Published online TBA

MINTyper: A method for generating phylogenetic distance matrices with long read sequencing data

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

[1]National Food Institute, Technical University of Denmark, Lyngby, Denmark , [2]Department of Bacteria, Parasites and Fungi, Statens Serum Institut, Copenhagen, Denmark

Received TBA, 2020; Revised TBA, 2020; Accepted TBA, 2020

ABSTRACT

In this paper we present a complete pipeline for generating a phylogenetic distance matrix from a set of sequencing reads. Importantly, the program is able to handle a mix of both short reads from the Illumina sequencing platforms and long reads from Oxford Nanopore Technologies' (ONT) platforms as input.

By employing automated reference identification, KMA alignment, optional methylation masking, recombination SNP pruning and pairwise distance calculations, we were able to build a complete pipeline for rapidly and accurately calculating the phylogenetic distances between a set of sequenced isolates with a presumed epidemiolocigal relation. Functions were built to allow for both high-accuracy base-called MinION reads (hac_m Q10) and fast generated lower-quality reads (fast Q8) to be used. The phylogenetical output when using different qualities of ONT data with correct input parameters were nearly identical, however a higher number of base pairs were excluded from the calculated distance matrix when fast Q8 reads were used.

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 shortcomings of these methods were the long diagnostic times and/or the 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 the tests (2). The introduction of DNA sequencing in the late 1970's by Sanger, and the subsequent improvements of the concept in the form of 2nd and 3rd generation sequencing, has allowed for better and faster analysis of microbes at a phylogenetic level (12). Illumina sequencing technologies 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. Oxford Nanopore Technologies' (ONT) MinION platform offers great potential due to the low cost of the machine and low average sequencing price per run, thus allowing for much smaller pools of samples to be sequenced (4).

One of the most significant factors that currently is preventing long read sequencing platforms from replacing the short read sequencers is the increased error rate of long read sequencing technologies (6). For some research purposes an increased error rate can be overcome, but especially when working with genetics and phylogeny, where the unit of measurement often is SNPs, error-prone sequences can be fatal to the analysis.

When working with bacterial outbreaks a widely used analytical method is to perform a SNP typing analysis (7). Assuming that the SNPs are the result of random mutations, i.e., not a result of recombination, SNP distances between isolates can be used as measurements of relatedness.

Here we present the first automated method to infer phylogenetic relations between isolates based solely on long read sequencing, with the same results as sole short read analysis. Furthermore, it has been designed to handle a mix of short read sequencing samples and long read sequencing samples, thus enabling a comparison of historical data produced on older sequencing platforms with new data being produced on modern sequencing platforms.

MATERIALS AND METHODS

Complete pipeline

MINTyper is a complete pipeline for determining the phylogenetical relationship between a set of sequenced isolate samples, including automatic identification of a bacterial reference sequence. Apart from the set of isolate samples, the program also requires a KMA-indexed (9) reference database comprised of complete reference genomes. For research purposes the option to give a single-contig reference file of the user's choosing was also implemented.

*To whom correspondence should be addressed. Email: plan@food.dtu.dk

For the complete usage guide, please see the usage and implementation section on MINTyper's Github page (https://github.com/MBHallgren/MINTyper) or use the web server version at (https://cge.cbs.dtu.dk/services/MINTyper).

Data

The data used to test and benchmark the performance of MINTyper originates from 12 *Escherichia coli* isolates that had been sequenced by Statens Serum Institut (SSI) in Denmark. The 12 isolates had previously been studied using Illumina sequencing data to perform MLST and cgMLST analysis, and it was found that all of the isolates were of the sequence type 410 (ST410) (8). Six of the isolates (Ec01-Ec06) were all sampled from patients who had been infected with the same bacterial clone in Denmark during the same outbreak. The remaining six isolates (Ec07-Ec12) were acquired from patients visiting other countries than Denmark. This knowledge of the isolates' phylogenetical relationship will be used to benchmark the quality of the final distance matrix calculated by MINTyper.

Each isolate was sequenced using both Oxford Nanopore's MinION sequencer and Illumina's MiSeq sequencer. The bacterial DNA for the Illumina sequencing was extracted using the DNeasy blood and tissue kit and the library preparation was done using the Nextra XT kit. After the sequence read trimming, adapter removal and quality filtering was done using Trimmomatic v0.36 (15). For the ONT sequencing the DNA extraction was done using the Agencourt GenFind v2 kit with a DynaMag-2 magnet. The library preparation was prepared with the 1D ligation barcoding kit followed by sequencing with a R9.4.1 flow cell on the MinION MK1B sequencer. Base-calling, demultiplexing and conversion to fastq format from the fast5 reads were done using Albacore v2.3.4. Adapter removal was performed with Porechop v0.2.3 (16). Finally, quality filtering were done using NanoFilt (17). For the high-accuracy Q10 ONT reads the base-calling was performed with Guppy 3.6.0 with high-accuracy methylation aware configuration.

Reference identification

The first step in the MINTyper pipeline is the identification of the best reference sequence from the given database to align the input sequences against. The best reference sequence to align against is defined as the reference that has the highest average template coverage with all the input samples. The program assumes that the user has chosen an input data set which has some sort of common denominator - otherwise a SNP based analysis is of no relevance. The reference is chosen from a KMA database of complete bacterial genomes (18), which has been indexed with "-Sparse ATG" as to greatly reduce the search time.

KMA Alignment

When the overall best reference has been found, KMA is used to align either Illumina short reads or ONT long reads (9). KMA has previously shown to be a good option for performing both long- and short read mapping and alignment (5).

The parser arguments used in the KMA alignment of the Illumina reads were "-ref_fsa, -ca, -dense and -bc 0.9". This combination of arguments allowed for the aligned consensus sequences to have "N" instead of gaps, no insertions, the same length as the reference and for only calling significant bases with more than 90% agreement. For the alignment of ONT MinION reads the parser arguments used with KMA were the same, except for the addition of "-bcNano -bc 0.7" to allow for a higher level of errors in the ONT data.

DCM Methylation masking

To ensure the quality and precision of the analysis done by MINTyper, regardless of whether the input is long or short reads, certain motifs within the input sequence need to be masked out. Long read sequencers have a higher error rate than short read sequencers, and thus the quality is traded off in return for speed, cost reductions and the ability to close chromosomes and plasmids (1). Additionally, according to David R. Grieg et al. 2019 as much as 95% of the discrepancies between the Illumina generated data and that of fast Q8 quality MinION generated data in two *E. coli* isolates could be attributed to DNA Cytosine Methylase (DCM) binding sites annotated by CC(A/T)GG (1). However, when employing time costly methylation aware high-accuracy (hac_m) base-calling using Guppy 3.6.0 many, but not all, methylation motifs such as DCM are correctly base-called and so not require additional masking (20).

An algorithm was implemented to mask the DCM sites in the consensus sequences by changing them to "N", thus excluding these columns from the calculation of the distance matrix. Furthermore, users of MINTyper can easily mask out motifs of their own choosing, if they find other motifs in their reads which are causing errors.

SNP Pruning

When working with large sets of sequencing data from different bacterial isolates the effect of recombination events can be an issue. However, it has been found that the impact of recombination can be mitigated by employing pruning of the aligned consensus sequences (10). A pruning algorithm was designed to remove SNPs within X (default 3) bases of each other and the intersecting bases between them as well.

Distance Matrix

An "All vs. All" calculation is carried out when calculating the distance matrix for all the consensus sequences. This entails excluding all columns of the matrix in which an "N" is found. Subsequently pairwise calculations are carried out between all of the samples of the remaining matrix, which results in the final distance matrix.

Additionally, the distance matrix is also converted into Newick-format using the implementation of Neighbour-joining in CCphylo(14), thereby allowing for visualization of the phylogenetical tree using software such as FigTree (11).

RESULTS AND DISCUSSION

Automated reference detection

The best matching reference for our dataset of 12 STA410 *E. coli* was identified as ”Escherichia coli strain AMA1167 chromosome, complete genome” with a reference length of 4767526 bases. This result was anticipated, as this reference sequence is the published complete genome from the same danish outbreak as six of the input samples (13).

Illumina vs. ONT

The resulting output from running the MINTyper program with the 12 ST410 *E. coli* isolates were a distance matrix and a Newick file for each of the three runs: One with no pruning or motif masking, one with prune length = 3 and DCM masking using fast Q8 ONT data, and one with using hac_m Q10 ONT data along with pruning). The SNP discrepancies between the Illumina reads and the ONT MinION reads can be seen in Table 1.

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 while performing recombination mitigating pruning but not any DCM methylation masking on fast Q8 data, recombination mitigating pruning and DCM methylation masking on fast Q8 data and recombination mitigating pruning but not any DCM methylation masking on hac_m Q10 data. All pruning lengths were set a 3 SNPs.*

Isolate Name	Δ SNP Q8	Δ SNP Q8 with masking	Δ SNP Q10
Ec01_ST410_CT587	31	0	0
Ec02_ST410_CT587	31	0	0
Ec03_ST410_CT587	31	0	0
Ec04_ST410_CT587	31	0	0
Ec05_ST410_CT587	31	0	0
Ec06_ST410_CT587	31	0	0
Ec07_ST410_CT527	34	1	4
Ec08_ST410_CT611	31	0	0
Ec09_ST410_CT512	33	2	3
Ec10_ST410_CT596	31	0	0
Ec11_ST410_CT523	33	0	3
Ec12_ST410_CT278	32	0	0

The results of Table 1 shows that 10 out of 12 isolates had completely identical consensus sequences after DCM masking and SNP pruning was used with fast base-called Q8 data. Sample Ec07_ST410_CT512 and Ec09_ST410_CT527 had 1 and 2 SNPs in difference, respectively.

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 and Ec11_ST410_CT527 still observed 3-4 SNPs.

Phylogenetical relationship

The generated Newick files from analyzing a mix of illumina and different types of base-called ONT data can be seen in table 1 have been visualized using FigTree. The phylogenetical trees can be seen in Figure 1, Figure 2 and Figure 3. It was known from previous studies that six of the isolates (Ec01-Ec06) were from a local outbreak in Denmark, whereas the other six isolates (Ec07-Ec12) were acquired by different patients visiting foreign countries.

The structural composition of the calculated phylogenetical trees was found not to be completely identical. In Figure 1, where only KMA alignment was used to determine the phylogenetical relationships, the closely related MinION

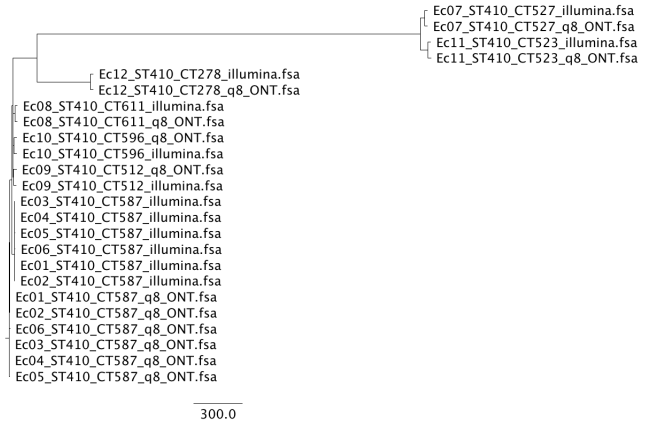

Figure 1. A visual representation of the generated distance matrix for the 12 isolates using fast Q8 MinION data and Illumina data, where no DCM masking or recombination pruning was used. Isolates Ec01-Ec06 are from an outbreak in Denmark, and Ec07-Ec12 originates from different foreign countries. A pairwise clustering of the isolates has occurred.

sequences and Illumina sequences have clustered separately. When working with closely related samples from the same local outbreak a few SNPs might be the only genetical difference. In situations like this the increased error rate in MinION sequences can result in an incorrect clustering. In Figure 2, where both DCM masking and SNP pruning was used, the MinION and Illumina sequences has clustered in pairs according to their sample numbers. Ideally, sequences from the different platforms should cluster together in pairs, since their consensus sequences should be identical.

In both Figure 1 and Figure 2 the isolates acquired independently in foreign countries cluster together correctly. This is due to their phylogenetical differences being greater than the margin of error introduced in the MinION sequencing. Additionally, when closely comparing the roots of these six isolates between the two trees, is is observed that the roots in Figure 2 are more separated which could indicate a clearer, more precise phylogenetical result.

In Figure 3 it is observed that the compositional structure is correct, and that the numerical distances are the same as in Figure 1.

Loss of data

When performing either SNP pruning or methylation masking 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 phylogenetical analysis is to try to only look at the correctly sequenced parts of the isolates. In this experiment, we both substituted insignificant base-calls (lower case base-call letters), DCM motifs and SNPs in proximity of 3 bases of each other, to ”N”. When using MINTyper with no motif masking or pruning on fast Q8 MinION data, thus only changing the insignificant base-calls, a total of 4259673 / 4767526 (89.35%) bases were included in the distance matrix. When masking the DCM motifs and performing SNP pruning on fast Q8 MinION data a total of 3504455 / 4767526 (73.50%) bases were included in the distance matrix.

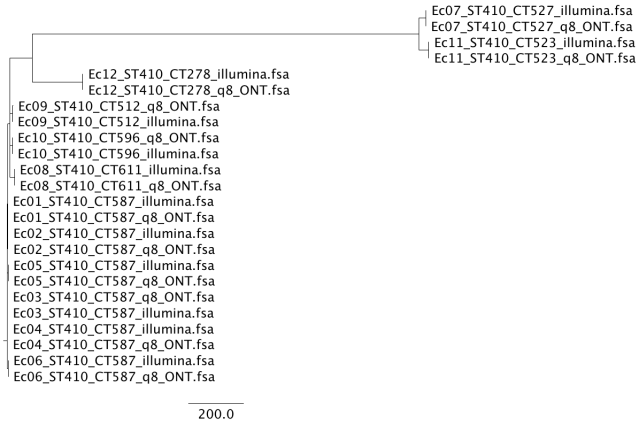

Figure 2. A visual representation of the generated distance matrix for the 12 isolates using fast Q8 MinION data and Illumina data, where both DCM masking and pruning of SNPs in proximity of 3 bases were used. Isolates Ec01-Ec06 are from an outbreak in Denmark, and Ec07-Ec12 originates from different foreign countries.. The relative SNP distances between the isolates are the nearly the same as in Figure 1, but the total number of included base pairs in the matrix is reduced, as can seen from the lenght of the scale bar.

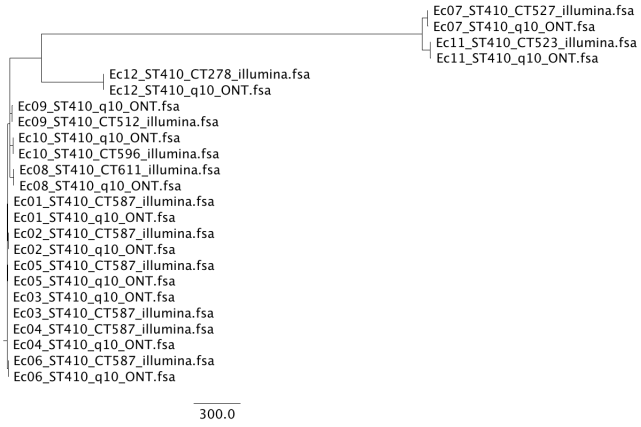

Figure 3. This figure shows the visual representation of the generated distance matrix for the 12 isolates using high accuracy base-called MinION data (Q10) and Illumina data. Here we observe that the total number of included base pairs resemble that of Figure 1. Additionally, the structural errors found in Figure 1 are not observed when using the hac_m Q10 data.

Naturally, losing data will always lead to a less confident result. However, as was shown in Figure 2, we can actually produce more accurate phylogenetical results when masking and pruning, even when it means dropping bases amounting to 15.85% of the total reference length. Thus, in the case of SNP typing analysis, it is more important to have a high quality of data rather than a high quantity.

Interestingly, if we only employ SNP pruning and instead use the methylation-aware high-accuracy Q10 MinION data, we are able to include 4276863/4767526 (89.71%) bases in the distance matrix. On top of this, as demonstrated in Figure 3, the structural errors found in Figure 1 no longer appears, and a greater fraction of base pairs are able to be included in the analysis.

CONCLUSION

After performing three separate experiments of MINTyper's ability to calculate phylogenetic distance matrices of a set of 12 *E. coli* isolates, it was found that by employing KMA alignment, recombination mitigating pruning and DCM methylation motif masking that ONT MinION long reads produced accurate phylogenetical results. It was found that in all 12 isolates the same systematic errors were occurring in the MinION fast Q8 data, and by masking the DCM motifs and pruning SNPs in close proximity or using all of these errors could be removed in 10 out of 12 samples with only one and two errors observed in the two remaining isolates. Running the analysis using methylation-aware high-accuracy Q10 MinION data instead of fast Q8 found four and three SNPs for three of the same isolates and an additional two SNPs on a third sample. These additional SNPs might have been true positives which were filtered out during the DCM masking of the fast Q8 run.

Even though the post-alignment masking function of the consensus sequences resulted in a 15.85% reduction the number of bases included in the distance matrix when using fast Q8 data, the phylogenetical analysis of the isolates' compositional relationship improved. Ideally, the base-calling errors should not occur, and when using data of a higher quality no motifs had to 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 a phylogenetical analysis of quickly generated MinION reads is desired. 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 the potential to be a game changer in the field of genomic epidemiology.

USAGE AND WEBSERVICE

The source code for MINTyper can be found at: https://github.com/MBHallgren/MINTyper.

A webserver service of MINTyper can be found at: https://cge.cbs.dtu.dk/services/MINTyper/.

The data set used in this article was uploaded to ENA project accession no. PRJEB38543.

ACKNOWLEDGEMENTS

We are grateful to Alfred F. Florensa for assisting in the setup of the web-server and Frank Hansen and Karin Sixhøj Pedersen for excellent technical help in the laboratory.

FUNDING

This project was supported by the European Union's Horizon 2020 research and innovation program under grant agreement no. 643476 (COMPARE), VEO grant agreement No. 874735 and The Novo Nordisk Foundation (NNF16OC0021856: Global Surveillance of Antimicrobial Resistance). Part of this work was supported by the Danish Ministry of Health.

The funding body did not play any role in the design of the study, writing of the manuscript nor did they have any

influence on the data collection, analysis or interpretation of the data and results.

REFERENCES

1. David R. Greig, Claire Jenkins, Saheer Gharbia and Timothy J Dallman (2019) Comparison of single-nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga toxin–producing Escherichia coli. *GigaScience*, **8**, 1-12.
2. Lauren M. Petersen,a Isabella W. Martin,a Wayne E. Moschetti,b Colleen M. Kershaw,c Gregory J. Tsongalisa (2019) Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing. *Journal of Clinical Microbiology*, **58:e01315-19.**.
3. Shokralla, Shadi and Porter, Teresita M. and Gibson, Joel F. and Dobosz, Rafal and Janzen, Daniel H. and Hallwachs, Winnie and Golding, G. Brian and Hajibabaei, Mehrdad (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform *Scientific Reports*, **5, 9687**.
4. URL: https://nanoporetech.com/products/minion *Oxford Nanopore Technologies*, **Date of visit: 14/1/2020**.
5. Jan H. Forth , Leonie F. Forth , Jacqueline King , Oxana Groza (2019) A Deep-Sequencing Workflow for the Fast and Efficient Generation of High-Quality African Swine Fever Virus Whole-Genome Sequences. *MDPI Viruses*, **11**, 846.
6. Wang et al (2018) FMLRC: Hybrid long read error correction using an FM-index. *Bmc Bioinformatics*, **19**, 50.
7. Pearce et al (2018) Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak *International Journal of Food Microbiology*, **274**, 1-11.
8. Roer L, Overballe-Petersen S, Hansen F, Schønning K, Wang M, Røder BL, Hansen DS, Justesen US, Andersen LP, Fulgsang-Damgaard D, Hopkins KL, Woodford N, Falgenhauer L, Chakraborty T, Samuelsen Ø, Sjöström K, Johannesen TB, Ng K, Nielsen J, Ethelberg S, Stegger M, Hammerum AM, Hasman H. (2018) Escherichia coli sequence type 410 is causing new international high-risk clones. *mSphere*, **337**, 18.
9. Philip T.L.C. Clausen, Frank M. Aarestrup Ole Lund, (2018) Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics*, **19**, 307.
10. Croucher N. J., Page A. J., Connor T. R., Delaney A. J., Keane J. A., Bentley S. D., Parkhill J., Harris S.R. (2014) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins *Nucleic Acids Research NAR*, **10**.
11. URL: https://github.com/rambaut/figtree/.
12. Brown, Clive G. and Clarke, James, (2016) Nanopore development at Oxford Nanopore *Nature Biotechnology*, **34**, 810-811.
13. Overballe-Petersen S, Roer L, Ng K, et al (2018). 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. *doi:10.1128/genomeA.01542-17*, **19**, 307.
14. URL: https://bitbucket.org/genomicepidemiology/ccphylo/src/master/ **Date of visit: 22/1/2020**.
15. Bolger, A. M., Lohse, M. Usadel, B., (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**.
16. https://github.com/rrwick/Porechop **Date of visit: 22/1/2020**.
17. https://github.com/wdecoster/nanofilt **Date of visit: 17/3/2020**.
18. http://www.cbs.dtu.dk/public/CGE/databases/KmerFinder/version/20190108_stable/b **Date of visit: 17/3/2020**.
19. https://github.com/nanoporetech **Date of visit: 16/5/2020**.
20. https://nanoporetech.github.io/medaka/methylation.html **Date of visit: 16/5/2020**.