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

Jérôme Ambroise [1*], Bertrand Bearzatto [1], Jean-Francois Durant [1], Leonid M. Irenge [1], Jean-Luc Gala [1]

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

* Corresponding author

Jérôme Ambroise : jerome.ambroise@uclouvain.be

Bertrand Bearzatto : bertrand.bearzatto@uclouvain.be

Jean-Francois Durant : jean-francois.durant@uclouvain.be

Leonid M. Irenge : leonid.irenge@uclouvain.be

Jean-Luc Gala : Jean-luc.gala@uclouvain.be

Abstract

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. The *MLVAType shiny* application was previously designed to extract MLVA profiles of *Vibrio cholerae* isolates from WGS data, and enable backward compatibility with traditional MLVA typing methods. The previous development and validation work was done on short (pair-end 300 and 150 nt long) reads from Illumina MiSeq and Hiseq sequencing. In this work, the use of the *MLVAType* application was validated on long reads generated with Oxford Nanopore Technologies (ONT) sequencing platforms. Accordingly, the MLVA profiles of *V. cholerae* isolates (n=9) from the Democratic Republic of the Congo were obtained using the *MLVAType*

application on WGS data. The WGS-derived MLVA profiles were extracted from canu (v.2.2) assemblies obtained after the MinION and GridION, ONT, sequencing. The results were compared to those obtained from SPAdes assemblies (v3.13.0; k-mer 175) generated from short-read (pair-end 300-bp) data obtained by MiSeq sequencing, Illumina, taken as a reference. For each isolate, the MLVA profiles were concordant for all three sequencing methods, demonstrating that the *MLVAType* application can accurately predict the MLVA profiles from assembled genomes generated with long-reads ONT sequencers.

Keywords:

MLVA profiles, sequencing, nanopore, long reads, WGS

Short title: MLVA profiles of *Vibrio cholerae* using Oxford Nanopore Technologies

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. Recently, we have shown that WGS data generated with short-read Illumina sequencing technology can be used to extract MLVA profiles of *V. cholerae* isolates from WGS data while ensuring backward compatibility with traditional MLVA typing methods [2]. MLVA profiles generated from WGS data were affected by censored estimations whose percentage 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 if the original SPAdes v.3.13.0 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. 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].

The main objective of the present study is to extent the field of application of our previous *MLVAtype* shiny application by comparing and validating MLVA results obtained when using WGS data obtained with MinION and GridION sequencing. Accordingly, we compared the MLVA profiles obtained with the three methods on a series of *V. cholerae* strains.

Method

Nine *V. cholerae* isolates were selected from a collection of isolates characterized in a recent study conducted between 2014 and 2017 in the DRC [1].

Whole genome sequencing of selected *V. cholerae* isolates was performed using three sequencers and two technologies: Illumina (Miseq) and ONT (MinION and GridION). Regarding Illumina technology, sequencing libraries were prepared starting from 70 ng of genomic DNA from DRC isolates according to Illumina DNA Prep protocol (Illumina, San Diego, CA, USA). Briefly genomic DNA were simultaneously fragmented and tagged with sequencing primers in a single step using Bead linked transposomes. Tagged DNAs were then amplified with a 12-cycle polymerase chain reaction (PCR) to generate sequencing ready DNA fragments by adding indexes and adapters. Final libraries were then cleaned up with AMPure beads, and subsequently loaded on a MiSeq for a paired-end 2 x 300 nt sequencing run using MiSeq Reagent kit V3 (600 cycles) (Illumina, San Diego, CA, USA).

ONT Long reads libraries were generated starting from 400 ng of high molecular weight genomic DNA (GQN > 8). DNA was firstly fragmented to an average fragment length of 11,6 kb using Covaris g-TUBES (Covaris, Woburn, WA, USA). Libraries were then prepared and barcoded according to the Ligation Sequencing genomic DNA – DNA Barcoding kit SQK-NBD112.24) protocol from ONT. The 9 libraries generated were multiplexed and loaded on 2 FLO-MIN112 (R10 version) flow cells. Sequencing was undertaken on a MinION Mk1C and a GridION for 72 hours.

WGS data generated with Illumina MiSeq were assembled into contigs using SPAdes v.3.13.0 [7] and a k-mer value of 175. Considering that this k-mer value is longer than what is offered by this version of the software, editing the software source code was required. WGS data generated with ONT MinION and GridION were assembled into contigs using canu v.2.2.

For each isolate and each sequencing platform, the MLVA profiles were extracted from the assembled contigs using the MLVAtype algorithm implemented in an R shiny application and freely available at https://ucl-irec-ctma.shinyapps.io/NGS-MLVA-TYPING/. This application enables the user to upload a list of draft genomes and the nucleotide sequences of the motifs. It was used to predict MLVA profiles associated to the *V. cholerae* loci listed in Table 1 as 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

Results

Table 2 and Table 3 provide an overview of sequence quality is reported for Illumina MiSeq, MinION and GridION , respectively. As expected, the forward Illumina MiSeq reads were of higher quality than the reverse reads.

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 x 2 228 605	2 x 300	0.983	0.751
CTMA-1427	2 x 1 521 677	2 x 300	0.944	0.761
CTMA-1432	2x 2 195 887	2 x 300	0.983	0.761
CTMA-1435	2 x 1 613 586	2 x 300	0.9	0.754
CTMA-1461	2 x 2 289 895	2 x 300	0.987	0.761
CTMA-1473	2 x 2 525 965	2 x 300	0.987	0.761
CTMA-1402	2 x 1 990 392	2 x 300	0.934	0.744
CTMA-1421	2 x1 949 341	2 x 300	0.987	0.754
CTMA-1424	2 x 1 955 144	2 x 300	0.924	0.764

Despite having lower quality than Illumina, both ONT platforms produced reads that were much longer (Table 2).

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

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

Discussion

Due to its low cost and rapid turnaround time, ONT sequencing MinION and GridION are attractive sequencing platforms for clinical laboratories with a clear potential to replace traditional typing methods. However, this type of analysis is not yet affordable in all institutions because it introduces many new challenges (*e.g.* data storage, computing power, bioinformatics expertise). Moreover, sequencing with ONT platforms still faces the problem of base-calling accuracy compared to other sequencing platforms such as Illumina short-reads sequencer [4]. Accordingly, the present study was designed to assess the impact of the lower sequencing accuracy of the ONT technology on assembled genomic region characterized by variable number of tandem repeats, using the MinION and GridION sequencing platforms. Considering that we previously demonstrated the accuracy of Illumina MiSeq-derived MLVA profiles on *V. cholerae* isolates, we compared the ONT results with the Illumina results taken as a reference. We used the same *V. cholerae* isolates as previously [1, 2] and assembled the contigs with SPAdes and a long kmer (i.e., k=175) to avoid

censored VNTR values. The perfect concordance of the results generated by the 3 sequencing platforms demonstrates the backward compatibility of ONT WGS sequencing with conventional MLVA typing.

References

1. Irenge LM, Ambroise 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:e0007642.
2. Ambroise J, Irenge 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:e0225848.
3. Delahaye C, Nicolas J. Sequencing DNA with nanopores: Troubles and biases. PLoS One. 2021;16:e0257521.
4. 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.

Financial support

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.

Conflict of Interest

None

Data availability statements

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

Ethical standards:

Not applicable