GigaScience, 8, 2019, 1–8

doi: 10.1093/gigascience/giy169
Advance Access Publication Date: 24 January 2019
Technical Note

TECHNICAL NOTE

NanoPipe—a web server for nanopore MinION sequencing data analysis

Victoria Shabardina[1], Tabea Kischka[1], Felix Manske[1], Norbert Grundmann[1], Martin C. Frith[2,3,4], Yutaka Suzuki[5] and Wojciech Makalowski[1,*]

[1]Institue of Bioinformatics, University of Muenster, Niels-Stensen-Strasse 14, Muenster, 48149, Germany, [2]Artificial Intelligence Research Center, AIST, 2-3-26, Aomi, Koto-ku, Tokyo, 135-0064, Japan, [3]Department of Computational Biology and Medical Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8562, Japan, [4]AIST-Waseda University Computational Bio Big Data Open Innovation Laboratory, 3-4-1 Ookubo, Shinjuku-ku, Tokyo, 169-8555, Japan and [5]Laboratory of Systems Genomics, Department of Computational Biology and Medical Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8562, Japan

[*]**Corresponding author.** Wojciech Makalowski, Institue of Bioinformatics, University of Muenster, Niels-Stensen-Strasse 14, Muenster, 48149, Germany,E-mail: wojmak@uni-muenster.de http://orcid.org/0000-0003-2303-9541

Abstract

Background: The fast-moving progress of the third-generation long-read sequencing technologies will soon bring the biological and medical sciences to a new era of research. Altogether, the technique and experimental procedures are becoming more straightforward and available to biologists from diverse fields, even without any profound experience in DNA sequencing. Thus, the introduction of the MinION device by Oxford Nanopore Technologies promises to “bring sequencing technology to the masses” and also allows quick and operative analysis in field studies. However, the convenience of this sequencing technology dramatically contrasts with the available analysis tools, which may significantly reduce enthusiasm of a “regular” user. To really bring the sequencing technology to every biologist, we need a set of user-friendly tools that can perform a powerful analysis in an automatic manner. **Findings:** NanoPipe was developed in consideration of the specifics of the MinION sequencing technologies, providing accordingly adjusted alignment parameters. The range of the target species/sequences for the alignment is not limited, and the descriptive usage page of NanoPipe helps a user to succeed with NanoPipe analysis. The results contain alignment statistics, consensus sequence, polymorphisms data, and visualization of the alignment. Several test cases are used to demonstrate the efficiency of the tool. **Conclusions:** Freely available NanoPipe software allows effortless and reliable analysis of MinION sequencing data for experienced bioinformaticians, as well for wet-lab biologists with minimum bioinformatics knowledge. Moreover, for the latter group, we describe the basic algorithm necessary for MinION sequencing analysis from the first to last step.

Keywords: sequencing technologies; long-reads sequencing; bioinformatics software; MinION; Oxford Nanopore

Received: 10 October 2018; **Revised:** 10 December 2018; **Accepted:** 23 December 2018

Background

Recent years have witnessed a DNA sequencing boom due to the constantly improving technologies and, consequently, the accessibility of sequencing to a large spectrum of customers including scientists and medical practitioners. Researchers in many fields, from metagenomics to plant physiology to medicine, have been implementing sequencing experiments into their research. Oxford Nanopore Technologies (ONT) essentially accelerated this process by introducing the MinION sequencer, a portable device with minimum requirements for technical skills and bioinformatics knowledge. Thus, DNA sequencing experiments became feasible, even in field studies, in small laboratories, and soon for medical applications in clinics [1].

The NCBI PubMed database includes 261 scientific articles containing the phrase “Oxford Nanopore” published between 2009 and 2018 (by September 1, 2018), the majority of which were published in the last three years (see Supplementary Fig. S1). This is pointing both at the increased popularity of the ONT sequencers and at the considerable improvement of the technology and sequencing quality in the last three years. For example, the R10 version of the MinION flow cell was recently announced and it promises to improve sequencing quality, also for homopolymer stretches. High throughput and long reads allow diverse applications of MinION: virology [2, 3], plant pathology and agriculture [4, 5], tuberculosis studies [6], metagenomics and diet [7], and veterinary research [8]; as a portable platform, field biodiversity studies [9], detection of Ebola virus in patients on the spot [10, 11], and sequencing in space [12]. Not to forget fundamental applications for long-read sequencing studies, such as *de novo* genome assembly, improvement of existing genome assemblies, and discovery of structural variants and long repeats [13–16]. ONT sequencing is favorable for microbiology research as small-sized bacterial genomes can be covered in just one MinION read [17, 18], thus providing high resolution in genome architecture.

ONT supplies its users with the necessary software to perform base calling, i.e., converting of MinION’s electrical signals into a sequence of nucleotides: the on-run MinKNOW and offline Albacore. Both applications utilize the complex, recurrent neural network algorithm, which is recently very popular in computer science. It allows the software to learn from existing data and improve its performance. It is worth noting that the base-calling process is central to improving the accuracy of ONT sequencing technology, and its algorithm is being constantly improved and updated. The output is a collection of FAST5 and/or FASTQ files containing the base-called sequences. These are the files that are used for any sequence analysis in bioinformatics, thus, the base-caller can be called a “gate” from MinION into data interpretation. Nevertheless, the range of ONT-provided analysis tools is limited and concerns only specific applications, excluding general processing, which is left to the user. For example, the EPI2ME software suite, based on the Metrichor platform, includes applications for barcode analysis, metagenomics, antimicrobial resistance analyses, and some technical tests [19]. Several research groups have been recently focusing on the development of MinION-specific bioinformatics tools [20–22], although most of these require considerable bioinformatics knowledge. These conditions impede benefits that MinION-based DNA sequencing could bring to medical practitioners and researchers with less information technology (IT) experience. To fill this gap, we developed NanoPipe, a web-driven automatic pipeline that can quickly and effortlessly process data produced by MinION, as well provide necessary files for further bioinformatics analysis if required.

Methods

NanoPipe can be conceptually divided into four stages: (1) data uploading in FASTA or FASTQ formats, (2) alignment of MinION reads against the target sequences, (3) alignment analysis, and (4) results display (see Fig. 1). It was developed with no-IT-experienced users in mind, hence, it provides a web-driven interface with the usage page describing the main features of the tool. The start-of-analysis page is simple and intuitive. A user must, first, choose a target genome from a NanoPipe’s list (see further) or upload their own target sequence. The next step is the essential part of the pipeline: mapping of sequencing reads to the target using version 946 of the LAST sequence aligner (LAST, RRID:SCR_006119). LAST accounts for the MinION-specific sequencing errors, thus it generates highly reliable results. It can determine the rates of insertion, deletion, and each kind of substitution in a type of data (e.g., MinION reads of AT-rich *Plasmodium* DNA) [23]. It then uses these rates to determine the most probable alignments. LAST also finds the most probable division of each read into one or more parts together with the most probable alignment of each part with the last-split function [24, 25], i.e., if LAST finds a better scoring alternative alignment of the read where it is being split into parts and mapped to different regions of the target, such alignment is submitted to the results. This is a principled way to handle complex DNA rearrangements, gene fusions in RNA, chimeric host/viral sequences, etc. The tasks of detecting polymorphisms and distinguishing viral serotypes can be performed more precisely for MinION data when based on a LAST alignment, because the tool estimates the probability that each base is correctly aligned. The probability is low if there is an ambiguity, i.e., the base could align to more than one place. The default NanoPipe parameters for the LAST alignment are efficient for most cases but can be easily adjusted by the user. We use last-train to find the optimal alignment parameters for MinION sequencing [23]. For advanced users, it is recommended, although not necessary, to be acquainted with the settings of LAST and last-train [26].

After the alignment is completed, NanoPipe evaluates the nucleotide variation for each position and, based on this analysis, generates a consensus sequence and a list of possible single nucleotide polymorphisms (SNPs). The minimum nucleotide count per position should be at least 10 (i.e., at least 10 reads), otherwise a gap will be assigned at the position in the consensus sequence. The consensus sequence is calculated based on the majority rule, i.e., the nucleotide with the higher count at a particular position is assigned to the consensus. If the counts for any two nucleotides differ from each other by not more than by 20%, both nucleotides are included in the consensus with the use of the IUPAC nomenclature [27]. Statistical evaluation of nucleotide variation is presented in a separate table and suggests a polymorphism candidate if an alternate nucleotide has coverage of at least 20% of all reads.

To distinguish between artifacts and true polymorphisms, we have set three additional filters. First, a SNP candidate should have read coverage of at least 30% from the maximum coverage in the respective contig, otherwise it is not listed in the polymorphisms table and not counted as a SNP. Second, the probability of a SNP at the position is calculated. For that, the relative nucleotide frequency is multiplied by a custom weight factor; transversions are weighted by the factor of 1 and transitions by

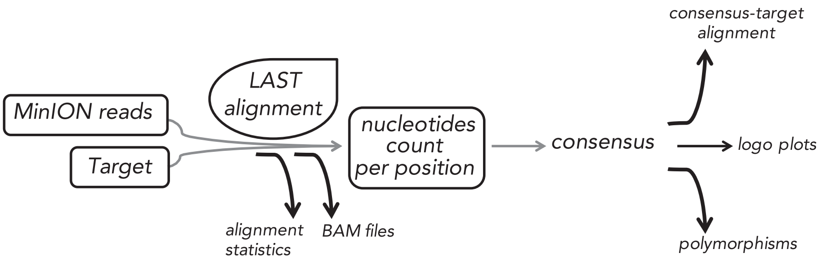


Figure 1: A schematic representation of the NanoPipe workflow.

the factor of 2. This is based on the assumption that transitions are two times more likely than transversions. This weighted probability for each SNP candidate is rescaled for convenience so that the maximum value is 1. The probability is displayed for each nucleotide in the polymorphisms table; a SNP candidate is most likely to be the true SNP if its joint probability is 1. Third, the analysis is refined by an assessment of the alignment quality around a potential SNP by estimating a p-error. The p-error is calculated based on the formula used in the LAST methods to calculate the probability of the alignment for a single nucleotide [28]. In NanoPipe the p-error is estimated over a region of a maximum of 10 nucleotides before and after the SNP position (excluding the SNP itself) and for all the read alignments at the region (Eq. 1). It is based on the LAST reliability score assigned to each base pair of the alignment. In addition, polymorphism candidates for human and *Plasmodium* are linked to the public SNP databases.

$$p - error \ = \frac{\sum_{1}^{i} 10^{-\left(\frac{ASCII_{value}-33}{10}\right)}}{i} \qquad (1)$$

Equation (1) is used in NanoPipe for the p-error calculations, where i is the total number of nucleotides around the SNP for all mapped reads, i.e., i = N(nucleotides around SNP)∗N(reads mapped to the evaluated region). ASCII values are extracted from the LAST alignment and indicate the reliability of each base's alignment; see [28] for more information about LAST quality symbols.

The results are supplemented with a number of useful pages, such as alignment of consensus sequences against target sequences, read length distribution and individual read's alignment length distribution, and nucleotide plots showing distribution of nucleotides from all reads at each position of the consensus sequence. The latter is the interactive visual representation of the results and enables the user to monitor any nucleotide variations by eye. NanoPipe also provides the necessary files for browsing the alignment in the IGV genome browser [29], where each individual read alignment is displayed.

Findings

The experience of our working group in conducting MinION sequencing and analysis workshops [30] for medical doctors and wet-lab researchers revealed the general gaps in bioinformatics knowledge among these groups. While the experimental part of the workshop was easily performed by the participants, even if mastered for the first time, the processing of the sequenced data proved to be the most enduring and difficult part. Therefore, keeping in mind the weakest spots of non-bioinformatician researchers, we describe here some features of NanoPipe usage in detail.

To start a new analysis, a query (sequencing reads) and a target (a genome or a region of interest) should be provided. The query (1D or 1D^2) can be uploaded via the NanoPipe's web interface from the user's computer in one FASTA or FASTQ file, also in archived format. As the ONT-provided base callers output sequenced reads in multiple files, it is important to know how to merge them into one file, on an operating system of the user (Windows, Mac OS, Linux). As an alternative, NanoPipe can also handle multiple FASTA or FASTQ files when they are archived (zipped), e.g., with "zip" command on Linux or WinZip tool on Windows. The maximum size of the query file should not exceed 3 GB. The target can be chosen from a drop-down menu or uploaded by the user in FASTA format. The list of NanoPipe precompiled targets includes reference genomes for human (hg38), RefSeq accession GCF_0 00001405.38; *Escherichia coli*, RefSeq accession GCF_0 00005845.2; *Caenorhabditis elegans*, RefSeq accession GCF_0 00002985.6; *Drosophila melanogaster*, RefSeq accession GCF_0 00001215.4; *Mus musculus*, RefSeq accession GCF_0 00001635.26; *Arabidopsis thaliana*, RefSeq accession GCF_0 00001735.4; *Plasmodium falciparum* strain 3D7, downloaded from plasmodb.org, version = 2013–03-01; a representative genome for *Camponotus floridanus*, RefSeq accession GCF_0 032 27725.1; and Dengue virus genome variants for serotyping (NC_0 01477.1, NC_0 01474.2, NC_0 01475.2, and NC_0 02640.1 for variant 1, variant 2, variant 3, and variant 4, respectively). Otherwise, DNA sequences of whole genomes, transcriptomes, or genes for different organisms can be accessed via particular species databases or in the corresponding databases at NCBI [31]. The NanoPipe-prepared targets include the precompiled, best-fitting alignment parameters and substitution matrices (based on last-train calculations). The substitution matrix is an important part of any alignment and contains information about the mismatch cost between any pair of nucleotides. Any uploaded targets will be used with the NanoPipe default parameters that are suitable for most cases. To avoid "noise," i.e., mapping of too short reads, the user can set the read length limit, e.g., to 200 nucleotides (depends on the experiment type and purpose). A unique name can be assigned to a job and used later within one month to retrieve the results from the NanoPipe server. After one month the data will be deleted from the server.

Depending on the query and target size, a NanoPipe analysis can last from several minutes to several hours, the server's memory used for the calculations is limited to 16 GB. The user will receive the notification via email (if the email address was provided) when his/her job is completed. The summary of the completed analysis depicts the LAST parameters that were used and the mapping statistics, i.e., how many reads were mapped altogether and the reads distribution per chromosomes/scaffolds of the target. This table can be sorted in increasing/decreasing order. Mapping distribution statistics show how long the reads in the query were and allow estimating whether the sequencing resulted in the expected read lengths. Thus, for whole-genome sequencing by MinION, a researcher can expect read lengths of more than 1 million nucleotides; targeted sequencing results in read lengths corresponding to the target length; transcriptome sequencing or RNA sequencing (RNA-seq) experiments should provide reads with length typical of particular species' transcript lengths (around one-to-two thousand nucleotides for most organisms). Alignment distribution statistics inform about the quality of the alignment, i.e., whether the whole read or a part of it could be mapped to the target.

Alignment results are visualized in NanoPipe in several ways: ordinal line-by-line pairwise alignment, a BAM file, and a graphical representation via nucleotide plots. Binary alignment (BAM) and indexed BAM files can be easily downloaded from the results page and further used for an interactive genome browsers, e.g., IGV. The target FASTA files are required by IGV and can be downloaded together with BAM files. IGV is free software and can be accessed from [32]. Nucleotide plots represent the colored mapping scheme at each position of the sequence, each nucleotide marked with its specific color. This enables easy monitoring for gaps and possible nucleotide substitutions. Navigating along the plot is enabled via right and left shift, as well by entering a nucleotide's coordinate in the search field. For long target regions, nucleotide plots provide a zoom-out preview at the bottom of the page.

Each chromosome/region of the target is supported with an individual consensus that can be seen and downloaded in FASTA format. Positions that cannot be defined (not enough information in the input data) are designated as "N"; gaps are designated as "-"; if a position is occupied by any of the four nucleotides, it is designated as "X"; other ambiguous positions are designated using IUPAC nomenclature.

The polymorphisms table lists SNP candidates and provides joint probabilities for each candidate (maximum = 1), as well raw counts for each nucleotide at the target position. If the data are available, the corresponding SNP IDs (identifiers) will be retrieved from the existing databases (currently available only for human [33] and *Plasmodium* [34]). The alignment quality around a SNP candidate is reported as p-error; the higher the p-error (maximum 1), the lower the alignment quality. Low alignment quality might indicate a region of sequencing errors around an individual SNP and, thus, signify the lower reliability of the candidate detection. However, a cluster of closely located SNPs within a distance of less than 10 nucleotides would have a similar effect. Therefore, the p-error is an additional parameter that might be taken into consideration by the user. Detailed analysis of the nucleotide plot or the alignment in the IGV viewer around questionable candidates may help in making a decision, including consideration of biological relevance.

To view the result pages for nucleotide plots, consensus, polymorphisms, and alignment, the user needs to choose a particular chromosome/region. This approach accelerates the data display and prevents the web browser from overloading.

Study Cases

Plasmodium polymorphism detection

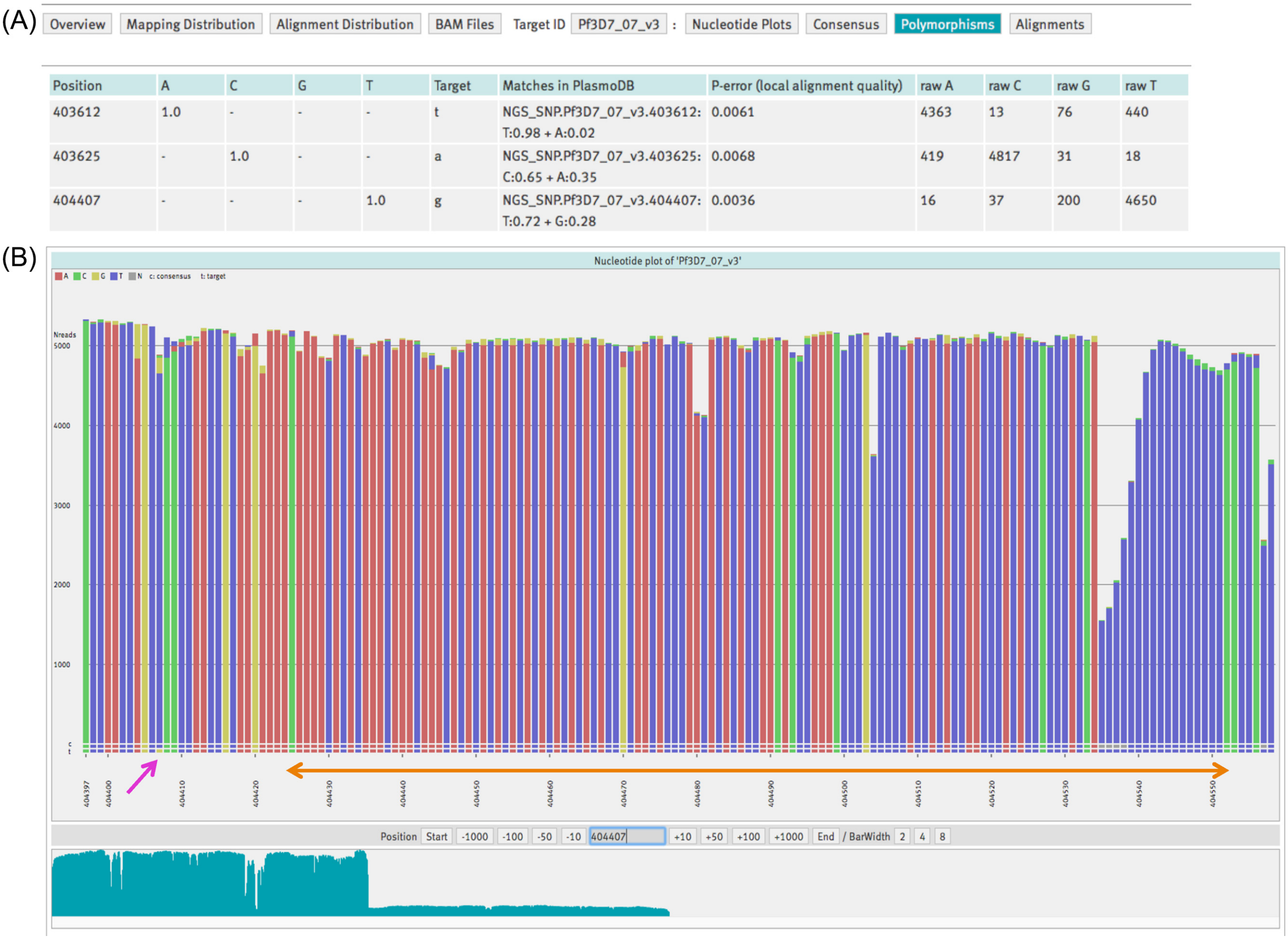
The targeted regions were first amplified using the standard PCR protocol [35]. The resulting amplicons were sequenced with the MinION using the ONT sequencing kit SDK-MAP005 or SQK-MAP006 for the library preparation and the flow cell version FLO-MAP003. The sequences were aligned against the *Plasmodium* genome (*P. falciparum* reference genome based on 3D7 strain). Exhaustive discussion of the ONT utility for *Plasmodium* SNP calling and the library preparation methods are presented elsewhere [35]. NanoPipe succeeded in mapping 99.9% of all query reads and detecting the expected mutations (see Fig. 2). The specific characteristic of the *Plasmodium* genome, multiple AT/T repeats, can be easily observed in the nucleotide plot. The full analysis of a 21 MB (FASTA format) query against the *P. falciparum* genome with NanoPipe took less than 4 minutes. This example demonstrates that the analysis of just 10,157 MinION reads with a high AT content on NanoPipe results in reliable data. Detailed screen shots of the NanoPipe results for this study case can be found in the Supplementary Material (Supplementary Figs. S2–S10).

Targeted sequencing of EGFR transcript from human lung adenocarcinoma cell line H1975

The region of the human EGFR cDNA corresponding to exons 17–22 was amplified using the primers CTAAGATCCCGTCCATCGCC (forward) and ACATATGGGTGGCTGAGGGA (reverse). The library preparation was performed following the manufacturer's recommendation using the SQK-LSK108 kit from ONT, and then sequencing was done with the MinION using flow cell version FLO-MIN107 R9. The raw 1D reads (900 MB, FASTA format) were uploaded to NanoPipe using the human reference genome as a target; the analysis took 135 minutes. A total of 79.9% of all reads were mapped to the EGFR gene on chromosome 7. Most of the remaining reads (13.5%) were mapped to chromosome 11, to regions corresponding to the cortactin gene (NM_138565) that is overexpressed in different cancers [36], and the gene encoding subunit 2 of the splicing factor 3b protein complex (NM_006842). This gene might be differentially expressed in tumor tissues [37], nevertheless, there is no definite research about possible roles of this gene in cancers. It may be the case that the gene is overexpressed in cell line H1975; therefore, its transcript was sequenced as an abundant contaminant, similar to the cortactin transcript. The four mutations expected within this region in this cell line [38] were detected, see Fig. 3. This example demonstrates the suitability of NanoPipe for cancer sequencing analysis.

RNA-seq analysis of the ant species *Camponotus maculatus*

Monarch Total RNA Miniprep kit (NEB) was used for the poly-A RNA extraction according to the manufacturer's recommendations. The library was prepared using the SQK-PCS108 kit from ONT and then sequenced with MinION using FLO-MIN107 R9 flow cell version following the manufacturer's recommendations. The raw 1D reads in FASTA format were uploaded to NanoPipe; the reference genome of *Camponotus floridanus* was used as the target genome, as neither genome nor transcriptome of *C. maculatus* are available. The analysis of the 1.5 GB query ran for 3 hours. Out of 1,814,750 raw reads, 1,773,747 (97.7%) were mapped to the target, spanning 431 scaffolds out of 657, including 150 scaffolds with coverage of more than 1,000 reads per scaffold. This result is consistent with the number of reads sequenced and the fact that the target was the genome of a different species. Thus, NanoPipe can be used for studying newly sequenced species.



Position	A	C	G	T	Target	Matches in PlasmoDB	P-error (local alignment quality)	raw A	raw C	raw G	raw T
403612	1.0	-	-	-	t	NGS_SNP.Pf3D7_07_v3.403612: T:0.98 + A:0.02	0.0061	4363	13	76	440
403625	-	1.0	-	-	a	NGS_SNP.Pf3D7_07_v3.403625: C:0.65 + A:0.35	0.0068	419	4817	31	18
404407	-	-	-	1.0	g	NGS_SNP.Pf3D7_07_v3.404407: T:0.72 + G:0.28	0.0036	16	37	200	4650

Figure 2: Sample case: *Plasmodium* polymorphism detection. **(A)** The polymorphisms table displays the three SNP candidates: two of them (at positions 403 625 and 404 407) are expected mutations leading to K76T and A220S amino acid changes and, as a consequence, to altered resistance of the parasite to chloroquine, mefloquine, and quinine drugs [35]. **(B)** Nucleotide plot. The purple arrow points to the G>T substitution at position 404 407 (GCC>TCC codon change). The orange arrow highlights an AT-rich region, the typical feature of the *P. falciparum* genome.

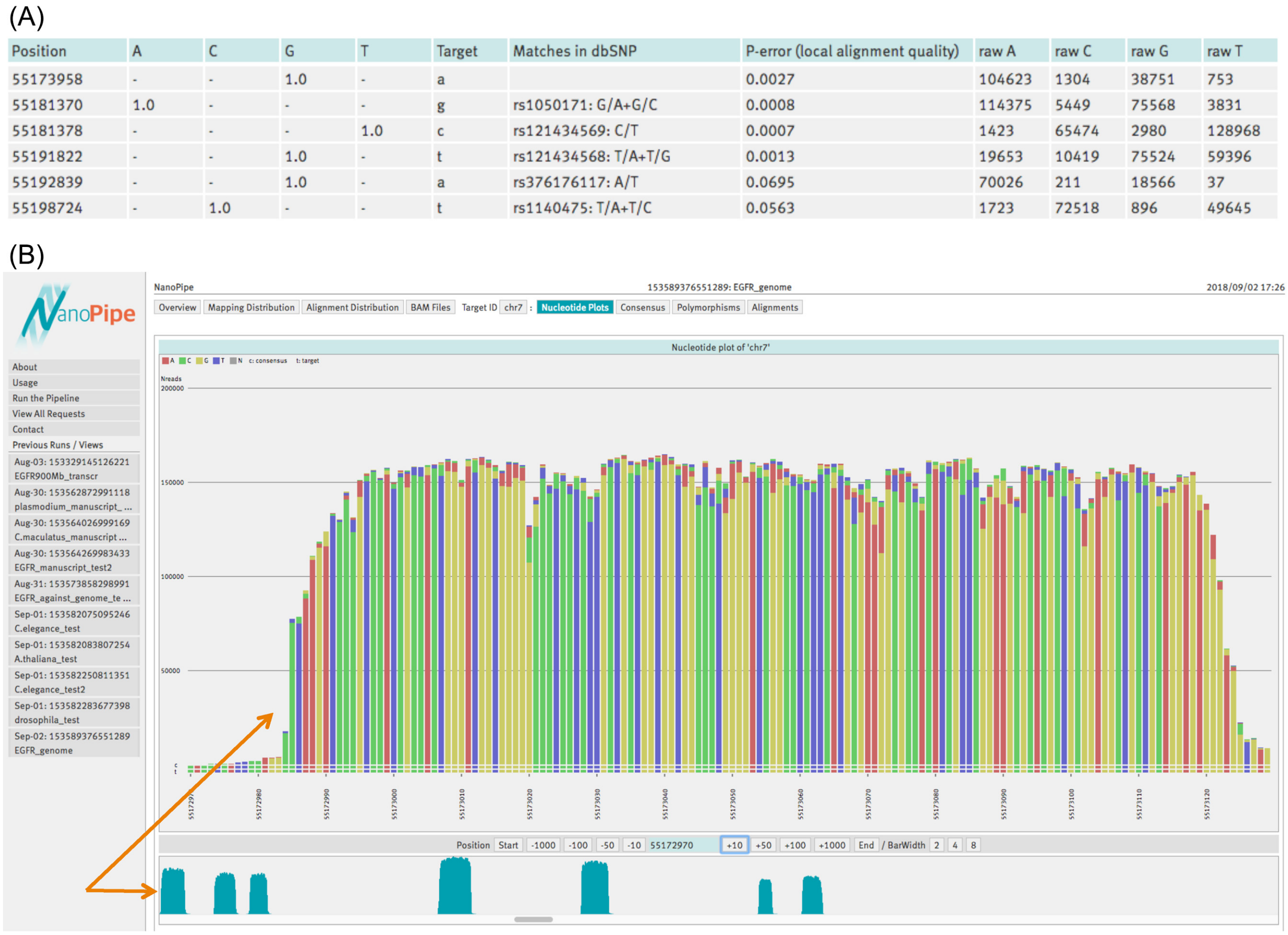
Direct RNA-seq of poxvirus isolated from the host cells of green monkey

We examined the recent MinION direct RNA-seq data of the *Vaccinia virus WR* mRNA isolated from the kidney fibroblast cells of *Chlorocebus sabeus* (for details, see [39]). Using the virus genome and the monkey genome as targets in the two separate runs, we could separate the reads coming from the two organisms. From the tested 29,846 reads, 1,314 were mapped to the virus genome (GenBank accession LT966077.1) and 14,714 reads were mapped to the *C. sabeus* genome (GenBank accession GCA_000 409795.2), which is consistent with the published results. The option of the direct target upload was also tested on the small and large input files (the virus genome size is 198 KB, and the green monkey genome is 2.82 GB). Both runs were finished successfully in 12 seconds and in 51 minutes, respectively, including the construction of the target databases. NanoPipe proved to be useful in analyzing the mixed long reads of virus and its host.

Discussion

We have developed NanoPipe, a web-driven application that enables easy analysis of MinION sequencing data that is suitable for both experienced bioinformaticians and biologists with limited IT knowledge. NanoPipe provides users with the consensus sequence of the studied DNA and a list of putative SNPs. In this work, we used four different experimental datasets: targeted sequencing of the AT-rich genome of *P. falciparum*; targeted sequencing of EGFR human cDNA; direct RNA-seq of *V. virus* and its host, green monkey *C. sabeus*; and RNA-seq of the ant *P. maculatus* that has, to date, no fully sequenced genome or transcriptome. These four test cases represent different tasks that biologists from different fields of study could be interested in: sequencing and SNP detection in repeat-rich genomes with low sequencing yield (study case with *P. falciparum*); mutation detection in human cancer samples (study case with lung adenocarcinoma cell line H1975); RNA-seq of a newly sequenced organism (study case with *C. maculatus*); and of a mixed species sample (virus-host study). NanoPipe proved to be a reliable tool for all of these tasks. It detected all expected mutations for *Plasmodium* and the EGFR transcript region and succeeded in mapping more than 90% RNA-seq reads from *C. maculatus* to the reference genome. In the case of virus-host direct RNA-seq, NanoPipe successfully segregated the mixed reads between the virus genome and its host genome.

We have tested NanoPipe during two workshops at the International Summer School for MinION sequencing in Bangkok (2017) and Manado (2018) with participants having no bioin-



(A)

Position	A	C	G	T	Target	Matches in dbSNP	P-error (local alignment quality)	raw A	raw C	raw G	raw T
55173958	-	-	1.0	-	a		0.0027	104623	1304	38751	753
55181370	1.0	-	-	-	g	rs1050171: G/A+G/C	0.0008	114375	5449	75568	3831
55181378	-	-	-	1.0	c	rs121434569: C/T	0.0007	1423	65474	2980	128968
55191822	-	-	1.0	-	t	rs121434568: T/A+T/G	0.0013	19653	10419	75524	59396
55192839	-	-	1.0	-	a	rs376176117: A/T	0.0695	70026	211	18566	37
55198724	-	1.0	-	-	t	rs1140475: T/A+T/C	0.0563	1723	72518	896	49645

(B)

NanoPipe
About
Usage
Run the Pipeline
View All Requests
Contact
Previous Runs / Views
Aug-03: 153329145126221 EGFR900Mb_transcr
Aug-30: 153562872991118 plasmodium_manuscript_ ...
Aug-30: 153564026999169 C.maculatus_manuscript ...
Aug-30: 153564269983433 EGFR_manuscript_test2
Aug-31: 153573858298991 EGFR_against_genome_te ...
Sep-01: 153582075095246 C.elegance_test
Sep-01: 153582083807254 A.thaliana_test
Sep-01: 153582250811351 C.elegance_test2
Sep-01: 153582283677398 drosophila_test
Sep-02: 153589376551289 EGFR_genome
NanoPipe
153589376551289: EGFR_genome
2018/09/02 17:26
Overview
Mapping Distribution
Alignment Distribution
BAM Files
Target ID
chr7
:
Nucleotide Plots
Consensus
Polymorphisms
Alignments
Nucleotide plot of 'chr7'
A C G T N c: consensus t: target
Nreads
200000
150000
100000
50000
c
t
55172970
55172980
55172990
55173000
55173010
55173020
55173030
55173040
55173050
55173060
55173070
55173080
55173090
55173100
55173110
55173120
Position
Start
-1000
-100
-50
-10
55172970
+10
+50
+100
+1000
End
/ BarWidth
2
4
8

Figure 3: Sample case: Targeted sequencing of EGFR transcript. **(A)** Polymorphism results. The two expected nucleotide substitutions are silent: G>A (CAG>CAA = Gln) at position 55 181 370 and T>C (ACT>ACC = Thr) at position 55 198 724; two other substitutions at position 55 181 378 (C>T leading to the amino acid change T745M) and at position 55 191 822 (T>G leading to the amino acid change L813R) are responsible for the sensitivity to anticancer drugs, in particular, to gefitinib and erlotinib [38]. **(B)** Nucleotide plot. The zoom-out scheme at the bottom depicts seven alignment picks; they represent seven sequenced exons of the transcript. The nucleotide plot is showing the first of these picks, pointed to by the orange arrows.

formatics experience from the medical and biological research fields. NanoPipe proved to be efficient and understandable for these users. Nevertheless, our team also noted during these practical studies that the participants lacked some basic knowledge necessary for conducting successful sequencing experiments. Therefore, we described here in detail the general scheme of working with long-read sequencing data. The flexibility of NanoPipe allows researchers to study any sequence of interest (including cancer samples in human patients), as the list of targets is not limited. With this software, we hope to make MinION sequencing even more accessible to medical researchers and biologists without sophisticated IT resources and expertise. The next step for the NanoPipe project will include more sophisticated statistics for SNP detection and evaluation. Also, we are planning to cover the microbiology field and implement an option for metagenomics analysis.

Availability of source code and requirements

Project name: NanoPipe
Project home page: http://bioinformatics.uni-muenster.de/tools/nanopipe, GitHub page: https://github.com/IOB-Muenster/nanopipe2. The NanoPipe package for local installation is available at the NanoPipe GitHub page. The explanation on installation and check-install procedures can be found in theGitHub directory.
Operating system: Unix
Programming language: Javascript, Python, Perl
License: Apache License 2.0
Other: there are no limitations for web browser type. The local version can be installed on Unix operating system
RRID: SCR_016852
The raw sequence data for the test runs and test jobs can be accessed from the homepage http://bioinformatics.uni-muenster.de/share/NanoPipe_test_data/
Test data description:
1. Data set *P. falciparum* sequencing
1.1 MinION targeted sequencing for *P. falciparum* sequencing
1.2 PfCRT and K-propeller genes were sequenced using long-read technologies within the experiments of testing MinION sequencer for SNP detection in *Plasmodium*
1.3 The sequencing was performed for the MinION sequencing workshop, 2017, in Bangkok, Thailand, at Mahidol University
1.4 Raw reads in a single FASTA file, archived using zgip command

1.5 *P. falciparum* culture, strain 3D7
1.6 Size: 21 MB, archived: 4.3 MB

2. Data set for EGFR sequencing
2.1 MinION targeted sequencing of the coding sequence for EGFR human gene, spanning exons 17–22
2.2 These data were generated to test MinION capacity for detecting SNP in human cancer DNA
2.3 The sequencing was performed for the MinION sequencing workshop, 2017, in Bangkok, Thailand, at Mahidol University
2.4 Raw reads in a single FASTA file, archived using zgip command
2.5 Human lung adenocarcinoma cell line H1975 (RRID:CVCL_1511)
2.6 Size: 900 MB, archived: 259 MB

3. Data set for *C. maculatus* sequencing
3.1 MinION RNAseq for ant *C. maculatus*
3.2 Poly-A RNA sequencing was performed for the given ant species at first time
3.3 The sequencing was performed in the Institute of Bioinformatics, Muenster, Germany; the library was prepared in the group of Juergen Gadau, University of Muenster
3.4 Raw reads in a single FASTA file, archived using zgip command
3.5 Larvae and adult individuals from the lab culture
3.6 Size: 1.5 GB, archived: 309 MB

Supporting data and snapshots of the GitHub are also archived in the *GigaScience* GigaDB repository [40]; the sequences used in the test cases can be accessed at the European Nucleotide Archive under the study ID PRJEB30433.

Additional files

Figure S1. Number of scientific publications that contain the phrase “Oxford Nanopore” in their abstracts.
Figures S2–S10. The results pages for the *P. falciparum* test case.

Abbreviations

BAM: binary alignment; IT: Information Technology; ONT: Oxford Nanopore Technologies; NCBI: National Center for Biotechnology Information; RNA-seq: RNA sequencing; SNP: single-nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.

Funding

The work has been funded by Institute of Bioinformatics, University of Muenster and University Clinic Muenster.

Author contributions

V.S. validated the software, participated in the concept design, and composed the manuscript draft. T.K. designed the software and participated in the code writing. F.M. wrote the code for polymorphism analysis and contributed to the manuscript writing. N.G. wrote the major part of the pipeline code and designed the software. M.F. participated in the improvement of the software, including LAST optimization, and contributed to the manuscript writing. Y.S. contributed to the analysis and provided the access to the MinION sequencing summer school. W.M. designed the concept and supervised the work. All the authors read and approved the manuscript.

Acknowledgements

The targeted sequencing of *P. falciparum* (strain 3D7) genes PfCRT_1, PfCRT_2, K13-propeller and EGFR human transcript was done for the MinION sequencing workshop in Bangkok, Thailand, in 2017 at Mahidol University. The RNA-seq library for *P. maculatus* was prepared by the group of Jürgen Gadau, Molecular Evolution and Sociobiology Group, University of Münster.

References

1. ONT, Flongle. https://nanoporetech.com/products/flongle. Accessed 5 September 2018.
2. Sasani TA, Cone KR, Quinlan AR, et al. Long read sequencing reveals poxvirus evolution through rapid homogenization of gene arrays. bioRxiv 2018:245373.
3. Karamitros T, Harrison I, Piorkowska R, et al. De novo assembly of human herpes virus type 1 (HHV-1) genome, mining of non-canonical structures and detection of novel drug-resistance mutations using short- and long-read next generation sequencing technologies. PLoS One 2016;**11**(6):e0157600.
4. Bronzato Badial A, Sherman D, Stone A, et al. Nanopore sequencing as a surveillance tool for plant pathogens in plant and insect tissues. Plant Dis 2018;**102**(8):1648–52.
5. Fleming MB, Patterson EL, Reeves PA, et al. Exploring the fate of mRNA in aging seeds: protection, destruction, or slow decay? J Exp Bot 2018;**69**(18):4309–21.
6. Bainomugisa A, Duarte T, Lavu E, et al. A complete high-quality MinION nanopore assembly of an extensively drug-resistant *Mycobacterium tuberculosis* Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. Microb Genomics 2018;**4** (7) :e000188.
7. Pearman W, Smith ANH, Breckell G, et al. New tools for diet analyses: nanopore sequencing of metagenomic DNA from stomach contents to quantify diet in an invasive population of rats. bioRxiv 2018:363622.
8. Theuns S, Vanmechelen B, Bernaert Q, et al. Nanopore sequencing as a revolutionary diagnostic tool for porcine viral enteric disease complexes identifies porcine kobuvirus as an important enteric virus. Sci Rep 2018;**8**(1):9830.
9. Krehenwinkel H, Pomerantz A, Henderson JB, et al. Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. bioRxiv 2018:358572. In press.
10. Quick J, Loman NJ, Duraffour S, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature 2016;**530**(7589):228–32.
11. Mbala Kingebeni P, Villabona-Arenas C-J, Vidal N, 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 2019;**68**(2):330–33.
12. Castro-Wallace SL, Chiu CY, John KK, et al. Nanopore DNA sequencing and genome assembly on the international space station. Sci Rep 2017;**7**(1):18022.
13. Solares EA, Chakraborty M, Miller DE, et al. Rapid low-cost assembly of the *Drosophila melanogaster* reference genome using low-coverage, long-read sequencing. G3 (Bethesda)

2018:g3.200162.2018.

14. Deschamps S, Zhang Y, Llaca V, et al. A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. Nat Commun 2018;**9**(1): 4844.
15. Pradhan B, Cajuso T, Katainen R, et al. Detection of subclonal L1 transductions in colorectal cancer by long-distance inverse-PCR and nanopore sequencing. Sci Rep 2017;**7**(1):14521.
16. Kuderna LFK, Lizano E, Julia E, et al. Selective single molecule sequencing and assembly of a human Y chromosome of African origin. bioRxiv 2018:342667.
17. Chacon-Vargas K, Chirino AA, Davis MM, et al. Genome sequence of *Zymomonas mobilis* subsp. *mobilis* NRRL B-1960. Genome Announc 2017;**5**(30).
18. Reis AC, Kroll K, Gomila M, et al. Complete genome sequence of *Achromobacter denitrificans* PR1. Genome Announc 2017;**5**(31):e00759–17.
19. Metrichor. METRICHOR LTD. https://metrichor.com/. Accessed 5 September 2018.
20. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 2015;**12**(8):733–5.
21. Srivathsan A, Baloğlu B, Wang W, et al. A MinION-based pipeline for fast and cost-effective DNA barcoding. Mol Ecol Resour 2018. doi: 10.1111/1755-0998.12890.
22. Leggett RM, Heavens D, Caccamo M, et al. NanoOK: multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. Bioinformatics 2015;**32**(1):btv540.
23. Hamada M, Ono Y, Asai K, et al. Training alignment parameters for arbitrary sequencers with LAST-TRAIN. Bioinformatics 2016;**33**(6):btw742.
24. Frith MC, Kawaguchi R. Split-alignment of genomes finds orthologies more accurately. Genome Biol 2015;**16**(1):106.
25. Frith MC, Khan S. A survey of localized sequence rearrangements in human DNA. Nucleic Acids Res 2018;**46**(4):1661–73.
26. Frith M. LAST: genome-scale sequence comparison. http://last.cbrc.jp/. Accessed 5 September 2018.
27. IUPAC. IUPAC codes. https://www.bioinformatics.org/sms/iupac.html. Accessed 5 September 2018.
28. Frith M. Last-Split. http://last.cbrc.jp/doc/last-split.html. Accessed 5 September 2018.
29. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013;**14**(2):178–92.
30. Suzuki Y. JSPS, Training seminar for MinION. http://www.cb.k.u-tokyo.ac.jp/suzukilab/minion/summary.html. Accessed 9 October 2018.
31. NCBI. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/. Accessed 5 September 2018.
32. IGV-page. Home | integrative genomics viewer. http://software.broadinstitute.org/software/igv/. Accessed 5 September 2018.
33. NCBI.dbSNP. dbSNP. https://www.ncbi.nlm.nih.gov/projects/SNP/. Accessed 10 October 2018.
34. PlasmoDB. PlasmoDB : The plasmodium genomics resource. http://plasmodb.org/plasmo/. Accessed 10 October 2018.
35. Runtuwene LR, Tuda JSB, Mongan AE, et al. Nanopore sequencing of drug-resistance-associated genes in malaria parasites, *Plasmodium falciparum*. Sci Rep 2018;**8**(1):8286.
36. Gattazzo C, Martini V, Frezzato F, et al. Cortactin, another player in the Lyn signaling pathway, is over-expressed and alternatively spliced in leukemic cells from patients with B-cell chronic lymphocytic leukemia. Haematologica 2014;**99**(6):1069–77.
37. Lantto TA, Laakso I, Dorman HJD, et al. Cellular stress and p53-associated apoptosis by *Juniperus communis* L. Berry extract treatment in the human SH-SY5Y neuroblastoma cells. Int J Mol Sci 2016;**17**(7):1113.
38. Suzuki A, Suzuki M, Mizushima-Sugano J, et al. Sequencing and phasing cancer mutations in lung cancers using a long-read portable sequencer. DNA Res 2017;**24**(6):585–96.
39. Tombácz D, Prazsák I, Szűcs A, et al. Dynamic transcriptome profiling dataset of vaccinia virus obtained from long-read sequencing techniques. GigaScience 2018. doi: 10.1093/gigascience/giy139.
40. Shabardina V, Kischka T, Manske F, et al. “Supporting data for “NanoPipe—a web server for nanopore MinION sequencing data analysis.” GigaScience Database. 2019, http://dx.doi.org/10.5524/100551.