

RESEARCH NOTE

Open Access



On the ability to extract MLVA profiles of *Vibrio cholerae* isolates from WGS data generated with Oxford Nanopore Technologies

Jérôme Ambroise^{1*}, Bertrand Bearzatto¹, Jean-Francois Durant¹, Leonid M. Irengé¹ and Jean-Luc Gala¹

Abstract

Objective Multiple-Locus Variable Number of Tandem Repeats (VNTR) Analysis (MLVA) is widely used to subtype pathogens causing foodborne and waterborne disease outbreaks. The *MLVAType* shiny application was previously designed to extract MLVA profiles of *Vibrio cholerae* isolates from whole-genome sequencing (WGS) data, and provide backward compatibility with traditional MLVA typing methods. The previous development and validation work was conducted using short (pair-end 300 and 150 nt long) reads from Illumina MiSeq and HiSeq sequencing. In this study, the *MLVAType* application was validated using long reads generated by Oxford Nanopore Technologies (ONT) sequencing platforms. In silico MLVA profiles of *V. cholerae* isolates ($n=9$) from the Democratic Republic of the Congo were generated using the *MLVAType* application on Nanopore WGS data. The WGS-derived in silico MLVA profiles were extracted from Canu (v.2.2) assemblies obtained through MinION and GridION sequencing by ONT. The results were compared to those obtained from SPAdes assemblies (v3.13.0; k-mer 175) generated from short-read (pair-end 300-bp) reference data obtained by MiSeq sequencing, Illumina.

Results For each isolate, the in silico MLVA profiles were concordant across all three sequencing methods, demonstrating that the *MLVAType* application can accurately predict the MLVA profiles from assembled genomes generated by long-reads ONT sequencers.

Keywords *In silico* MLVA profiles, Sequencing, Nanopore, Long reads, WGS

Background

Multiple-Locus Variable Number of Tandem Repeats (VNTR) Analysis (MLVA) is widely used by laboratory-based surveillance networks for subtyping pathogens causing foodborne and water-borne disease outbreaks. We recently demonstrated that WGS data generated with short-read Illumina sequencing technology can be used

to extract in silico MLVA profiles of *V. cholerae* isolates from WGS data while maintaining backward compatibility with traditional MLVA typing methods [1]. The percentage of censored estimations in MLVA profiles generated from WGS data was inversely proportional to the k-mer parameter used during genome assembly. However, preventing censored estimation was possible by using a longer k-mer size (e.g. 175) even though the original SPAdes v.3.13.0 [2] software did not propose this k-mer size.

Both MinION and GridION ONT sequencers are quickly gaining popularity because the long sequence reads enable to assemble contiguous microbial genome.

*Correspondence:

Jérôme Ambroise
jerome.ambroise@uclouvain.be

¹Center for Applied Molecular Technologies (CTMA), Institute of Clinical and Experimental Research (IREC), Université catholique de Louvain (UCLouvain), Brussels, Belgium



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

However, their base-calling accuracy is significantly lower than that obtained with Illumina short reads, although the resolution of this shortcoming is steadily improving. More specifically, it is well known that ONT sequencers have difficulty in accurately sequencing low-complexity regions, such as homopolymers [3].

Recent studies have looked into methods for deriving in silico MLVA profiles from long-read sequencing data for several bacterial species. For multidrug-resistant organisms such as *Klebsiella pneumoniae*, *Escherichia coli*, *Enterobacter cloacae*, *Citrobacter freundii*, *Pseudomonas aeruginosa*, *Acinetobacter baumannii*, and methicillin-resistant *Staphylococcus aureus*, perfect concordance was achieved between in silico MLVA profiles derived from long- and short-read data, as well as conventional MLVA typing [4]. Lower concordance rates were observed for *Bacillus anthracis* [5], where Nanopore and Illumina sequencing yielded an 88% and 83% concordance, respectively.

To the best of our knowledge, the accuracy of in silico MLVA typing using Nanopore data has not yet been assessed on *V. cholerae*. This species seems to be undergoing unprecedented genetic changes, with climate change possibly acting as a trigger factor [6, 7]. These changes pose an increasing threat to public health in cholera-affected regions. Therefore, this study aimed to compare and validate MLVA results obtained with WGS data from *V. cholerae* using MinION, GridION and Illumina sequencing, in order to expand the scope of application of our previous *MLVAtype* shiny application. We analysed the in silico MLVA profiles derived from the three methods on a series of *V. cholerae* strains. Given that that we had previously demonstrated the accuracy of MLVA profiles derived from Illumina MiSeq on *V. cholerae* isolates [1], we compared the ONT results to the Illumina results, which served as benchmark.

Method

Sample collection and sequencing technology

Nine *V. cholerae* isolates were selected from a collection of isolates characterised in a recent study conducted in the DRC between 2014 and 2017 [8].

Two technologies were used to sequence the whole genomes of selected *V. cholerae* isolates: Illumina (MiSeq) and ONT (MinION and GridION). Regarding Illumina

technology, whole genome assemblies were generated from paired-end 300 nt long reads, as previously detailed [1]. In brief, sequencing libraries were prepared using 70 ng of *V. cholerae* genomic DNA following the Illumina DNA Prep protocol (Illumina, San Diego, CA, USA). In brief, genomic DNA from *V. cholerae* isolates was simultaneously fragmented and tagged with sequencing adapters in a single step using Nextera transposome (Nextera XT DNA Library Preparation Kit, Illumina, San Diego, CA, USA). Tagged DNA was then amplified with a 12-cycle polymerase chain reaction (PCR), cleaned up with AMPure beads, and subsequently loaded on a MiSeq for a paired-end 2 × 300 nt sequencing run using MiSeq reagent kit V3 (600 cycles) (Illumina, San Diego, CA, USA).

ONT long-read libraries were generated using 400 ng of high molecular weight genomic DNA (GQN > 8). The DNA was initially fragmented to an average fragment length of 11.6 kb using Covaris g-TUBES (Covaris, Woburn, MA, USA). Libraries were then prepared and barcoded according to ONT's Ligation Sequencing genomic DNA – DNA Barcoding kit SQK-NBD112.24 protocol. The nine libraries were multiplexed and loaded into two FLO-MIN112 (R10 version) flow cells. Sequencing took 72 h on a MinION Mk1C and a GridION.

Sanger-Derived MLVA Typing was used as a reference method for resolving MLVA discrepancies. Sanger-derived MLVA typing was performed by sequencing amplicons on both strands on the ABI 3130 GA, using the BigDye Terminator v1.1 cycle sequencing kit (Applied Biosystems, USA). Motif repeats were counted manually and translated into MLVA profiles.

WGS assembly and MLVA profiling

WGS data from Illumina MiSeq were assembled into contigs using SPAdes v.3.13.0 [2] with a k-mer value of 175 and other default settings. WGS data from ONT MinION and GridION were assembled into contigs using Canu v.2.2 with genome size = 4 m and other default settings [9].

For each isolate and each sequencing platform, the in silico MLVA profiles were extracted from the assembled contigs using the *MLVAtype* algorithm, which has been implemented in an R shiny application. This application is freely available at <https://ucl-irec-ctma.shinyapps.io/NGS-MLVA-TYPING/>. It enables users to upload a list of draft genomes and the nucleotide sequences of the motifs. The application was used to predict MLVA profiles for *V. cholerae* loci listed in Table 1, as demonstrated in our previous study.

Table 1 Loci and motifs characterising the MLVA profiles of *V. Cholerae*

| Locus | Motif |
|---------|-----------|
| VC0147 | Aacaga |
| VC0437 | Gacccta |
| VC1650 | Ataatccag |
| VCA0171 | Gctgtt |
| VCA0283 | Ccagaa |

Table 2 Quality control metrics of Illumina MiSeq reads

| Isolate | Number of reads | Length of reads | Positions in forward reads with median Phred score > 30 | Positions in reverse reads median Phred score > 30 |
|-----------|-----------------|-----------------|---|--|
| CTMA-1426 | 2×2 228 605 | 2×300 | 0.983 | 0.751 |
| CTMA-1427 | 2×1 521 677 | 2×300 | 0.944 | 0.761 |
| CTMA-1432 | 2×2 195 887 | 2×300 | 0.983 | 0.761 |
| CTMA-1435 | 2×1 613 586 | 2×300 | 0.9 | 0.754 |
| CTMA-1461 | 2×2 289 895 | 2×300 | 0.987 | 0.761 |
| CTMA-1473 | 2×2 525 965 | 2×300 | 0.987 | 0.761 |
| CTMA-1402 | 2×1 990 392 | 2×300 | 0.934 | 0.744 |
| CTMA-1421 | 2×1 949 341 | 2×300 | 0.987 | 0.754 |
| CTMA-1424 | 2×1 955 144 | 2×300 | 0.924 | 0.764 |

Table 3 Quality control metrics of ONT MinION and GridION long reads

| ONT sequencing platform | Isolate | Number of reads | Length of reads min – max | median of reads length | Positions with median phred score > 10 |
|-------------------------|-----------|-----------------|---------------------------|------------------------|--|
| MinION | CTMA-1426 | 190 186 | 69–63 195 | 1599 | 0.996 |
| | CTMA-1427 | 87 929 | 70–43 611 | 1492 | 0.942 |
| | CTMA-1432 | 112 216 | 68–61 372 | 759 | 0.985 |
| | CTMA-1435 | 143 620 | 70–53 139 | 1444 | 0.990 |
| | CTMA-1461 | 163 440 | 68–109 071 | 1612 | 0.978 |
| | CTMA-1473 | 98 596 | 72–104 391 | 4162 | 0.989 |
| | CTMA-1402 | 113 252 | 75–44 694 | 2098 | 0.989 |
| | CTMA-1421 | 95 725 | 74–62 565 | 2604 | 0.991 |
| | CTMA-1424 | 1 372 705 | 65–127 859 | 624 | 0.999 |
| GridION | CTMA-1426 | 122 597 | 67–43 560 | 1840 | 0.993 |
| | CTMA-1427 | 57 920 | 70–54 700 | 1764 | 0.982 |
| | CTMA-1432 | 71 818 | 73–43 924 | 857 | 0.995 |
| | CTMA-1435 | 93 995 | 69–63 410 | 1683 | 0.991 |
| | CTMA-1461 | 10 6175 | 69–53 447 | 1869 | 0.995 |
| | CTMA-1473 | 68 365 | 71–47 336 | 4593 | 0.996 |
| | CTMA-1402 | 74 836 | 80–54 138 | 2450 | 0.995 |
| | CTMA-1421 | 63 781 | 70–78 397 | 3095 | 0.999 |
| | CTMA-1424 | 829 644 | 70–138 439 | 639 | 0.999 |

Results

Tables 2 and 3 summarise the sequence quality reported for Illumina MiSeq, MinION, and GridION, respectively. As expected, forward reads from Illumina MiSeq exhibited higher quality than reverse reads.

Despite having lower quality than Illumina, both ONT platforms produced significantly longer reads (Table 3).

As expected, assembled genomes generated by SPAdes using Illumina MiSeq reads were more fragmented with contigs (ranging from 74 to 89 contigs, compared to those produced by Canu with MinION and GridION reads, which ranged from 2 to 5 contigs). As shown in Table 4, MLVA profiles were generated using the *MLVAType* algorithm on WGS data from nine previously reported isolates [1, 8]. The results were perfectly concordant across all sequencing platforms.

Discussion

Due to its low cost and rapid turnaround time, ONT sequencing platforms such as MinION and GridION are appealing to clinical laboratories, with the clear potential to replace traditional typing methods. However, this type of analysis is not yet affordable in all institutions due to several new challenges, including data storage, computing power, and bioinformatics expertise. Moreover, sequencing with ONT platforms still faces the issue of base-calling accuracy when compared to other sequencing platforms such as Illumina short-reads sequencer [10]. Accordingly, the current study was designed to assess the impact of the lower sequencing accuracy of the ONT technology on assembled genomic region of *V. cholerae*, which is characterised by a variable number of tandem repeats, using both ONT platforms.

Table 4 MLVA profiles of *V. Cholerae* isolates obtained from WGS data using the *MLVAtype* shiny application

| Isolate | Miseq-derived MLVA profile | MinION-derived MLVA profile | GridION-derived MLVA profile |
|-----------|----------------------------|-----------------------------|------------------------------|
| CTMA-1402 | (9,7,7,10,16) | Idem | Idem |
| CTMA-1421 | (9,7,7,11,17) | Idem | Idem |
| CTMA-1424 | (10,7,7,11,15) | Idem | Idem |
| CTMA-1426 | (10,7,6,26,20) | Idem | Idem |
| CTMA-1427 | (10,7,6,24,21) | Idem | Idem |
| CTMA-1432 | (10,7,6,16,20) | Idem | Idem |
| CTMA-1435 | (10,7,6,23,18) | Idem | Idem |
| CTMA-1461 | (11,7,7,13,16) | Idem | Idem |
| CTMA-1473 | (10,7,7,12,16) | Idem | Idem |

Given that we had previously demonstrated the accuracy of Illumina MiSeq-derived MLVA profiles on *V. cholerae* isolates, we compared the ONT results to the Illumina-MiSeq results, which served as a reference. Notably, in this study, MiSeq-derived MLVA profiles were free of censored values due to two factors: (i) the limited number of repetitions per motif, with a maximum of 26 for the 4th motif of CTMA-1426, and (ii) the use of a long k-mer size (175) during genome assembly with SPAdes v.3.13.0.

We used the same nine DRC *V. cholerae* isolates as in our previous study [1] and assembled the Illumina reads into contigs with the same assembler (SPAdes). Our MLVA studies were therefore conducted in several phases, during which *Vibrio cholerae* strains were recultured between 2019 and 2023. During this process, minor variations in MLVA profiles were observed in two strains (CTMA-1424 and CTMA-1426) across passages, confirmed by Sanger sequencing. The long-term stability of MLVA profiles across passages has been explored by Kendall et al. [11] and Garcia et al. [12] in large-scale studies, where microevolution was observed in *V. parahaemolyticus* after multiple passages. In contrast, the limited number of passages (≤ 3) in our study and the low mutation rate (on the order of 10–4 mutant per generation) observed during culture by Garcia et al. [12] make it unlikely that significant MLVA changes occurred. We therefore believe that experimental conditions are more likely responsible for the observed MLVA variations. While the same initial colonies were used in our 2019 and 2023 studies, different glycerol stocks were employed. Although we cannot conclusively demonstrate this, the few differences in MLVA profiles are more plausibly attributable to these technical factors than to microevolutionary processes. A large-scale study on *V. cholerae*, involving many strains and numerous passages (e.g., $n > 30$), falls beyond the current scope of this study but is planned for the near future.

In conclusion, the perfect concordance of results across short- and long-reads sequencing platforms in this study demonstrates the reliability and accuracy of in silico MLVA typing using Nanopore WGS data. While ongoing

technological progress is expected to improve ONT base-calling accuracy in the near future, this study confirms that the currently reported lower accuracy of ONT long-read sequencing, compared to short-read Illumina sequencing, does not affect MLVA typing results.

Limitations

As demonstrated in this study, using short- and long-read sequencing for backward comparison with historical MLVA profiles obtained through traditional methods can introduce bias due to unpredictable MLVA profile variations across passages of the same strain. These variations may result from technical factors (e.g., reculturing strains from different aliquots or randomly analysing different colonies from the same culture plate), genetic microevolution, or a combination of both. Consequently, further investigation is needed to assess the variability among multiple colonies from the same culture plate and the long-term stability of MLVA profiles across numerous passages in a larger-scale study.

Although the limited number of *V. cholerae* isolates in this study could be considered a genuine limitation, this is counterbalanced by the broad diversity of MLVA profiles included in the analysis and the perfect concordance observed across traditional, short-read, and long-read sequencing methods for MLVA profiling.

Abbreviations

| | |
|------|--|
| ONT | Oxford Nanopore Sequencing |
| WGS | Whole Genome Sequencing |
| MLVA | Multiple-Locus Variable Number of Tandem Repeats (VNTR) Analysis |

Acknowledgements

Not applicable.

Author contributions

J.A. and J.L.G. wrote the main manuscript text. B.B. J.F.D and L.I. collected the isolates and performed short and long-reads sequencing. All authors reviewed the manuscript.

Funding

This study was funded by the Belgian Cooperation Agency of the ARES (Académie de Recherche et d'Enseignement Supérieur) [grant COOP-CONV-20-022]. The funder did not play any role in the study design, collection, analysis, and interpretation of data, manuscript writing, or the decision to submit the paper for publication.

Data availability

All NGS data are available from the European Nucleotide Archive (ENA, <http://www.ebi.ac.uk/ena>), available under study accession number PRJEB55717.

Declarations

Ethics approval and consent to participate

The study protocols, including the method for collecting rectal swab samples and conducting genomic analysis, were approved by the Ethical Review Board (ERB) of the Institut Supérieur des Techniques Médicales de Bukavu, RDC (ISTM-BUKAVU/CRPS/CIES/ML/0016/2023). The ERB explicitly granted a waiver of written informed consent, citing compliance with international ethical guidelines for research conducted during severe outbreaks (e.g., CIOMS 2016). These guidelines recommend the use of oral informed consent for participants with low literacy, particularly in public health emergencies where written consent may be impractical. In accordance with this waiver, all participants provided oral informed consent prior to sample collection. The study also adhered to national regulations of the Democratic Republic of Congo governing research ethics in public health emergencies. To ensure participant confidentiality, all collected samples were fully anonymised during the genomic analysis process.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 13 September 2024 / Accepted: 8 January 2025

Published online: 16 January 2025

References

1. Ambrose J, Irengre LM, Durant JF, Bearzatto B, Bwire G, Stine OC, et al. Backward compatibility of whole genome sequencing data with MLVA typing using a new MLVAtype shiny application for *Vibrio cholerae*. *PLoS ONE*. 2019;14(12):e0225848.
2. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455–77.
3. Delahaye C, Nicolas J. Sequencing DNA with nanopores: troubles and biases. *PLoS ONE*. 2021;16(10):e0257521.
4. Landman F, Jamin C, de Haan A, Witteveen S, Bos J, van der Heide HG, et al. Genomic surveillance of multidrug-resistant organisms based on long-read sequencing. *medRxiv*. 2024;202402:18–24301916.
5. Linde J, Brangsch H, Hölzer M, Thomas C, Elschner MC, Melzer F, et al. Comparison of Illumina and Oxford Nanopore Technology for genome analysis of *Francisella tularensis*, *Bacillus anthracis*, and *Brucella suis*. *BMC Genomics*. 2023;24(1):258.
6. Irengre LM, Ambrose J, Bearzatto B, Durant J-F, Bonjean M, Wimba LK et al. Genomic evolution and rearrangement of CTX- Φ prophage elements in *Vibrio cholerae* during the 2018–2024 cholera outbreaks in eastern Democratic Republic of the Congo. *Emerging Microbes & Infections*. 2024(just-accepted):2399950.
7. Chowdhury FR, Nur Z, Hassan N, von Seidlein L, Dunachie S. Pandemics, pathogenicity and changing molecular epidemiology of cholera in the era of global warming. *Ann Clin Microbiol Antimicrob*. 2017;16(1):10.
8. Irengre LM, Ambrose J, Mitangala PN, Bearzatto B, Kabangwa RKS, Durant JF, et al. Genomic analysis of pathogenic isolates of *Vibrio cholerae* from eastern Democratic Republic of the Congo (2014–2017). *PLoS Negl Trop Dis*. 2020;14(4):e0007642.
9. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 2017;27(5):722–36.
10. Petersen LM, Martin IW, Moschetti WE, Kershaw CM, Tsongalis GJ. Third-generation sequencing in the Clinical Laboratory: exploring the advantages and challenges of Nanopore sequencing. *J Clin Microbiol*. 2019;58(1).
11. Kendall EA, Chowdhury F, Begum Y, Khan AI, Li S, Thierer JH, et al. Relatedness of *Vibrio cholerae* O1/O139 isolates from patients and their household contacts, determined by multilocus variable-number tandem-repeat analysis. *J Bacteriol*. 2010;192(17):4367–76.
12. Garcia K, Gavilan RG, Hofle MG, Martinez-Urtaza J, Espejo RT. Microevolution of pandemic *Vibrio parahaemolyticus* assessed by the number of repeat units in short sequence tandem repeat regions. *PLoS ONE*. 2012;7(1):e30823.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.