34	0	2

Alignment trimming was performed with a minimum distance of 10 between accepted SNPs.
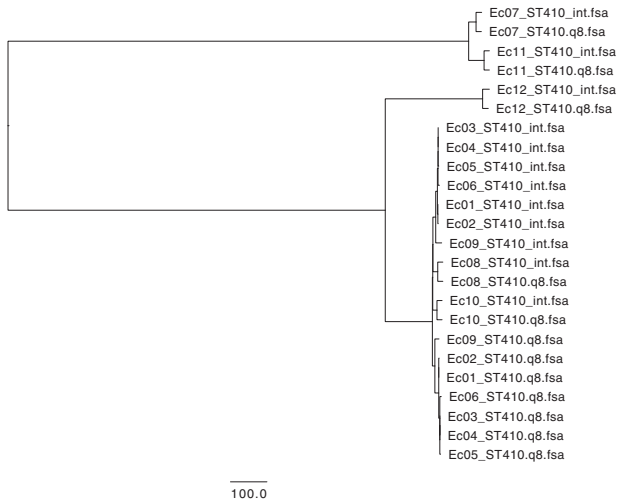
Figure 1: clustering of sequences from Illumina (denoted int) and fast base-called Q8 ONT sequences (denoted Q8) of 12 *E. coli*, based on core genome SNPs without alignment trimming. Isolates Ec01–Ec06 are from an outbreak in Denmark, while Ec07–Ec12 originate from different foreign countries.
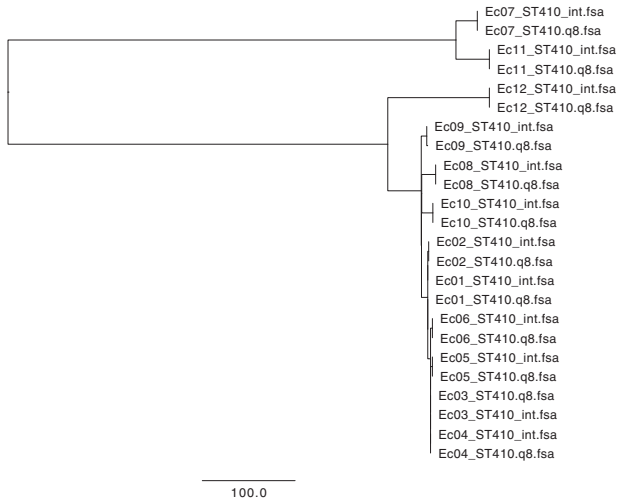

Figure 2: clustering of sequences from Illumina (denoted int) and fast base-called Q8 ONT sequences (denoted Q8) of 12 *E. coli*, based on core genome SNPs. SNPs were trimmed away if they were within a proximity of 10, together with masking of DCM methylation-sites. Isolates Ec01–Ec06 are from an outbreak in Denmark and Ec07–Ec12 originate from different foreign countries.

and were visualized using FigTree v1.4.4. Without alignment trimming, it is clear that the systematic errors from the different technologies are stronger than the true relationships between the samples (see Fig. 1). Trimming the alignments of Illumina and fast Q8 ONT data, with a minimum distance between accepted SNPs of 10, and masking out DCM methylation-sites revealed the true clustering (see Fig. 2). Each isolate, groups together between sequencing technologies, and the six outbreak strains (Ec01–Ec06) are clustered together differentiating them from the remaining isolates (Ec07–Ec12) of the same ST type. This coincide with previous studies and the epidemiological data that concluded that the isolates Ec01–Ec06 are from the same outbreak, whereas the isolates Ec07–Ec12 were acquired from different sources in different foreign countries [6]. Likewise, the clustering of Illumina and hac_m Q10 ONT data revealed the same clustering, where alignment trimming was performed with a minimum length of 10 between called SNPs
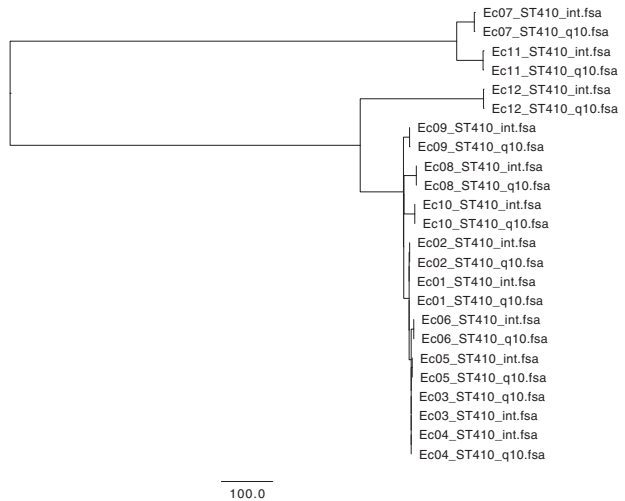

Figure 3: clustering of sequences from Illumina (denoted int) and high-accuracy methylation-aware (hac_m) base-called Q10 ONT sequences (denoted Q10) of 12 *E. coli*, based on core genome SNPs. SNPs were trimmed away if they were within a proximity of 10. Isolates Ec01–Ec06 are from an outbreak in Denmark and Ec07–Ec12 originate from different foreign countries.

(see Fig. 3). The Newick files from the trees in Figs 1–3 can be seen in Supplementary File S2. The same topology was achieved using IQtree and FastTree on the three combinations of Illumina, fast base-called ONT Q8, and hac_m ONT Q10 data (see Supplementary File S3). This suggests that the alignments and trimming of alignments determine the clustering and phylogeny to a larger degree than the methods of clustering and phylogeny themselves. This coincides with the assumptions made by most maximum-likelihood methods, which trust the alignments and are not built to differentiate between long closely related sequences, such as this study contains [22, 23]. For distantly related sequences, the maximum-likelihood methods will produce more precise results, whereas the MINTyper pipeline will fail mostly due to the reference guided multiple alignment that does not account for large rearrangements of the genome.

Loss of data

When performing either SNP or methylation-motif masking, a certain part of the data is excluded from the analysis. Since the errors in the ONT MiNION sequences are derived at the sequencing/base-calling stage, the best option to make a good clustering/phylogenetic analysis is to only look at the correctly sequenced parts of the isolates. In this study, we masked both insignificant base-calls (lower-case base-call letters), DCM motifs, and SNPs in proximity of 10 bases of each other. When using MINTyper with no motif or proximity masking on fast Q8 MiNION data, a total of 4183458/4767526 (87.7%) reference-genome bases were included in the distance matrix. When masking the DCM motifs and performing proximity masking on fast Q8 MiNION data, a total of 3835782/4767526 (80.5%) bases were included in the distance matrix.

Naturally, loosing data will lead to a less sensitive result, where true differences are at risk of being masked out. However, as was shown in Figs 1–3, we can produce more accurate results when trimming the alignments, even when it means loosing bases amounting to 8.3% of the included core-genome positions. Thus, in the case of SNP typing closely related isolates, it is more important to have a high quality of data rather than a high coverage of the core genome.

If we only employ masking of SNPs within a proximity of 10 and use the methylation-aware high-accuracy Q10 MinION data, we are able to include 4267999/4767526 (89.5%) bases in the analysis. As demonstrated in Fig. 3, the structural errors found in Fig. 1 no longer appear, and a greater fraction of base pairs is included in the analysis.

Performance evaluation and comparison

To evaluate the performance of MINTypers ability to resolve outbreaks, we challenged MASH and CSIPhylogeny with the same combinations of Illumina and ONT data as MINTyper. MASH was tested with the sketch sizes: 1024, 1048576, and 4194304 in combination with a minimum threshold of k-mer count of: 1, 2, 8, 16, and 32. For all combinations of sketch size and minimum k-mer count, MASH were not able to cluster the outbreak sequences (Ec01–Ec06) together, but clustered mostly based on sequencing technology (see Supplementary Files S4 and S5). CSIPhylogeny crashed when analyzing both combinations of the ONT data after 4 h with a peak memory of 256 GB. The crash was due to a malformed bam file, due to a flaw in the bam-format that cannot handle alignment cigars longer than 65535 operations.

In addition to the test above, MINTyper was tested with a Nanopore-only assembly as reference, using Unicycler v0.4.8-beta (minimap2 v2.17-r941, miniasm v0.3-v179, Racon 1.3.1) with default options [31–34], and concurrent polishing using Medaka v1.2.1 (minimap2 v2.17-r941, samtools v1.10) with option "-m r941_min_high_g360" [26, 32], thus skipping the automatic reference identification. Using the Nanopore-only assembly as reference revealed the same clusters as with the Illumina-Nanopore hybrid assembly of the "E. coli strain AMA1167 chromosome, complete genome" reference. Nanopore-only assembly, distance matrices, and Neighbor-Joining trees have been included in Supplementary Files S6–S8, respectively. The computational requirements, in terms of CPU hours and peak memory, of the different methods are shown in Table 2. All parts of the MINTyper pipeline can run multi-threaded, except for the automatic reference identification. This can drastically lower the wall-time compared with the CPU time. The automatic reference identification accounted for 3.3–7.6% of the total run time of MINTyper, and was the component responsible for the peak memory, except when FastTree was used to construct the trees. The alignments accounted for most of the CPU time used by MINTyper, where the Illumina samples had an average run time of 1.1 CPU-min, the fast base-called ONT Q8 used 7.9 CPU-min, and the hac_m ONT Q10 used 5.3 CPU-min on average.

Both CSIPhylogeny, MASH, and MINTyper correctly clustered the A. baumannii data set, while CSIPhylogeny and MINTyper correctly clustered the C. freundii and K. pneumoniae data sets as well (see Supplementary Files S9–S11).

Conclusion

After performing separate experiments of MINTyper's ability to cluster a set of 12 E. coli isolates with known relationships, it was found that ONT MinION long reads produced accurate clustering of the outbreak isolates with few discrepancies between sequencing technologies. This was achieved by employing KMA alignment, alignment trimming, and DCM-methylation motif-masking. It was detected that in all 12 isolates the same systematic errors occurred in the MinION fast Q8 data. By masking the DCM motifs and trimming SNPs in close proximity all of these

Table 2: computational requirements of tested methods against 12 E. coli isolates sequenced on Illumina and ONT MinIon with fast base-called Q8 and high-accuracy methylation-aware Q10 base-calling data

Method	Correct clustering	CPU time (h:mm:ss)	Peak memory
Illumina and fast base-called Q8 ONT data			
MINTyper[nj]	No	1:56:06	10.7 GB
MINTyper[1, nj]	Yes	1:56:08	10.7 GB
MINTyper[iq]	No	2:14:46	10.7 GB
MINTyper[1, iq]	Yes	1:57:58	10.7 GB
MINTyper[ft]	No	3:47:52	24.9 GB
MINTyper[1, ft]	Yes	3:34:46	23.9 GB
MINTyper[3, nj, *]	Yes	1:54:56	1.5 GB
MASH[5, nj]	No	0:20:14–0:48:32	2.8 MB–2.3 GB
MASH[6, nj]	No	0:24:06–4:12:51	0.3–11.6 GB
MASH[7, nj]	No	0:44:26–5:03:50	1.3–29.6 GB
Illumina and high-accuracy Q10 ONT data			
MINTyper[2, nj]	Yes	1:23:01	10.1 GB
MINTyper[2, iq]	Yes	1:27:47	10.1 GB
MINTyper[2, ft]	Yes	2:54:16	24.1 GB
MINTyper[4, nj, *]	Yes	1:27:48	1.8 GB
MASH[5, nj]	No	0:17:45–0:33:43	3.0 MB–1.2 GB
MASH[6, nj]	No	0:22:16–2:40:07	0.3–5.9 GB
MASH[7, nj]	No	0:41:15–3:28:45	1.3–16.3 GB

1: Alignment trimming; core-genome SNPs with a minimum distance of 10 between called SNPs and DCM-methylation masking, 3: Alignment trimming; core-genome SNPs with a minimum distance of 10 between SNPs and DCM-methylation masking, 4: Alignment trimming; core-genome SNPs with a minimum distance of 10 between SNPs, 5: Sketch size of 1024 with minimum k-mer occurrence thresholds varying from [1–32], 6: Sketch size of 1048576 with minimum k-mer occurrence thresholds varying from [1–32], 7: Sketch size of 4194304 with minimum k-mer occurrence thresholds varying from [1–32], nj: Neighbor-Joining tree-construction, iq: IQtree was used for tree-construction, ft: FastTree was used for tree-construction, *Ec01 from the ONT data was assembled with Unicycler, polished with Medaka, and used as reference.

errors could be removed in 11 out of 12 samples, with only one SNP discrepancy in the remaining isolate. Running the analysis using methylation-aware high-accuracy Q10 MinION data, instead of fast Q8 data, lead to one to three SNP discrepancies in four samples.

Even though the trimming of alignments and the methylation-motif masking of fast Q8 ONT data resulted in a 8.3% data reduction, compared with no alignment trimming of the same data, the clonal clustering improved. When using ONT data of a higher quality, less alignment trimming were needed, and thus the methylation-motif masking could be excluded from the analysis. However, generating MinION data of a quality greater than fast Q8 can be extremely time consuming, and thus simply masking out error-generating motifs can be an effective tool when an urgent clustering is needed. Until the sequencing technology improves to allow for consistent, quick, and precise sequencing and base-calling, MINTyper's approach to apply long-read sequencing-data of lesser quality in outbreak detection has proven itself useful in the field of genomic epidemiology.

Supplementary data

Supplementary data are available at *Biology Methods and Protocols* online.

Data availability

The source code for MINTyper is available at: https://bitbucket.org/genomicepidemiology/mintyper.git. A web-server service of MINTyper is available at:https://cge.cbs.dtu.dk/services/MINTyper/. The data set used in this article was uploaded to ENA project accession no. PRJEB38543.

References

1. Petersen LM, Martin IW, Moschetti WE *et al*. Third-generation sequencing in the clinical laboratory: exploring the advantages and challenges of nanopore sequencing. *J Clin Microbiol* 2019;**58**:e01315–19.
2. Brown CG, Clarke J. Nanopore development at oxford nanopore. *Nat Biotechnol* 2016;**34**:810–1.
3. Shokralla S, Porter TM, Gibson JF *et al*. Massively parallel multiplex DNA sequencing for specimen identification using an Illumina miseq platform. *Sci Rep* 2015;**5**:9687.
4. Wang JR, Holt J, McMillan L et al. FMLRC: hybrid long read error correction using an FM-index. *BMC Bioinformatics* 2018;**19**: 50.
5. Pearce ME, Alikhan NF, Dallman TJ *et al*. Comparative analysis of core genome MLST and SNP typing within a European *Salmonella serovar enteritidis* outbreak. *Int J Food Microbiol* 2018; **274**:1–11.
6. Roer L, Overballe-Petersen S, Hansen F *et al*. *Escherichia coli* sequence type 410 is causing new international high-risk clones. *Msphere* 2018;**3**:e00337.
7. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 2014;**30**: 2114–20 (doi: 10.1093/bioinformatics/btu170).
8. Wick RR, Judd LM, Gorrie CL *et al*. Completing bacterial genome assemblies with multiplex minion sequencing. *Microb Genomics* 2017a;**3**:e000132.
9. de Coster W, D'Hert S, Schultz DT *et al*. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 2018; **34**:2666–9 (https://doi.org/10.1093/bioinformatics/bty149).
10. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2015;**44**:D7–D19 (https://doi.org/10.1093/nar/gkv1290).
11. Clausen PTLC, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with kMA. *BMC Bioinformatics* 2018;**19**:307 (doi: 10.1186/s12859-018-2336-6).
12. Forth JH, Forth LF, King J *et al*. A deep-sequencing workflow for the fast and efficient generation of high-quality African swine fever virus whole-genome sequences. *Viruses* 2019;**11**: 846 (https://www.mdpi.com/1999-4915/11/9/846).
13. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000;**17**:540–52.
14. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* 2007;**56**:564–77.
15. Besser J, Carleton HA, Gerner-Smidt P *et al*. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin Microbiol Infect* 2018;**24**: 335–41.
16. Besser JM, Carleton HA, Trees E *et al*. Interpretation of whole-genome sequencing for enteric disease surveillance and outbreak investigation. *Foodborne Pathog Dis* 2019;**16**:504–12.
17. Leekitcharoenphon P, Kaas RS, Thomsen MCF *et al*. snptree—a web-server to identify and construct SNP trees from whole genome sequence data. *B M C Genomics* 2012;**13**:S6–14712164.
18. Kaas RS, Leekitcharoenphon P, Aarestrup FM et al. Solving the problem of comparing whole bacterial genomes across different sequencing platforms. *PLoS ONE* 2014;**9**:e104984.
19. Greig, DR JenkinsCGharbia, S. *et al*. Comparison of single-nucleotide variants identified by illumina and oxford nanopore technologies in the context of a potential outbreak of Shiga toxin-producing *Escherichia coli*. 2019;**8**:giz104 (doi: 10.1093/gigascience/giz104).
20. Croucher NJ, Page AJ, Connor TR *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using gubbins. *Nucleic Acids Res* 2015;**43**:e15.
21. Saitou N, Nei M. The neighbor-joining method—a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;**4**:406–25.
22. Minh BQ, Schmidt HA, Chernomor O *et al*. Corrigendum to: iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 2020;**37**:2461.
23. Price MN, Dehal PS, Arkin AP. Fasttree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* 2010;**5**:e9490.
24. Ondov BD, Treangen TJ, Melsted P *et al*. Mash: fast genome and metagenome distance estimation using minhash. *Genome Biol* 2016;**17**:132.
25. Li H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. arXiv 2013;1303.3997.
26. Li H, Handsaker B, Wysoker A et al.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9 (https://doi.org/10.1093/bioinformatics/btp352).
27. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2 (https://doi.org/10.1093/bioinformatics/btq033).
28. Roberts LW, Forde BM, Hurst T *et al*. *Genomic Surveillance, Characterization and Intervention of a Polymicrobial Multidrug-Resistant Outbreak in Critical Care* 2021 (https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000530).
29. Hammerum AM, Hansen F, Nielsen HL *et al*. Use of WGS data for investigation of a long-term NDM-1-producing *Citrobacter freundii* outbreak and secondary in vivo spread of blaNDM-1 to *Escherichia coli*, *Klebsiella pneumoniae* and *Klebsiella oxytoca*. *J*

Antimicrob Chemother 2016;**71**:3117–24 (https://doi.org/10.1093/jac/dkw289).

30. Overballe-Petersen S, Roer L, Ng K *et al*. Complete nucleotide sequence of an *Escherichia coli* sequence type 410 strain carrying blaNDM-5 on an IncF multidrug resistance plasmid and blaoxa-181 on an incx3 plasmid. *Genome Announc* 2018;**6**:e01542–17.

31. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* 2016;**32**:2103–10.

32. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100 (https://doi.org/10.1093/bioinformatics/bty191).

33. Vaser R, Sović I, Nagarajan N et al. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 2017;**27**:737–46.

34. Wick RR, Judd LM, Gorrie CL et al. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017b;**13**:e1005595.