Nanopore sequencing of a monkeypox virus strain isolated from a pustular lesion in the Central African Republic

Mathias Vandenbogaert, Aurélia Kwasiborski, Ella Gonofio, Stéphane Descorps-Declère, Benjamin Selekon, Andriniaina Andy Nkili Meyong, Rita Sem Ouilibona, Antoine Gessain, Jean-Claude Manuguerra, Valérie Caro, et al.

To cite this version:

Mathias Vandenbogaert, Aurélia Kwasiborski, Ella Gonofio, Stéphane Descorps-Declère, Benjamin Selekon, et al.. Nanopore sequencing of a monkeypox virus strain isolated from a pustular lesion in the Central African Republic. Scientific Reports, 2022, 12 (1), pp.10768. 10.1038/s41598-022-15073-1 . pasteur-03720772

HAL Id: pasteur-03720772
https://pasteur.hal.science/pasteur-03720772

Submitted on 12 Jul 2022

scientific reports

OPEN

Nanopore sequencing of a monkeypox virus strain isolated from a pustular lesion in the Central African Republic

Mathias Vandenbogaert[1,7], Aurélia Kwasiborski[1,7], Ella Gonofio[2,7], Stéphane Descorps-Declère[3], Benjamin Selekon[2], Andriniaina Andy Nkili Meyong[4], Rita Sem Ouilibona[2], Antoine Gessain[5], Jean-Claude Manuguerra[1], Valérie Caro[1], Emmanuel Nakoune[2,8] & Nicolas Berthet[1,6,8]✉

Monkeypox is an emerging and neglected zoonotic disease whose number of reported cases has been gradually increasing in Central Africa since 1980. This disease is caused by the monkeypox virus (MPXV), which belongs to the genus *Orthopoxvirus* in the family *Poxviridae*. Obtaining molecular data is particularly useful for establishing the relationships between the viral strains involved in outbreaks in countries affected by this disease. In this study, we evaluated the use of the MinION real-time sequencer as well as different polishing tools on MinION-sequenced genome for sequencing the MPXV genome originating from a pustular lesion in the context of an epidemic in a remote area of the Central African Republic. The reads corresponding to the MPXV genome were identified using two taxonomic classifiers, Kraken2 and Kaiju. Assembly of these reads led to a complete sequence of 196,956 bases, which is 6322 bases longer than the sequence previously obtained with Illumina sequencing from the same sample. The comparison of the two sequences showed mainly indels at the homopolymeric regions. However, the combined use of Canu with specific polishing tools such as Medaka and Homopolish was the best combination that reduced their numbers without adding mismatches. Although MinION sequencing is known to introduce a number of characteristic errors compared to Illumina sequencing, the new polishing tools allow a better-quality MinION-sequenced genome, thus to be used to help determine strain origin through phylogenetic analysis.

Monkeypox is an emerging and neglected disease of zoonotic origin that presents with maculopapular rashes—particularly on the palms of the hands and soles of the feet—sometimes very similar to those of smallpox[1,2]. Infection can also be associated with adenopathy. This infection is caused by the monkeypox virus (MPXV), which belongs to the genus *Orthopoxvirus* in the family *Poxviridae*. This large virus, whose genome is around 200 kb, was first isolated in 1958 from a monkey (*Macaca fascicularis*) originating from Singapore and imported to Copenhagen, Denmark that had caused an outbreak in captive *Cynomolgus* monkeys[3]. However, the precise animal reservoirs of MPXV have yet to be identified, although this virus was isolated once from a symptomatic squirrel (*Funisciurus anerythrus*) caught in 1985 in the Democratic Republic of Congo (DRC) near a village where a human case had been reported previously and once from a sooty mangabey (*Cercocebus atys*) in 2012 in Ivory Coast[4,5]. Although orthopoxvirus antibodies are not specific to MPXV, they have been detected in a large number of animal species living in Africa, including numerous non-human primates and rodents, suggesting the

[1]Unité Environnement et Risque Infectieux, Cellule d'Intervention Biologique d'Urgence, Institut Pasteur, Paris, France. [2]Institut Pasteur de Bangui, Bangui, Central African Republic. [3]Centre of Bioinformatics, Biostatistics and Integrative Biology (C3BI), Institut Pasteur, Paris, France. [4]Centre Interdisciplinaire de Recherches Médicales de Franceville (CIRMF), Franceville, Gabon. [5]Unité d'Epidémiologie et Physiopathologie des Virus Oncogènes, Département de Virologie, UMR3569, Institut Pasteur, Centre National de la Recherche Scientifique (CNRS, Paris, France. [6]The Center for Microbes, Development and Health, CAS Key Laboratory of Molecular Virology and Immunology, Institut Pasteur of Shanghai-Chinese Academy of Sciences, Discovery and Molecular Characterization of Pathogens, No. 320 Yueyang Road, XuHui District, Shanghai 200031, China. [7]These authors contributed equally: Mathias Vandenbogaert, Aurélia Kwasiborski and Ella Gonofio. [8]These authors jointly supervised this work: Emmanuel Nakoune and Nicolas Berthet. ✉email: nicolas.berthet@pasteur.fr

presence of orthopoxviruses in wild animals[6]. In Central or West Africa, outbreaks of monkeypox are generally reported in remote populations that depend on hunting and consume bushmeat[7–9]. Although the reasons are not clear, there has been a gradual increase in the number of monkeypox cases since 1980 in Central Africa (mainly in the DRC) and the Central African Republic (CAR)[10–13]), but also more recently in West Africa, particularly in Nigeria since 2017–2018[14,15] where the previous case was reported several decades ago.Unlike the strains belonging to the West African clade, in particular that of the last epidemic in Nigeria, which have already spread in several countries outside the African continent and to Cameroon[16–18], the strains belonging to the Congo Basin clade (Central Africa) have never been identified outside their geographical region of origin. Indeed, genomic analyses of the main MPXV strains detected in the CAR between 2001 and 2018 confirmed that they belonged to three lineages closely related to those found in DRC[19]. Finally, molecular data obtained from viral genomes can trace the origin of strains and establish relationships between strains isolated from different outbreaks in the countries affected by this disease.

The rapid identification of pathogens responsible for infectious diseases and access to high quality genome sequences are clearly essential for better disease management[20]. Molecular characterization of the pathogen is generally carried out using second-generation sequencing (SGS) techniques. However, these techniques only generate short sequences and are biased according to GC content, limiting their efficiency[21]. To overcome these limits, third-generation sequencing (TGS) techniques, including Oxford Nanopore technologies (ONT) and PacBio platforms, have been widely developed and applied to study pathogens such as bacteria and viruses[22]. Unlike PacBio, ONT sequencers, especially the MinION, are better adapted to field sequencing, because they are highly portable, being small in size and weight (< 100 g). In addition, library preparation is simpler than that used for SGS. All these technical advantages of TGS have made it possible to sequence genomes rapidly in regions with extreme climate conditions[23,24].The most elegant demonstration of the usefulness of rapid MinION sequencing is that of the field for a rapid molecular characterization of Ebola virus during the last West African outbreak[25]. Real-time monitoring of the Ebola virus and the associated genomic analysis shed much-needed light on the spread of the virus and helped establish infection control strategies[26]. MinION sequencing has also been successfully used to study other viruses in epidemics, such as the Zika virus in South America, where MinION sequencing results helped link congenital malformations to Zika fever, or in the recent Lassa virus epidemic in Nigeria in 2019[27]. In addition to these examples, TGS has also been used to better understand the evolution of emerging pathogens in animals, and in particular to better characterize the African swine fever virus (ASFV). Although this virus is endemic to sub-Saharan Africa, China was the first Asian country to be affected by an ASFV epidemic in August 2018 and this virus has since spread to neighboring countries. The data suggest a single origin of the epidemic, but SGS and TGS have shown variation in genomes across cities and provinces in China[28]. TGS has not yet been used alone to sequence an MPXV genome; however, it has already been combined with SGS to sequence several strains, including the genome of the MPXV strain detected in a patient who discarded rodent carcasses at his home in Port Harcourt (Nigeria) before returning to Israel a week later[29,30].

The aim of our study was to evaluate the performance and added value of the MinION real-time TGS sequencing device for sequencing the complete genome (around 200 kb) of an MPXV strain, obtained directly from a pustular lesion sampled in a remote area of Central Africa during an outbreak. At the time of writing, this study represents the first attempt to sequence using only MinION a member of the MPXV, for which there are only a small number of reference genomes available for comparison. Due to the sparsity of available reference genomes, the reads (directly obtained from the lesion) were taxonomically binned to identify the species-origin of the reads. This binning approach is similar to a metagenomics setup using two taxonomic classifiers. Assembly of the corresponding TGS reads was done using state-of-the-art assembly methods, as implemented in the Canu tool[31]. For validation, the resulting genome sequence was compared with that obtained using a different, but more conventional method (SGS). The results we obtained illustrate the usefulness of this sequencing approach to quickly obtain a whole virus genome sequence at a level of accuracy that is precise enough for phylogeographic determination of the origin of this zoonotic virus.

Materials and methods

Organization of suspected case notification, collection of biological samples, DNA extraction of the monkeypox virus and molecular assays.

Whenever a case of monkeypox is reported in CAR, standardized data collection procedures have been developed and validated by the Ministry of Health and the World Health Organization (WHO). These procedures consist of notification of suspected cases, collection of biological samples for diagnosis and data collection in the field by the investigation team. The biological samples were then sent to the IPB in the best possible storage conditions for virological investigations. Although independent samples of several pustules were collected, DNA extraction was performed on a single pustular lesion using the QIAamp DNA Mini kit, according to the manufacturer's instructions. Extracted DNA was quantified using a Qubit dsDNA High-Sensistivity Assay kit (Invitrogen) following the protocol according to the manufacturer's guidelines and stored at −20 °C until use in molecular investigations. MPXV was detected using a quantitative polymerase chain reaction (PCR), as previously described [32,33].

MinION library preparation and sequencing.

Barcoded sequencing libraries were prepared from 400 ng DNA using the Rapid Barcoding Sequencing kit SQK-RBK004 (Oxford Nanopore Technologies) following the manufacturer's protocol. The optional clean-up steps using AMPure XP beads (Beckman Coulter) were carried out to increase throughput. The library was loaded onto an R9.4 flow cell (FLO-MIN106) and sequenced on MinION Mk1B device within 8 h. ONT MinKNOW software (version 19.05.0) was used to collect raw sequencing data and ONT's cloud-based basecaller based on Guppy (version 3.2.8) was used to perform on-site and real-time basecalling during sequencing runs. Subsequently, the "What is my pot?" (WIMP) workflow

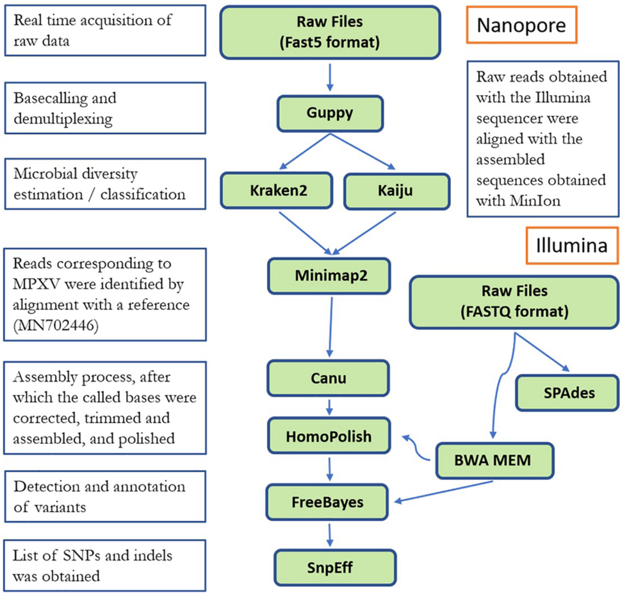
Real time acquisition of raw data
Raw Files (Fast5 format)
Nanopore
Basecalling and demultiplexing
Guppy
Raw reads obtained with the Illumina sequencer were aligned with the assembled sequences obtained with MinIon
Microbial diversity estimation / classification
Kraken2
Kaiju
Illumina
Reads corresponding to MPXV were identified by alignment with a reference (MN702446)
Minimap2
Raw Files (FASTQ format)
Assembly process, after which the called bases were corrected, trimmed and assembled, and polished
Canu
SPAdes
HomoPolish
BWA MEM
Detection and annotation of variants
FreeBayes
List of SNPs and indels was obtained
SnpEff

Figure 1. Workflow for the analysis of raw data, from real-time acquisition to the listing of SNPs and indels.

was launched for real-time species classification and estimation of the species (metagenomics) diversity of the sample (based on Centrifuge software).

Bioinformatics analyses. Unlike the previous steps which were carried out at the Institut Pasteur of Bangui (IPB), the following steps were carried out at the Institut Pasteur in Paris after the transfer of the raw data acquired in real time in files containing 4000 sequences in FAST5 format. These files were basecalled and demultiplexed using Guppy (version 3.4.1) (Fig. 1). Microbial diversity was determined by using (1) the Kraken2 taxonomic classifier based on a custom extended RefSeq database (containing the RefSeq reference libraries for viral, bacterial, fungal, archaea, protozoan and plasmid genomes/proteins, the human GRCh38 human genome/proteins, as well as the NCBI non-redundant nucleotide database containing sequences from large environmental sequencing projects), and (2) Kaiju, a protein-level classifier, using an equivalent protein database[34–36] (Fig. 1). The taxonomic classification of raw reads results were visualized on the Krona web interface[37]. The identification of reads corresponding to MPXV were confirmed by alignment with a reference MPXV (NC_003310) using Minimap2 (version 2.9) for consensus sequence process, after which the called bases were corrected, trimmed, assembled and read-corrected using Canu (version 1.8)[31,38]. A caveat of long-read technology is the introduction of sequencing errors (mostly short indels), which was thus verified and corrected by polishing using HomoPolish (Release v0.3) (Fig. 1). The workflow used for the detection and annotation of variants in the MinION dataset consisted of an alignment of the reads using Minimap2 (version 2.9)[38] to the reference sequence of the same virus (GenBank Accession MN702446) obtained by Illumina sequencing and assembled using SPAdes (version 3.10). Then, from this alignment, the list of SNPs and indels was obtained using Freebayes (version 1.1)[39]. Annotation was performed using SnpEff (version 4.3) according to the annotations of the Illumina sequence (MN702446)[40]. Finally, the raw reads obtained with the Illumina sequencer (BioProject number PRJNA680806) were compared with the assembled sequences obtained with the MinION sequencer using BWA MEM (version 0.7.4)[41,42] (Fig. 1).

Phylogenetic analyses. The genome of the MPXV 38c strain sequenced with the MinION was aligned with the 65 MPXV genomes available at NCBI using MAFFT v7.471 (2020/Jul/3)[43,44]. Sequences from two other poxvirus genomes (cowpox virus, Grisham 1990, X94355 and horsepox virus, Mongolia 1976, DQ792504) were used as outgroups. The resulting alignment was manually edited to remove poorly aligned portions, in particular the extremities containing missing data, as well as many insertions or deletions (indels). IQ-TREE was then used to reconstruct a global MPXV bootstrapped likelihood phylogenetic tree with 1000 replicates using the "best model" indicated in the IQ-TREE internal model fitter[45].

Ethical statement. The investigation of this case of monkeypox virus infection in the CAR was approved by both the Institut Pasteur of Bangui Scientific Committee and the CES (*Comité Ethique et Scientifique*; University of Bangui). All the experiments carried were done in accordance with relevant guidelines and regulations. In addition, the adult patient gave his oral and written informed consent.

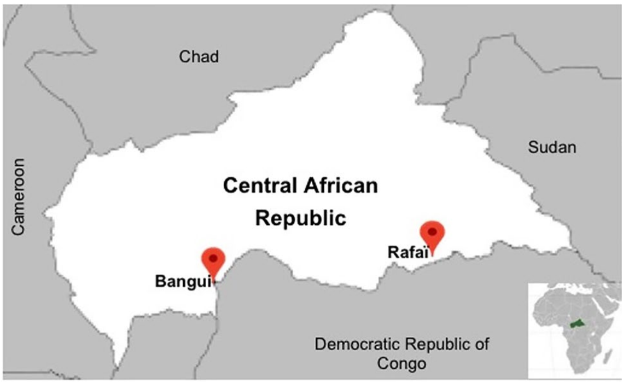

Figure 2. Map of the Central African Republic in Africa. The map was edited in Microsoft PowerPoint (Version 2020).

	Kraken2	Kaiju
	Number of reads	Number of reads
Eukaryota		
Homo sapiens	134,702	63,192
Other	0	13,730
Total	134,702 (92.86%)	76,922 (64.02%)
Bacteria		
Proteobacteria	1384	24,088
Terrabacteria group	23	7555
Chlamydiae	0	1077
Other	0	3382
Total	1407 (0.97%)	36,102 (30.02%)
Double-stranded DNA virus, no RNA stage		
Family *Poxviridae*	2168	2195
Other	–	127
Total	2198 (1.51%)	2322 (1.93%)
Unassigned	–	4841 (4.03%)
Bacterial metagenomes	6761 (4.66%)	–
Total number of reads assigned to a taxon	145,068 (100%)	120,187 (100%)

Table 1. Summary of the taxonomic assignments of the reads for each of the two classifiers used (Kraken2 and Kaiju).

Results

Description of the clinical case.

A 31-year-old street vendor living in the city of Rafai (Mbomou province, CAR) was admitted to a private health center with skin rashes on his face, chest, palms and genital mucosa in April 2018 (Fig. 2). He also had nausea, inguinal adenopathies, chills, difficulty swallowing, mouth ulcers, itchy lesions and myalgia. Following his admission, two women, aged 21 and 22, also developed skin rashes on their face, chest and genital mucosa. Although the infection of the two young women was clearly linked to the sexual relations they had had with the street vendor, epidemiological investigations were not able to determine the origin of infection for the man. Virological qPCR investigations carried out at IPB determined that the cycle threshold (CT) value for the street vendor was 19.7, and 19.8 and 18.3 for the 21- and 22-year-old women, respectively.

Taxonomic assignment of reads obtained by MinION sequencing.

A total of 146,920 raw reads were obtained with sizes ranging from 77 bp to about 68 kb for a median size of 1946 bp. A taxonomic assignment was obtained for 145,068 reads (98.7%) with the Kraken2 classifier, but only 120,187 reads (81.8%) were assigned with the Kaiju protein-level classifier (Table 1). Regardless of the classifier, the vast majority of the reads were assigned as eukaryotes. Kraken2 assigned 93% of the total reads to the *Homo sapiens* host and Kaiju assigned only 82% (63,192/76,922) of the reads to *Homo sapiens* among all the eukaryote-assigned reads (76,922/121,026). Kraken2 and Kaiju assigned 1423 and 36,102 reads respectively to the superkingdom Bacteria, including 97.26% (138/1423) and 66.72% (24,088/36,102) respectively to the phylum Proteobacteria (Table 1).

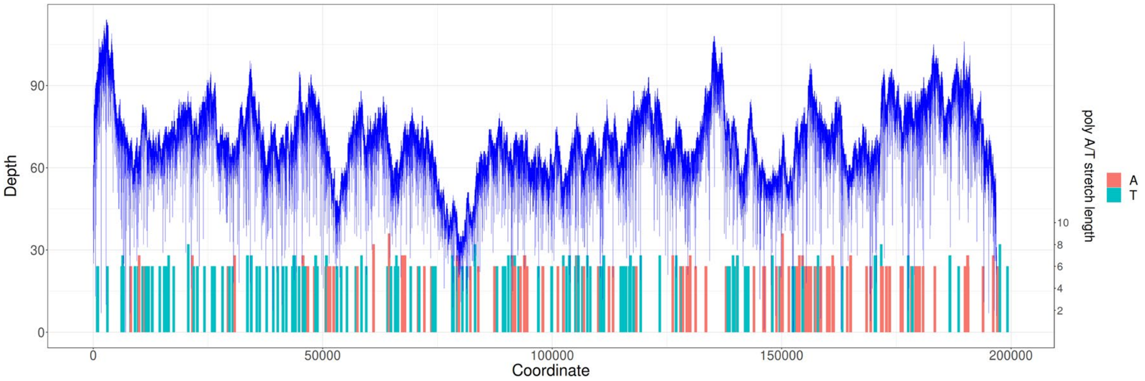

Figure 3. Genome sequencing coverage and location of A/T homopolymers in the monkeypox virus (MPXV) genome. Homopolymer repeat length at each homopolymer position (coordinates) in the genome.

On the other hand, only Kraken2 assigned 6761 reads as belonging to bacterial metagenomes. Finally, the two classifiers Kraken2 and Kaiju made it possible to highlight respectively 2198 and 2322 reads of double-stranded DNA virus sequences with no RNA stage. Of these reads, Kraken2 assigned 2168 (98.6%) and 2111 (96.4%) as belonging respectively to the family *Poxviridae* and the genus *Orthopoxvirus* and Kaiju followed the same pattern, assigning 2195 (94.5%) and 2137 (92.0%) to the same family and genus respectively. Although Kraken2 assigned 99.2% of the reads (2094/2111) to the MPXV species, Kaiju assigned only 51.9% (1110/2137) of the reads to the MPXV species. In addition, 43% (919/2137) and 5.1% (108/2137) of the reads classified in the genus *Orthopoxvirus* by Kaiju were assigned respectively as an unassigned orthopox or another orthopox species such as cowpox (Table 1). Examination of the taxonomic assignments obtained from these two classifiers showed that the total reads assigned to the family of *Poxviridae* was 2223 reads, of which 2140 reads were common to both classifiers. However, some reads were assigned to the family *Poxviridae* by only one classifier. Analyses of the taxonomic classification results showed that 28 reads were assigned by Kraken2 only, while 55 reads were assigned by Kaiju only. In parallel with the use of these classifiers, mapping all 146,920 raw reads using Minimap2 on a reference MPXV sequence (NC_003310) identified 2171 matching reads, representing about 1.47% of the total reads. Finally, the size of these reads varied between 168 bp and 44 kb for the largest reads.

Comparison of Illumina- and MinION-sequenced genomes. All 2171 MinION reads corresponding to MPXV produced a sequence (MPXV-M) of 196,956 bp with a depth ranging from 12× to 57×. In 94% of the nucleic base positions, the depth was between 30× and 59× with an average depth of 39.72×. Canu was used to both assemble reads classified as belonging to the MonkeyPox genome into draft genomes and then to also remove random sequencing errors. Comparison with the Illumina sequence showed that the two sequences aligned well, but that the MPXV-M sequence was longer than the one initially obtained with the Illumina data (MPXV-I-MN702446), which had a length of only 190,357 bp. The MPXV-I sequence begins at base 573 and ends at base 190,634 with respect to the MPXV-M sequence. However, a retrospective analysis of raw Illumina data found a total of 5924 reads that matched the two 'missing ends' of the MPXV-I sequence, even though the majority of these mapped to the terminal repeated regions of both ends of the genome (data not shown). The comparison between the MPXV-I sequence (MN702446) and the MPXV-M sequence revealed insertions or deletions located in homopolymers of at least 6 bases that have been identified in this genome (Table S1 and Fig. 3). However, the major remaining differences between the two genomic sequences were between positions 178,570 and 178,664 with MPXV-M as the reference sequence. Detailed analyses of these differences showed that these 95 bp of the MPXV-I sequence actually corresponded to a region between bases 196,170 and 196,265 of the MPXV-M sequence. However, retrospective analysis using the raw data from Illumina sequencing showed that no reads corresponding to these 95 bp from the region between 178,570 and 178,664 were available. Finally, analysis of the reference sequence (KP849471) that was used for probe design for our targeted enrichment also did not contain this 95 bp region. The absence of this 95 bp region in this reference sequence explains the absence of a capture probe targeted in this genomic region and consequently the absence in our Illumina-sequenced genome (MPXV-I).

In addition to those differences, the central coding region sequence (CRS; roughly between 56,000–120,000) in the MPXV genome is highly conserved and is flanked by variable ends that contain inverted terminal repeats (ITRs). Those ITR's comprise as much as 1% of the genome, are prone to hair-pin loop-outs [46,47] and contain at least 4 ORFs [48,49]. As such, the ITR regions represent global repeats (i.e., long sequence which is duplicated throughout the genome) and contain local repeated sequences. In the MPXV genome, local tandem repeats are found both within the ITR regions and outside these regions. In contrast to global repeats, the local repeat contains simple sequences, that are tandem duplicated, and therefore pose a special challenge to genome-assembly tools. The GenBank reference MN702446 was obtained using an Illumina short-read sequencing approach. The analysis of this sequence using Tandem Repeat Finder (TRF; [50]) showed that the left-most ITR sequences could be identified, but showed that the corresponding right-most ITR is missing in the final sequence. Figure 4 shows the tandem repeat locations on the x-axis in the MPXV genome, and the y-axis shows their period size (green bars) and the length (red bars) of the tandem repeat region. This in turn highlights the shortcomings of the assembly of such a repeat-rich sequencing dataset based on short-read sequencing technology. The sequence obtained in

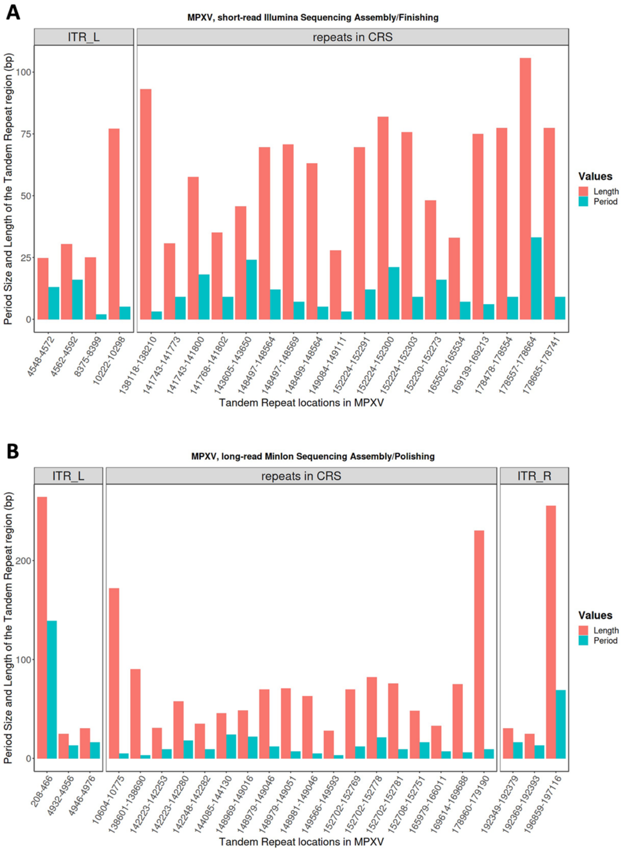

Figure 4. Tandem repeat locations in monkeypox virus genome for short-reads Illumina (**A**) and long-reads MinION (**B**) sequenced genomes using Tander Repeat Finder tool.

the present study, using long-read MinIon technology, analyzed using TRF with the same parameters, showed that the repeats could also be identified in the right-most ITR sequence Fig. 4 shows the equivalent Tandem Repeats found by TRF for the Illumina sequence (Genbank MN702446). The long reads spanning both ITRs could in fact more easily discriminate both ITR's and allow to finish the genome.

However, despite the presence of these differences in the MPXV-M sequence compared with the MPXV-I sequence, phylogenetic analyses showed that the MPXV-M sequence is positioned in the CAR1 lineage of group II strains belonging to the Central African clade. The position of the MPXV-M sequence, along with the Bao (A4 and A5) and Bangassou (B2) sequences of 2016 and 2017, respectively, is identical to the phylogenetic analyses previously performed in a previous study using all CAR MPXV sequences obtained with Illumina sequencing[19] (Fig. 5).

Evaluation of polishing tools on MinION-sequenced genome.. A major issue in nanopore sequencing is the basecalling in homopolymer-rich regions. Basecallers often tend to collapse homopolymers into shorter stretches if the homopolymer length exceeds the number of bases simultaneously influencing the measured current in the nanopore-device, resulting in a higher deletion rate in nanopore reads. As a consequence, the homopolymer containing stretches need to be checked with a dedicated polisher to increase accuracy in downstream steps of the genome finishing process. Therefore, basecalled reads are often assembled into a consensus sequence after which they are mapped back to the assembly to improve the consensus by so-called polishers. Although Canu in fact by default performs a read-correction step, dedicated polishing tools are used in combination with Canu, such as Medaka, Racon or Homopolish, alone or in combination. In order to remove the remaining systematic errors, the efficiency of the polishing operation was evaluated using the sequence for this genome, obtained by Illumina-sequencing, as a reference (MN702446). The different combinations of polishing tools show differences in the number of indels corrected. Indeed, the different combinations of Canu for assembly with polishing tools, such as Canu/Homopolish, but also Canu/Medaka/Homopolish and Canu/Racon/Medaka/Homopolish, give a better correction of indels compared to the other combinations tested. Indeed, the number of indels compared to the use of Canu alone is reduced. It is reduced from 433 to a number varying from 156 to 159 (Table 2). The other combinations tested, such as Canu/Medaka, Canu/Racon or Canu/Racon/Medaka, improve the error correction compared to Canu alone, but less efficiently. Indeed, there are still between 236 and 375 indels depending on the combinations used (Table 2). Even if the assembly performed with Canu is not satisfactory compared to the combinations with the other tools in terms of indels (433 compared to 156 to 375), no mismatch was observed compared to the reference sequence. In contrast, the

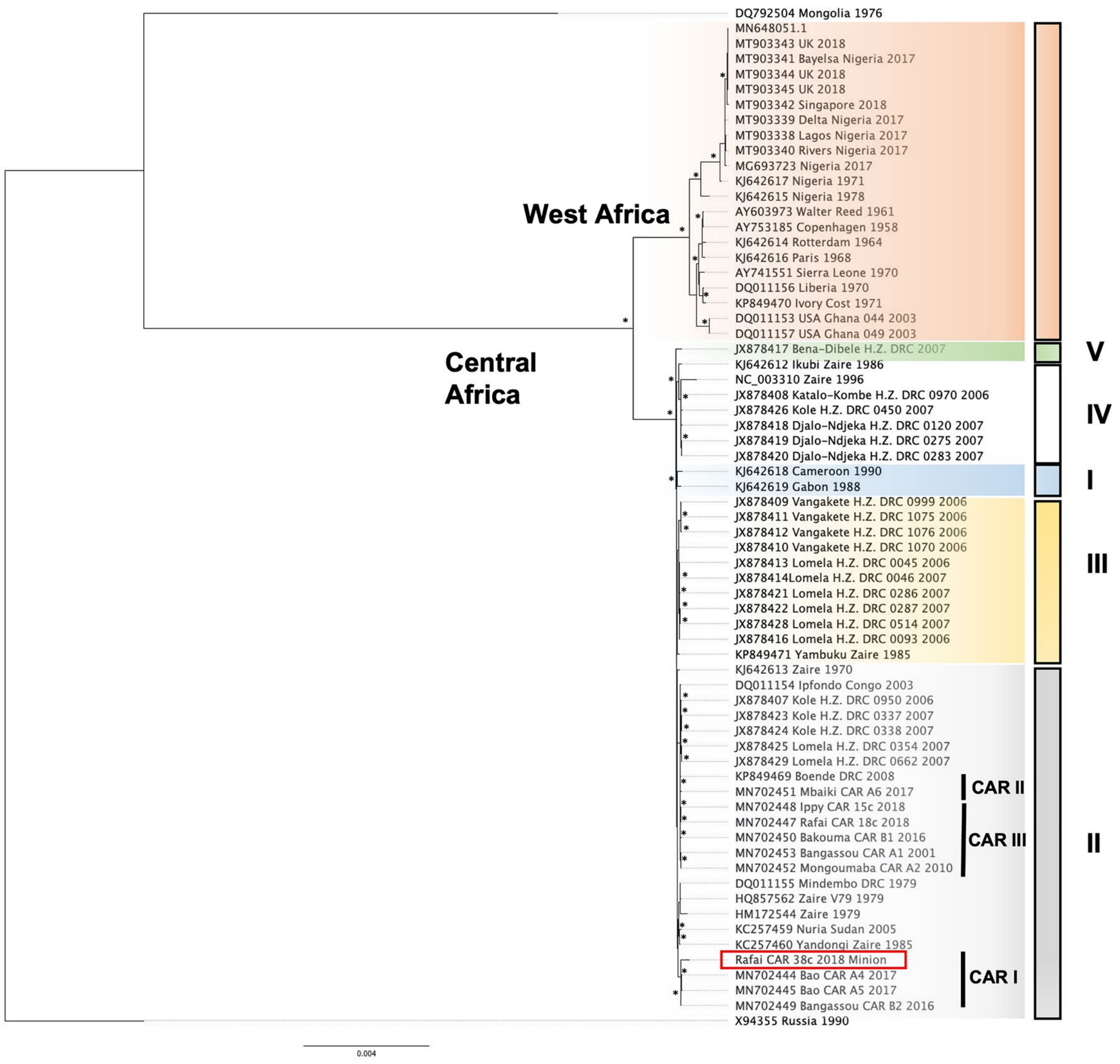
DQ792504 Mongolia 1976
MN648051.1
MT903343 UK 2018
MT903341 Bayelsa Nigeria 2017
MT903344 UK 2018
MT903345 UK 2018
MT903342 Singapore 2018
MT903339 Delta Nigeria 2017
MT903338 Lagos Nigeria 2017
MT903340 Rivers Nigeria 2017
MG693723 Nigeria 2017
KJ642617 Nigeria 1971
KJ642615 Nigeria 1978
AY603973 Walter Reed 1961
AY753185 Copenhagen 1958
KJ642614 Rotterdam 1964
KJ642616 Paris 1968
AY741551 Sierra Leone 1970
DQ011156 Liberia 1970
KP849470 Ivory Cost 1971
DQ011153 USA Ghana 044 2003
DQ011157 USA Ghana 049 2003
West Africa
Central Africa
JX878417 Bena-Dibele H.Z. DRC 2007
KJ642612 Ikubi Zaire 1986
NC_003310 Zaire 1996
JX878408 Katalo-Kombe H.Z. DRC 0970 2006
JX878426 Kole H.Z. DRC 0450 2007
JX878418 Djalo-Ndjeka H.Z. DRC 0120 2007
JX878419 Djalo-Ndjeka H.Z. DRC 0275 2007
JX878420 Djalo-Ndjeka H.Z. DRC 0283 2007
KJ642618 Cameroon 1990
KJ642619 Gabon 1988
JX878409 Vangakete H.Z. DRC 0999 2006
JX878411 Vangakete H.Z. DRC 1075 2006
JX878412 Vangakete H.Z. DRC 1076 2006
JX878410 Vangakete H.Z. DRC 1070 2006
JX878413 Lomela H.Z. DRC 0045 2006
JX878414Lomela H.Z. DRC 0046 2007
JX878421 Lomela H.Z. DRC 0286 2007
JX878422 Lomela H.Z. DRC 0287 2007
JX878428 Lomela H.Z. DRC 0514 2007
JX878416 Lomela H.Z. DRC 0093 2006
KP849471 Yambuku Zaire 1985
KJ642613 Zaire 1970
DQ011154 Ipfondo Congo 2003
JX878407 Kole H.Z. DRC 0950 2006
JX878423 Kole H.Z. DRC 0337 2007
JX878424 Kole H.Z. DRC 0338 2007
JX878425 Lomela H.Z. DRC 0354 2007
JX878429 Lomela H.Z. DRC 0662 2007
KP849469 Boende DRC 2008
MN702451 Mbaiki CAR A6 2017
MN702448 Ippy CAR 15c 2018
MN702447 Rafai CAR 18c 2018
MN702450 Bakouma CAR B1 2016
MN702453 Bangassou CAR A1 2001
MN702452 Mongoumaba CAR A2 2010
DQ011155 Mindembo DRC 1979
HQ857562 Zaire V79 1979
HM172544 Zaire 1979
KC257459 Nuria Sudan 2005
KC257460 Yandongi Zaire 1985
Rafai CAR 38c 2018 Minion
MN702444 Bao CAR A4 2017
MN702445 Bao CAR A5 2017
MN702449 Bangassou CAR B2 2016
X94355 Russia 1990
CAR II
CAR III
CAR I
V
IV
I
III
II
0.004

Figure 5. Phylogeny of monkeypox viruses (MPXV) based on complete genomes. The West African clade as well as the 5 groups of the Central African clade were highlighted by different colors: West African clade (orange), Central African group I (blue), II (grey), III (yellow), IV (white) and V (green). The final phylogenetic was generated using FigTree version 1.4.4.

Assembly/polishing methods	Q-score	Number of mismatches	Number of InDels
Canu	26.16	0	433
Canu/Homopolish	30.45	2	159
Canu/Medaka	26.78	0	375
Canu/Medaka/Homopolish	30.51	0	159
Canu/Racon	28.71	1	236
Canu/Racon/Medaka	28.39	1	254
Canu/Racon/Medaka/Homopolish	30.5	1	156

Table 2. Assembly and Polishing results using Illumina-sequenced genome (MN702446).

other combinations, with the exception of Canu/Medaka and Canu/Medaka/Homopolish, introduced 1 or 2 mismatches during the polishing process (Table 2). These mismatches are relatively rare, although they occur in regions of relatively constant sequencing depth (around 70×). The 2 mismatches resulting from polishing with Homopolish, which are 2 A–T transversions, occur in the first half of the genome (about 20–26%, in genome coordinates), while the singular mismatches occurring using Racon, which are A-C transversions, occur in the second identical half of the genome (about 83%, in genome coordinates), and are thus apparently the same SNP

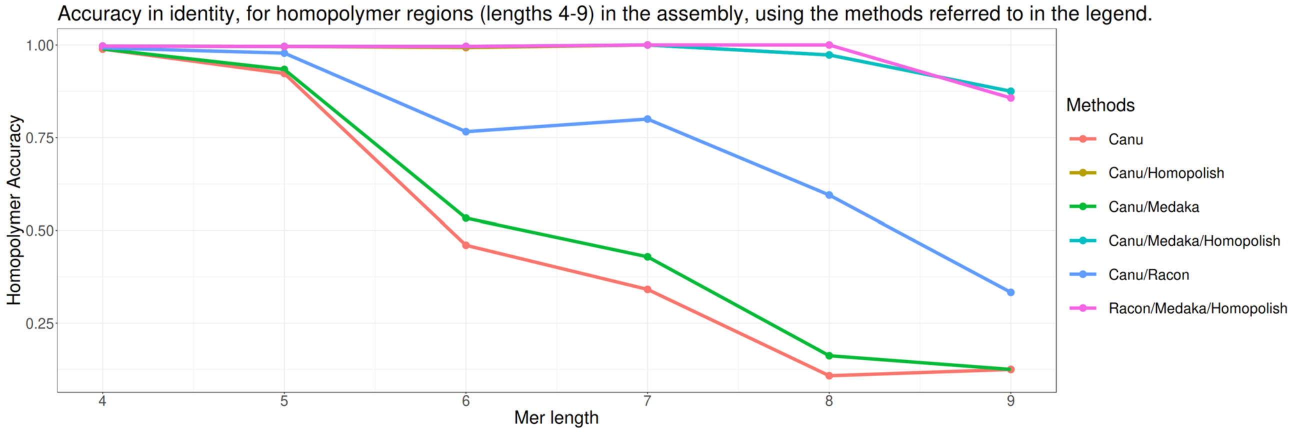

Figure 6. Accuracy of correction of the different polishing tools in combination with Canu for homopolymer sizes between 4 and 9 nt.

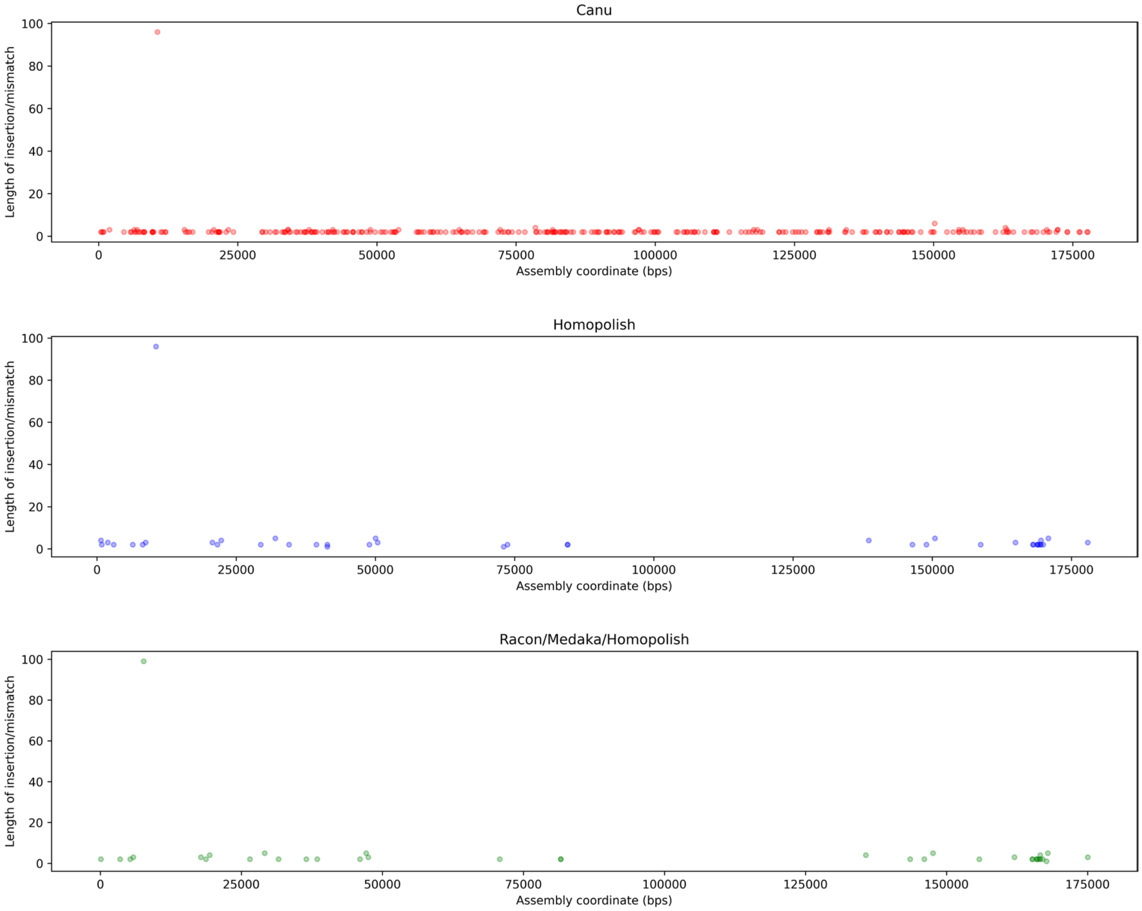

Figure 7. Scatterplot showing effect of polishing using Homopolish (all polymer occurrences).

events. The combinations tested also have variable efficiency at homopolymeric regions of size 9. Indeed, the Canu/Homopolish combination provides the best accuracy (87.5%) in the polishing procedure in these regions (Fig. 6). Moreover, it is significantly better than the Canu/Medaka combination or Canu/Racon alone (Fig. 6). In contrast, the combination of Medaka with Homopolish did not provide any improvement in error correction in these homopolymeric regions. Furthermore, the genomic coordinates of the homopolymeric sequences appear to be distributed over the entire length of the final assembly after polishing, and are therefore not restricted to the beginning and end of the assembly as illustrated with Figs. 7 and 8, showing all polymeric sequences and aberrations in the length of the deleted polymeric sequences with the precise genomic location of the remaining polymeric sequences. In conclusion, polishing of the raw and corrected reads improved the final assembly, most essentially using the Homopolish tool, notable through the Q-values that are maxing out to 30 Phred value,

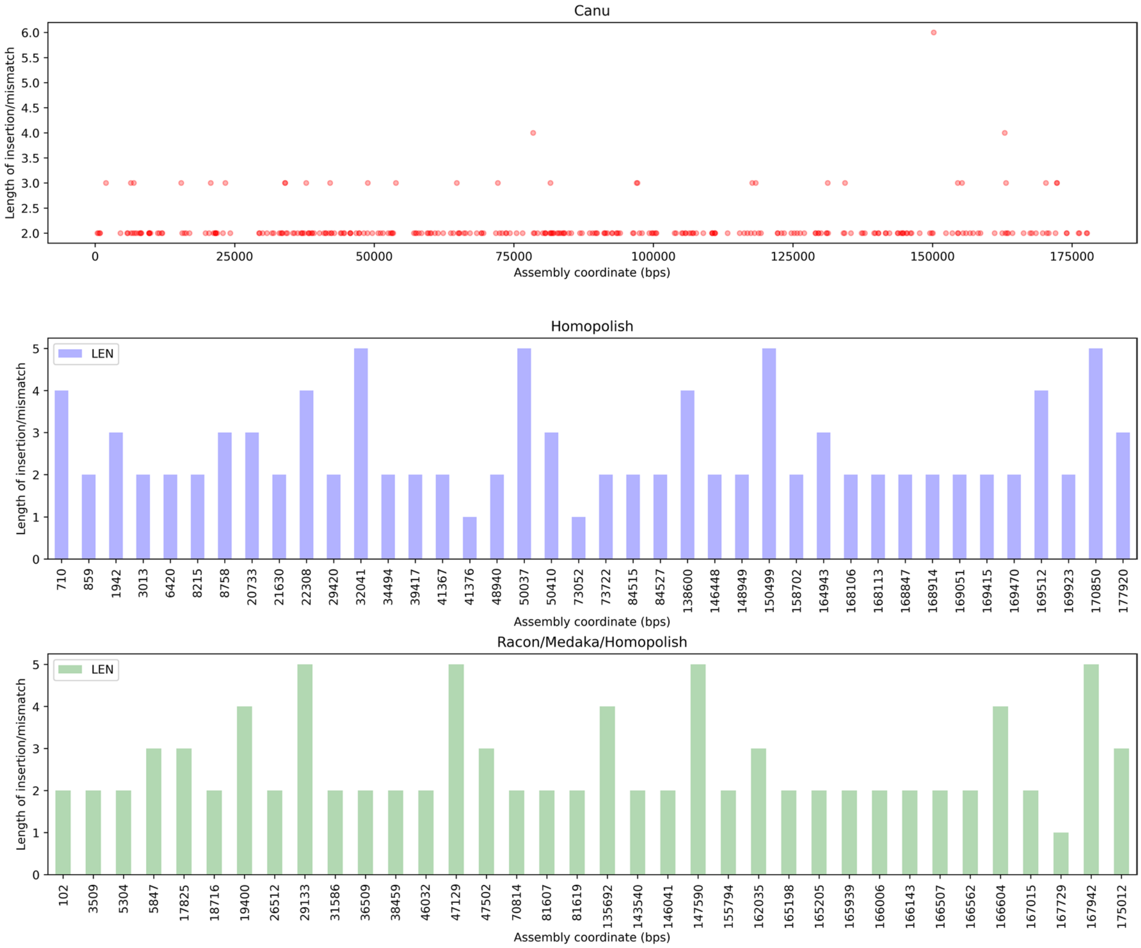

Figure 8. Scatterplot/Barplots showing Effect of polishing using Homopolish (occurrences of polymer of length < 10 bp).

while increasing only to a small extent the percent identity of the final consensus sequence. The larger number of reads available through the ONT sequencing approach, altogether with the relatively small size of the target genome (200kbp), may contribute to this effect.

Discussion

In this study, the portable real-time ONT sequencer (MinION) was used for the first time to target and sequence the full MPXV genome from DNA isolated from a human clinical sample in the CAR. Genomic data acquisition and analysis, whether for SGS or TGS, depend on a number of computer-intensive steps. After acquisition, basecalling is the initial step and is ideally performed on a dedicated computer system, given the large amount of data and metadata generated. ONT offers real-time basecalling using cloud services. However, even today, field laboratories in resource-constrained or remote locations often lack stable internet connections, making downstream analyses problematic and severely hindering sample identification in the field. Thus, for field studies, it is necessary to perform private, offline analyses of MinION data. The necessary tools include software for diversity analysis and for classification to rapidly identify species. The most common tools for MinION TGS sequences are carried out using sequence classification tools such as Centrifuge, Kraken2 and Kaiju; reference mapping tools such as MiniMap2 and BWA; and a set of variant analysis tools. Kraken2 was selected among all available DNA-to-DNA classifiers (Kraken2, KrakenUniq, k-SLAM, MegaBLAST, metaOthello, CLARK, CLARK-S, GOTTCHA, TaxMaps, Prophyle, PathSeq, Centrifuge and Karp), because it was available on a dedicated Galaxy platform with an appropriate database (MiniKraken)[35,36,51–60]. In addition, Kraken2 has good performance footprint measures and is very fast on a large number of samples. From the available DNA–protein classifiers (DIAMOND, Kaiju and MMseqs2), Kaiju was chosen because it generally has a much faster classification speed and lower memory footprint requirements than the DIAMOND and MMseqs2 classifiers, without compromising performance[34,61,62]. Most state-of-the-art tools, such as Kraken2 and Kaiju are designed for short reads. They use pseudo alignments, i.e., the exact or approximate matches from reads to a reference as signals to perform classification. Kraken2 utilizes spaced seeds in the storage and querying of minimizers to improve classification accuracy. To run efficiently, Kraken2 requires enough free memory to hold the database (primarily the hash table) in RAM (default database size, 29 GB). The custom extended RefSeq database we used (containing the RefSeq reference libraries for viral, bacterial, fungal, archaea, protozoan and plasmid genomes/proteins, the human GRCh38 human genome/proteins, as well as the NCBI non-redundant nucleotide database containing sequences from large environmental sequencing projects) required a server that had at least 60 GB of RAM, which is currently fairly common. Kraken2 was used in multi-threading mode, and typically required (at the time of writing) approximately 50 min to process a typical sequencing run. For Kaiju, the reference index was built using an equivalent NCBI protein database due to its specifically designed read classification approach, and required approximately 40 GB of RAM. Run times were slightly increased compared with Kraken2, on a comparable number of CPUs. However, given that the two selected classifiers (Kraken2 and Kaiju) operate on different principles, it was almost impossible to compare the accuracy of the classification of each method without using another reference method. Kraken uses k-mers of fixed but variable length identified in the reads. Then, it matches them to the k-mers as defined in its index constructed from reference genomes. Kaiju finds the maximum number of exact matches between the reads and the indexed database. These differences mean that Kaiju should be able to classify more reads, but with less accuracy than Kraken2 [63]. Likewise, Kaiju classified 36,102 reads as bacteria, whereas Kraken2 classified only 1423 as bacteria. However, the number of reads classified by both methods as belonging to the *Poxviridae* family was similar (2168 versus 2195, respectively, for Kraken2 and Kaiju). Although a few reads were specifically assigned to this family by either method, Kraken2 did not identify any other viral reads, unlike Kaiju, which identified 127. Despite these small differences in the results of viral read classification, both classifiers gave similar results in more than 98% of the cases. Although BLASTN alone would be more efficient compared with either Kraken2 and Kaiju, its use in field conditions is not possible [34]. Finally, the combination of results obtained with these two classifiers compensated for the lack of precision and the lack of sensitivity by making it possible to identify all the viral reads corresponding to MPXV and to use them for further analysis.

Although this TGS type of technology is known to have a higher error rate compared with SGS on the Illumina platform, few errors were detected in the genome sequence of our MPXV strain. The difference in genome size obtained using the MinION and Illumina sequencing approaches was significant given that the assembly based on the short reads obtained with Illumina was suboptimal in terminal repeat regions. Therefore, the generation of longer reads (with the MinION) provides an advantage in closing the gap on repeat regions such as terminal repeats that are more difficult to assemble with a short-read sequencing approach. In addition, TGS was able to 'correct' a region that had not been correctly sequenced and assembled from the data generated by Illumina sequencing. A study on an old African swine fever virus (ASFV) strain from 2007 resequenced using TGS also improved the length of the inverted terminal repeat sequences compared with the genome obtained previously with SGS (Illumina)[64]. In addition, a major issue in nanopore sequencing is the basecalling in homopolymer-rich regions. Basecaller can accurately call homopolymers up to six bases in length, but without significant improvement for longer homopolymers. Current basecallers often include homopolymer correction settings, although the effect appears to be small, also in our dataset. Therefore, in order to resolve mismatched stretches related to homopolymer sequences, polishers such as Nanopolish and HomoPolish propose to map the reads back to the assembly in order to improve the assembly-consensus, which after polishing is in general very high, but depending on the genome size, it can be a computationally expensive process. As in this study, the 71 detected errors were homopolymers[64]. Polishing the Canu assemblies using NanoPolish (v 0.13.2) is recommended to improve the accuracy of the sequence. Unfortunately, here, polishing did not improve either the final assembly or the accuracy of the final sequence, due to the mapping-based read ordering described above, which was primarily intended to facilitate and guide the assembly process. In contrast, HomoPolish (https://github.com/ythuang0522/homopolish)[65], a tool relying on a machine-learning model, trained to correct systematic errors that occur in Nanopore sequencing, corrected a number of mismatches and provided a polished version of the assembly compared with the original reference sequence (NC_003310). However, using the reference sequence (obtained with Illumina) did not correct the errors observed in the homopolymers. In addition, using an alternative assembly method based on SPAdes (v. 3.10) to check the robustness of the assembly approach did not produce better assembly results (data not shown).

Despite the differences observed in the homopolymer regions when sequencing the MPXV and ASFV genomes compared with SGS, TGS identifies and characterizes these viral genomes much faster than SGS, and allows to more clearly discriminate ITR sequences, which in turn allows to obtain a longer finished assembly. With SGS, the raw data can only be analyzed after sequencing has been completed, but the MinION data can be analyzed in real time. For example, the ASFV virus was identified as early as 6 min after the start of sequencing and the whole genome within hours[66]. Many examples of the use of TGS have shown that real-time data analysis is the main advantage of this technology over SGS technologies. Rapid identification is a major advantage, especially for viruses that spread very quickly and have serious life-threatening consequences. Although TGS is not the most suitable tool for rapid differential diagnosis of MPXV to distinguish it from Varicella-Zoster virus, its main advantage resides in the very fast identification of the origin of the strain. Our phylogenetic analysis showed that the MPXV-M sequence was sufficiently precise to position it correctly in the CAR1 lineage[19]. Molecular data obtained with TGS makes it possible to identify the relationships between detected cases more rapidly, especially those identified in the same city within a very short timeframe. For instance, in the city of Rafai, several cases had been reported in the space of a few weeks between March and April 2018[67], but only phylogenetic analyses were able to resolve the specific relationship between them[19]. Similarly, early differential diagnosis of MPXV, distinguishing it from Varicella-Zoster virus, and as close as possible to the index case can help limit its spread to the immediate environment. Secondary cases frequently occur either among family members or among staff at the healthcare facility where the cases are treated. Rapid identification of monkeypox within a few hours can facilitate the isolation of the patient, where possible, and the increased use of personal protective equipment to prevent the spread of the disease. Even though the use of MinION technology is relatively simple and compatible with field use, the handling of biological samples (blood, scabs or pustular lesions, etc.) where a potentially highly contagious agent is present may expose the investigator, without adequate protective equipment, to infection until the sample under investigation is fully inactivated.

In conclusion, this study confirms the usefulness of MinION technology for sequencing the genome of an MPXV virus in the context of an outbreak. Here, we show that the data obtained from directly sequencing DNA extracted from a lesion is sufficient to obtain the complete genome of the virus. The quality of the sequence obtained is suitable to provide information on the origin of the virus with sufficient accuracy despite minor errors related to the acquisition of reads observed in the homopolymeric regions.

Data availability

The corresponding MinION raw data are available under the BioProject ID (PRJNA762014).

Received: 8 January 2021; Accepted: 17 June 2022
Published online: 24 June 2022

References

1. McCollum, A. M. & Damon, I. K. Human monkeypox. *Clin. Infect. Dis.* **58**, 260–267. https://doi.org/10.1093/cid/cit703 (2014).
2. Jezek, Z., Szczeniowski, M., Paluku, K. M. & Mutombo, M. Human monkeypox: Clinical features of 282 patients. *J. Infect. Dis.* **156**, 293–298. https://doi.org/10.1093/infdis/156.2.293 (1987).
3. Pv, M., Andersen, E., Petersen, K. & Birch-Andersen, A. A pox-like disease in cynomolgus monkeys. *Acta Pathol. Microbiol. Scand.* **46**(2), 156–176 (1959).
4. Khodakevich, L., Jezek, Z. & Kinzanzka, K. Isolation of monkeypox virus from wild squirrel infected in nature. *Lancet (London, England)* **1**, 98–99 (1986).
5. Radonic, A. *et al.* Fatal monkeypox in wild-living sooty mangabey, Cote d'Ivoire, 2012. *Emerg. Infect. Dis.* **20**, 1009–1011. https://doi.org/10.3201/eid2006.13-1329 (2014).
6. Reynolds, M. G., Doty, J. B., McCollum, A. M., Olson, V. A. & Nakazawa, Y. Monkeypox re-emergence in Africa: A call to expand the concept and practice of One Health. *Expert Rev. Anti Infect. Ther.* **17**, 129–139. https://doi.org/10.1080/14787210.2019.1567330 (2019).
7. Hutin, Y. J. *et al.* Outbreak of human monkeypox, Democratic Republic of Congo, 1996 to 1997. *Emerg. Infect. Dis.* **7**, 434–438. https://doi.org/10.3201/eid0703.010311 (2001).
8. Khodakevich, L., Jezek, Z. & Messinger, D. Monkeypox virus: Ecology and public health significance. *Bull. World Health Organ.* **66**, 747–752 (1988).
9. Meyer, H. *et al.* Outbreaks of disease suspected of being due to human monkeypox virus infection in the Democratic Republic of Congo in 2001. *J. Clin. Microbiol.* **40**, 2919–2921. https://doi.org/10.1128/jcm.40.8.2919-2921.2002 (2002).
10. Berthet, N. *et al.* Maculopapular lesions in the Central African Republic. *Lancet (London, England)* **378**, 1354. https://doi.org/10.1016/s0140-6736(11)61142-2 (2011).
11. Nakoune, E. *et al.* A nosocomial outbreak of human monkeypox in the Central African Republic. *Open Forum Infect. Dis.* **4**, 168. https://doi.org/10.1093/ofid/ofx168 (2017).
12. Rimoin, A. W. *et al.* Major increase in human monkeypox incidence 30 years after smallpox vaccination campaigns cease in the Democratic Republic of Congo. *Proc. Natl. Acad. Sci. USA* **107**, 16262–16267. https://doi.org/10.1073/pnas.1005769107 (2010).
13. Sklenovska, N. & Van Ranst, M. Emergence of monkeypox as the most important orthopoxvirus infection in humans. *Front. Public Health* **6**, 241. https://doi.org/10.3389/fpubh.2018.00241 (2018).
14. Eteng, W. E. *et al.* Notes from the field: Responding to an outbreak of monkeypox using the one health approach—Nigeria, 2017–2018. *MMWR Morb. Mortal. Wkly Rep.* **67**, 1040–1041. https://doi.org/10.15585/mmwr.mm6737a5 (2018).
15. Faye, O. *et al.* Genomic characterisation of human monkeypox virus in Nigeria. *Lancet. Infect. Dis* **18**, 246. https://doi.org/10.1016/s1473-3099(18)30043-4 (2018).
16. Erez, N. *et al.* Diagnosis of imported monkeypox, Israel, 2018. *Emerg. Infect. Dis.* **25**, 980–983. https://doi.org/10.3201/eid2505.190076 (2019).
17. Sadeuh-Mba, S. A. *et al.* Monkeypox virus phylogenetic similarities between a human case detected in Cameroon in 2018 and the 2017–2018 outbreak in Nigeria. *Infect. Genet. Evolut. J. Mol. Epidemiol. Evolut. Genet. Infect. Dis.* **69**, 8–11. https://doi.org/10.1016/j.meegid.2019.01.006 (2019).
18. Vaughan, A. *et al.* Two cases of monkeypox imported to the United Kingdom, September 2018. *Euro Surveill.* https://doi.org/10.2807/1560-7917.es.2018.23.38.1800509 (2018).
19. Berthet, N. *et al.* Genomic history of human monkey pox infections in the Central African Republic between 2001 and 2018. *Sci. Rep.* **11**, 13085. https://doi.org/10.1038/s41598-021-92315-8 (2021).
20. Gardy, J., Loman, N. J. & Rambaut, A. Real-time digital pathogen surveillance—The time is now. *Genome Biol.* **16**, 155. https://doi.org/10.1186/s13059-015-0726-x (2015).
21. Pop, M. & Salzberg, S. L. Bioinformatics challenges of new sequencing technology. *Trends Genet.* **24**, 142–149. https://doi.org/10.1016/j.tig.2007.12.006 (2008).
22. Keller, M. W. *et al.* Direct RNA sequencing of the coding complete influenza A virus genome. *Sci. Rep.* **8**, 14408. https://doi.org/10.1038/s41598-018-32615-8 (2018).
23. Castro-Wallace, S. L. *et al.* Nanopore DNA sequencing and genome assembly on the international space station. *Sci. Rep.* **7**, 18022. https://doi.org/10.1038/s41598-017-18364-0 (2017).
24. Goordial, J. *et al.* In situ field sequencing and life detection in remote (79 degrees 26'N) Canadian high Arctic permafrost ice wedge microbial communities. *Front. Microbiol.* **8**, 2594. https://doi.org/10.3389/fmicb.2017.02594 (2017).
25. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232. https://doi.org/10.1038/nature16996 (2016).
26. Mbala-Kingebeni, P. *et al.* Rapid confirmation of the Zaire Ebola virus in the outbreak of the Equateur Province in the Democratic Republic of Congo: Implications for public health interventions. *Clin. Infect. Dis.* **68**, 330–333. https://doi.org/10.1093/cid/ciy527 (2019).
27. Kafetzopoulou, L. E. *et al.* Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. *Science (New York, N.Y.)* **363**, 74–77. https://doi.org/10.1126/science.aau9343 (2019).
28. Jia, L. *et al.* Nanopore sequencing of African swine fever virus. *Sci. China Life Sci.* **63**, 160–164. https://doi.org/10.1007/s11427-019-9828-1 (2020).
29. Cohen-Gihon, I. *et al.* Identification and whole-genome sequencing of a monkeypox virus strain isolated in Israel. *Microbiol. Resour. Announc.* https://doi.org/10.1128/MRA.01524-19 (2020).
30. Mauldin, M. R. *et al.* Exportation of monkeypox virus from the African continent. *J. Infect. Dis.* https://doi.org/10.1093/infdis/jiaa559 (2020).
31. Koren, S. *et al.* Canu: Scalable and accurate long-read assembly via adaptive, javax.xml.bind.JAXBElement@1d401af-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

32. Meyer, H., Ropp, S. L. & Esposito, J. J. Gene for A-type inclusion body protein is useful for a polymerase chain reaction assay to differentiate orthopoxviruses. *J. Virol. Methods* **64**, 217–221 (1997).
33. Panning, M., Asper, M., Kramme, S., Schmitz, H. & Drosten, C. Rapid detection and differentiation of human pathogenic orthopox viruses by a fluorescence resonance energy transfer real-time PCR assay. *Clin. Chem.* **50**, 702–708. https://doi.org/10.1373/clinchem.2003.026781 (2004).
34. Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **7**, 11257–11257 (2016).
35. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257–257 (2019).
36. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46–R46 (2014).
37. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *BMC Bioinform.* **12**, 385–385 (2011).
38. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)* **34**, 3094–3100 (2018).
39. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. arXiv 1207.3907. *Proc. R. Soc. B* **286**, 8–8 (2012).
40. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–1760 (2009).
42. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078–2079 (2009).
43. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780. https://doi.org/10.1093/molbev/mst010 (2013).
44. Nakamura, T., Yamada, K. D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics (Oxford, England)* **34**, 2490–2492. https://doi.org/10.1093/bioinformatics/bty121 (2018).
45. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534. https://doi.org/10.1093/molbev/msaa015 (2020).
46. Likos, A. M. *et al.* A tale of two clades: monkeypox viruses. *J. Gen. Virol.* **86**, 2661–2672. https://doi.org/10.1099/vir.0.81215-0 (2005).
47. Tulman, E. R. *et al.* Genome of horsepox virus. *J. Virol.* **80**, 9244–9258. https://doi.org/10.1128/JVI.00945-06 (2006).
48. Hendrickson, R. C., Wang, C., Hatcher, E. L. & Lefkowitz, E. J. Orthopoxvirus genome evolution: the role of gene loss. *Viruses* **2**, 1933–1967. https://doi.org/10.3390/v2091933 (2010).
49. Shchelkunov, S. N. *et al.* Human monkeypox and smallpox viruses: Genomic comparison. *FEBS Lett.* **509**, 66–70. https://doi.org/10.1016/s0014-5793(01)03144-1 (2001).
50. Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580. https://doi.org/10.1093/nar/27.2.573 (1999).
51. Ainsworth, D., Sternberg, M. J. E., Raczy, C. & Butcher, S. A. k-SLAM: Accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. *Nucleic Acids Res.* **45**, 1649–1656 (2017).
52. Breitwieser, F. P., Baker, D. N. & Salzberg, S. L. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol.* **19**, 198–198 (2018).
53. Břinda, K., Salikhov, K., Pignotti, S. & Kucherov, G. ProPhyle 0.3.1.0. *Zenodo* (2017).
54. Freitas, T. A. K., Li, P.-E., Scholz, M. B. & Chain, P. S. G. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Res.* **43**, e69–e69 (2015).
55. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
56. Kostic, A. D. *et al.* PathSeq: Software to identify or discover microbes by deep sequencing of human tissue. *Nat. Biotechnol.* **29**, 393–396 (2011).
57. Liu, X. *et al.* A novel data structure to support ultra-fast taxonomic classification of metagenomic sequences with k-mer signatures. *Bioinformatics (Oxford, England)* **34**, 171–178 (2018).
58. Ounit, R. & Lonardi, S. Higher classification sensitivity of short metagenomic reads with CLARK-S. *Bioinformatics (Oxford, England)* **32**, 3823–3825 (2016).
59. Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: Fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* **16**, 236–236 (2015).
60. Reppell, M. & Novembre, J. Using pseudoalignment and base quality to accurately quantify microbial community composition. *PLoS Comput. Biol.* **14**, e1006096–e1006096 (2018).
61. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
62. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
63. Corvelo, A., Clarke, W. E., Robine, N. & Zody, M. C. taxMaps: Comprehensive and highly accurate taxonomic classification of short-read data in reasonable time. *Genome Res.* **28**, 751–758 (2018).
64. Forth, J. H. *et al.* A deep-sequencing workflow for the fast and efficient generation of high-quality African swine fever virus whole-genome sequences. *Viruses* https://doi.org/10.3390/v11090846 (2019).
65. Huang, Y. T., Liu, P. Y. & Shih, P. W. Homopolish: A method for the removal of systematic errors in nanopore sequencing by homologous polishing. *Genome Biol.* **22**, 95. https://doi.org/10.1186/s13059-021-02282-6 (2021).
66. O'Donnell, V. K. *et al.* Rapid sequence-based characterization of African swine fever virus by use of the Oxford nanopore MinION sequence sensing device and a companion analysis software tool. *J. Clin. Microbiol.* https://doi.org/10.1128/jcm.01104-19 (2019).
67. WHO. *Weekly Bulletin on Outbreaks and Other Emergencies, Week 26 2018* (2018). http://apps.who.int/iris/bitstream/handle/10665/272981/OEW26-2329062018.pdf. Accessed 09 July 2018 (2018).

Acknowledgements

We acknowledge Dr. Carolyn Engel-Gautier for editing the English in the manuscript. This study was supported by the World Organisation for Animal Health (OIE) through the European Union (capacity building and surveillance for Ebola virus disease; EBO-SURSY, project reference: FOOD/2016/379-660). This project was also supported by *Agence National de la Recherche* (Grant AFRIPOX) as well as by the Chinese Academy of Sciences, a Shanghai Municipal Science and Technology Major Project (Grant No. 2019SHZDZX02) and External Cooperation Program of Chinese Academy of Sciences (Grant No. 153211KYSB20160001).

Author contributions

N.B. and E.N. conceived and designed the study. A.K., E.G., B.S. and R.S.O. performed all the molecular biology assays, library preparation and sequencing. M.V. performed the bioinformatics analysis whereas S.D.D. and

A.A.N.M. performed phylogenetic analysis. All authors analyzed the data. M.V. and N.B. wrote the manuscript. V.C., A.G. and E.N. participated in the critical reading of the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-15073-1.

Correspondence and requests for materials should be addressed to N.B.

Reprints and permissions information is available at www.nature.com/reprints.

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