Journal of Human Genetics (2020) 65:61–67
https://doi.org/10.1038/s10038-019-0659-4

REVIEW ARTICLE

Bioinformatics of nanopore sequencing

Wojciech Makałowski[1] · Victoria Shabardina[1]

Received: 7 June 2019 / Revised: 26 July 2019 / Accepted: 5 August 2019 / Published online: 26 August 2019

Abstract
Nanopore sequencing is one of the most exciting new technologies that undergo dynamic development. With its development, a growing number of analytical tools are becoming available for researchers. To help them better navigate this ever changing field, we discuss a range of software available to analyze sequences obtained using nanopore technology.

Introduction

Beginning of twenty-first century witnessed dynamic development of sequencing technology. First, so-called "Next Generation Sequencing" brought increased sequencing yield and decline of sequencing cost. However, it was with a certain price, namely the length of the reads, which is much shorter than in traditional Sanger sequencing. As a result, we are flooded with a number of genomes (as of May 30, 2019, there are over 200,000 prokaryotic and almost 9000 eukaryotic genomes deposited at the NCBI database). However, most of these genomes are in the so-called draft form. It means that chromosomes are presented in rather small pieces for which order and orientation on a chromosome is unknown. Moreover, gene annotation in these genomes is quite poor or does not exist at all.

Understandably, the emergence of the third generation of sequencing technology that enables a single molecule long reads was met with a great excitement within the community. There are currently two competing products on the market, namely single molecule, real-time sequencing by PacBio [1] and nanopore sequencing by Oxford Nanopore Technologies (ONT) [2]. The latter is especially exciting technology thanks to its portability and very low initial cost of hardware [3], which brings the sequencing "democratization" one step closer. Initially, ONT did not provide any analytical tools for the sequences obtained by nanopore sequencing. Moreover, the sequences delivered by the native base caller were in a peculiar format (FAST5) that none of the existing at that time software could handle. As a matter of fact, the very first tool developed outside of ONT was a toolkit called poretools, which main function was a format conversion from FAST5 to more familiar FASTQ or FASTA formats [4]. Nevertheless, long reads with relatively low-sequencing accuracy require different computational approaches than short highly accurate reads. Consequently, many algorithms and analytical tools have been developed to aid nanopore sequences, some of them are specific to this particular method and others are more generic that can be used to any long reads.

Here, we describe a few dozens of tools developed in recent years that are suitable for the nanopore sequences' analyzes. Although the list is by no means comprehensive, we try to cover the whole range of software that reflects diversity of the nanopore sequencing applications. For the reading convenience, we group these programs by tasks they can perform. Please note that some more generic programs are listed several times as they can handle different tasks.

Base calling

Base calling is a crucial step in any sequencing method. It is a process of transforming a raw signal obtained from a sequencer into a string of nucleotides. In the case of nanopore sequencing, it is a computational processing of electric signal collected from an ONT instrument (MinION, GridION, or PromethION). The accuracy of base calling is influenced by two factors. First, the chemistry used can affect a signal-to-noise ratio. If the ratio is low, determination of underlying DNA sequence may not be possible [5]. The second factor is how well the signal can be

✉ Wojciech Makałowski
wojmak@uni-muenster.de

[1] Institute of Bioinformatics, Faculty of Medicine, University of Münster, 48149 Münster, Germany

Table 1 Base callers developed for nanopore sequencing

Tool	Read qscore[a]	Consensus qscore[a#]	Availability
Albacore	9.2	21.9	Only to ONT customers
BasecRAWller	N/A	N/A	https://basecrawller.lbl.gov/ (seems to be down)
Chiron	7.7	21.4	https://github.com/haotianteng/Chiron
DeepNano	N/A	N/A	https://bitbucket.org/vboza/deepnano/src/master/
FastQC	A quality control tool for high throughput sequence data.		https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Flappie	9.6	22.0	https://github.com/nanoporetech/flappie
Guppy	9.7	23.0	Only to ONT customers
Metrichor	N/A	N/A	Only to ONT customers
Nanocall	N/A	N/A	https://github.com/mateidavid/nanocall
Scrappie	9.3	22.4	https://github.com/nanoporetech/scrappie

[a]based on [8]

interpreted by a software used for base calling. To discriminate between signal and noise a specific training dataset is used, which may not be optimal for interpretation of real DNA molecule if the latter has, for instance, strong nucleotide composition bias. For example, the genome of malaria-causing parasite is 80% AT rich and nanopore base calling of reads from this genome is usually far from optimal, especially within homopolymers stretches.

Since, the base calling is a very important step in obtaining a useful, for a customer, data, it is not surprising that the ONT is developing such software and is constantly working on its improvement. The original ONT base caller used Hidden Markov Models (HMM) approach but all current programs use neural networks machine learning approach, following the path paved by the DeepNano software [6]. Table 1 lists ten base callers developed specifically for nanopore sequencing. Since nanopore-sequencing technology is developing dynamically, base callers need to keep up the pace and understandably some of them became obsolete, e.g., albacore or metrichore. The base callers are usually available for a range of operating systems, including Linux, MacOS, and Windows. However, since base calling is computationally intensive, it is a good idea to run the software on a multiple-processor machine or a server. It is also beneficial to perform a base calling on GPUs (graphics processing unit) instead of CPUs (central processing unit). Their highly parallel structure makes them more efficient in specialized tasks. In fact, ONT's computing unit, MinIT, designed for MinION and computing module of PromethION are equipped with 256-core GPU (https://nanoporetech.com/products/minit). For example, Chiron [7] is about 80 times faster on a GPU when compared with a single CPU performance. Wick et al. recently compared several base callers for nanopore sequencing [8]. In short, they all perform very similarly regarding quality score of both single reads and consensus sequences with the exception of Chiron who performed a bit worst on a single-reads level (see Table 1). Interestingly, previous version of Chiron achieved better quality scores than the current version. Chiron is also much slower than other base callers compared by Wick et al. [8]. On the other end of speed distribution lies Guppy [8]. They also tested Guppy with different models and concluded that developing custom-trained models can improve base calling significantly [8]. Furthermore, the accuracy of a consensus sequence can be improved by employing Nanopolish software [9].

Mapping

Obtaining a sequence is only the beginning of the analysis. Depending on a biological question we ask or why the sequencing was done at the first place, there are several avenues that one may take. However, aligning raw reads to existing sequences is often the first task on the "to do" list. This is especially true if a sequencing project involves organisms for which genomes has been already decoded. Often, the task of aligning raw-sequencing reads to already determined sequences, for instance a genome, is called mapping raw reads to the target. Aligning is such a basic task in the molecular sequence analysis that several algorithms that deal with the problem were developed long before bioinformatics field existed [10, 11]. Although these early algorithms are very elegant and guarantee optimal solution to the problem, they are computational heavy and not very practical when one has to deal with the vast number of sequences. Consequently, heuristic algorithms have been developed to conquer speed limitation of exhaustive algorithms. Probably, the most successful and the best known is the BLAST algorithm [12]. It is worth to mention that BLAST was developed for database similarity searches and its goal is to find all the instances of similar sequences in a database. In mapping, however, the goal is different, namely to find an exact position on a genome or a

Table 2 Selected aligners suitable for long reads

Tool	Algorithm	Availability
BWA	Burrows–Wheeler Aligner's Smith–Waterman Alignment	http://bio-bwa.sourceforge.net
GraphMap	Gapped spaced seeds	https://github.com/isovic/graphmap
Kart	Divide and conquer	https://github.com/hsinnan75/Kart
LAMSA	Sparse dynamic programming (SDP)-based split alignment	https://github.com/hitbc/LAMSA
LAST	Adaptive seeds approach	http://last.cbrc.jp/
Minimap2	Hash table approach	https://github.com/lh3/minimap
NanoPipe	A pipeline that includes a consensus sequence calculation based on LAST alignment to a reference sequence	http://bioinformatics.uni-muenster.de/tools/nanopipe2/index.hbi https://github.com/IOB-Muenster/nanopipe2
NGMLR	k-mer search followed by a banded Smith–Waterman alignment algorithm	https://github.com/philres/ngmlr

transcript from which a given read comes from. In practice, BLAST and other similar tools report many alignments of a query to a set of sequences but mappers, in ideal situation, report just one alignment. In fact, if a sequencing read aligns to multiple positions in a genome, usually it will be discarded from further analysis.

Nevertheless, with the advent of the new generation of sequencing methods and increased volume of raw reads early heuristic solutions turned out to be also too slow and the new generation of aligning/mapping software has been developed (reviewed in [13]). These programs take advantage of short length but high accuracy of the sequencing reads offered by these technologies. However, with the third generation of sequencing technology we took a step back regarding sequencing accuracy. It quickly became clear that many heuristic algorithms used for the next generation sequencing are not good enough for long reads with relatively high-sequencing error rate. Interestingly, algorithms developed for the alignment of relatively distant (dissimilar) sequences proved to be very useful. For instance, LAST [14] became very popular aligner within ONT community. Some developers adapted successful next generation-sequencing aligners to the new data, see for example "-l" option in bwa software [15].

Table 2 lists several programs that have been successfully used for aligning nanopore sequences. Unfortunately, there is no standard developed, or rather there are several standards, for the mapping software output, which makes down-the-line analyses more complicated. For example, one has to be aware of numbering style that is used by the specific aligner, i.e., zero- or one-based coordinates. Although one-based coordinates seem to be more natural (first position in the alignment gets number 1 assigned), some software use the latter system, where position one in the alignment gets coordinate 0. This may lead to erroneous results down the line in an analysis pipeline. The aligned raw reads can be used for further analyses, such as single-nucleotide polymorphism (SNP) finding or calculation of a consensus sequence. Finally, the mapped sequences can be visualized in a graphical viewer, such as Integrative Genomics Viewer [16], if the mapper's output can serve as an input for the viewer, for instance in the form of BAM file.

Sequence assembly

The longest recorded DNA read to date is over 2.3 Mb (https://nanoporetech.com/about-us/news/longer-and-longer-dna-sequence-more-two-million-bases-now-achieved-nanopore; https://www.biorxiv.org/highwire/filestream/96273/field_highwire_adjunct_files/1/312256-2.gz). The longest known human transcript is one of the isoforms of titin gene (*TTN*) with more than 108 kb. One may conclude that with nanopore technology we can sequence any transcript in its entirety. However, even such long reads are not long enough to cover the whole bacterial genome or the whole eukaryotic chromosome by a single read. From the very beginning of sequencing approach, an assembly of raw reads was required in order to obtain a complete, contiguous sequences. Originally, the sequence assemblers used overlaps to merge and order raw sequences [17]. This approach is called Overlap-Layout-Consensus (OLC). However, for the NGS data volume and short length of reads, it proved to be prohibitive to use the OLC algorithms effectively, although there are some exemptions from this rule [18–20]. In early 2000s, using de Bruijn graphs to solve assembly problems has been proposed [21, 22]. Algorithms based on de Bruijn graphs turned out to be very useful for short reads assembly and most of modern assemblers developed for the next generation sequencing use this approach. Interestingly, these algorithms are not very well suited for the noisy, long reads and for these sequences the interest has shifted back to the OLC approach. One of these assemblers is Canu [23] developed based on Celera Assembler [24], which has been used for a successful assembly of the variety of genomes, such as bacteria [25], fungi [26], fruit fly [27], cotton [28], or fish [29]. Recently, de Bruijn graph approach was implemented for long reads by Pevzner

Table 3 Selected software for sequence assembly and scaffolding

Tool	Description	Availability
ABruijn	De novo assembler for long and noisy reads	https://github.com/bioreps/ABruijn
Canu	A hierarchical assembly pipeline based on Celara Assembler	https://github.com/marbl/canu
Cobbler	Gap filling with long sequences	https://github.com/bcgsc/RAILS
Flye	De novo assembler for single-molecule sequencing reads	https://github.com/fenderglass/Flye
HINGE	A long-read assembler based on an idea called hinging	https://github.com/HingeAssembler/HINGE
LINKS	Application for scaffolding genome assemblies with long reads	https://github.com/bcgsc/LINKS
MECAT	An ultra-fast mapping, error correction and de novo assembly tool for long reads	https://github.com/xiaochuanle/MECAT
Medaka	A tool to create a consensus sequence of nanopore-sequencing data using neural networks	https://nanoporetech.github.io/medaka/index.html
Miniasm	OLC-based de novo assembler	https://github.com/lh3/miniasm
NanoPipe	A pipeline that includes a consensus sequence calculation based on LAST alignment to a reference sequence	http://bioinformatics.uni-muenster.de/tools/nanopipe2/index.hbi https://github.com/IOB-Muenster/nanopipe2
Nanopolish	Software package for signal-level analysis of Oxford Nanopore-sequencing data, including consensus sequence calculation	https://github.com/jts/nanopolish
npScarf	A program that scaffolds and completes draft genomes assemblies in real time with Oxford Nanopore sequencing	https://github.com/mdcao/npScarf
PBJelly	A pipeline that aligns long sequencing reads to high-confidence draft assembles to fills the gaps	https://sourceforge.net/projects/pb-jelly/
Racon	A standalone consensus module to correct raw contigs generated by rapid assembly methods	https://github.com/isovic/racon
RAILS	Radial assembly improvement by long sequence scaffolding	https://github.com/bcgsc/RAILS
SMART denovo	It produces an assembly from all-vs-all raw read alignments without an error correction stage.	https://github.com/ruanjue/smartdenovo
SPAdes	Hybrid assembler that can handle different input data, including long and short reads, and preassembled contigs.	http://cab.spbu.ru/software/spades/
wtdbg2	Fast long reads assembler based on fuzzy–Bruijn graphs	https://github.com/ruanjue/wtdbg2

group [30] but it has not been wildly used so far. The ABruijn assembler includes a sequence polishing module and repeat analysis step, which apparently improves the structural accuracy of the final assembly [30]. Table 3 lists assemblers suitable for long but noisy reads along with other tools useful for assembly improvement, such as consensus sequence polishing.

Variant detection

Each individual of a given species is different at the genome level. For instance, each of us differs from the rest of humans by about 3 million bp or 0.1%. The reference genomes are usually represented as a consensus sequence based on several individuals. However, the subtle changes in a genome are responsible for individual traits or disease susceptibility. Genetic variants consist of SNPs, small insertions or deletions (InDels) and structural variants, such as large InDels and complex rearrangements, including translocations and inversions. In principle any mapping tool can be used for variant detection. However, some additional steps are required to evaluate mapping results and validate potential variants. Noisy nature of current long reads, including nanopore reads, makes the process quite difficult. In the case of most eukaryotic organism, potential-sequencing errors are overlaid over diploid genomes and consequently one has to deal with heterozygous loci. Moreover, in the cases such as cancer genomes, the variants might be obscured by mix of cancer and healthy cells that were used to isolate-sequencing material. Therefore, statistical methods have been developed to evaluate likelihood of the observed differences between nanopore reads and a reference sequence to be a true genomic variant. Table 4 lists some tools related to variant analysis, including variant calling and variant phasing. These tools employ different approaches for variant calling. While most of the programs try to detect variants after base calling and often by comparing a consensus sequence with a reference, Nanopolish is doing so using raw signal information [31].

Table 4 Variant calling and variant phasing software

Tool	Description	Availability
Clair	Deep neural network-based variant caller	https://github.com/HKU-BAL/Clair
HapCUT2	It is a maximum-likelihood-based tool for assembling haplotypes.	https://github.com/vibansal/HapCUT2
IDP-ASE	Haplotyping and quantification of allele-specific expression	http://augroup.org/IDP-ASE/IDP-ASE
Medaka	An experimental pipeline to call SNPs	https://nanoporetech.github.io/medaka/index.html
NanoPipe	A pipeline that includes a consensus sequence calculation based on LAST alignment to a reference sequence	http://bioinformatics.uni-muenster.de/tools/nanopipe2/index.hbi https://github.com/IOB-Muenster/nanopipe2
Nanopolish	Software package for signal-level analysis of Oxford Nanopore-sequencing data, including SNP and indel calling	https://github.com/jts/nanopolish
PBHoney	An implementation of variant-identification designed for long reads	https://sourceforge.net/projects/pb-jelly/
Sniffles	Sniffles is a structural variation (over 10 bp) caller using third generation sequencing	https://github.com/fritzsedlazeck/Sniffles
WhatsHap	It is a software for phasing genomic variants	https://whatshap.readthedocs.io/en/latest/

Miscellaneous tools

Nanopore sequencing is still developing technology and researchers keep finding new applications for it. These usually require new, specialized software. One such an interesting application is studying DNA methylation, which is involved in many biological processes, such as gene regulation and cell differentiation [32]. Nanopore technology enables studying DNA methylation directly, as ionic current signal should be different for methylated and unmethylated nucleotides. Two different software, Nanopolish [9] and SignalAlign [33], are using HMMs to identify C5-methylcytosine (5mC) with high accuracy. The newest addition to the methylation toolbox is DeepMod, which is using raw electric signals and a bidirectional recurrent neural network to detect 5mC and N6-methyldeoxyadenosine (6 mA) [34]. Interestingly, the recent ONT's base caller Flappie is able to call 5mC in CpG context for R9.4.1 on PromethION platform.

Another promising application of the nanopore sequencing lies within metagenomic studies. Short reads often cannot distinguish between closely related species or microbial strains, as rRNA often used for the bar coding is a very conservative molecule. Several reports showed that long reads might be a solution to that problem and few dedicated software were developed, including MetaG and the ONT's own EMPI2ME.

One of the advantages of nanopore sequencing is possibility of direct sequencing of RNA molecules (see ONT's white paper at https://nanoporetech.com/resource-centre/rna-sequencing-white-paper-value-full-length-transcripts-without-bias). This includes defining complexity of alternative transcripts, unbiased quantification of transcriptome and detection of methylated nucleotides. Several, specialized software appeared recently to specifically deal with noisy long reads. For instance, SQANTI [35] was developed to decipher complexity of alternative transcripts. Although developed and tested with PacBio reads it can be used for any long reads, including nanopore, since it takes FASTA files as an input. FLAIR (Full-Length Alternative Isoform analysis of RNA) enables the correction, isoform definition, and alternative splicing analysis of noisy reads [36]. ONT developed a set of tools called pinfish for long transcriptomics data analyses, which was inspired by Mandalorion pipeline [37]. LoReAn is a tool developed for eukaryotic genome annotation utilizing short- and long-read cDNA sequencing, protein evidence, and ab initio gene prediction [38].

Table 5 lists few other applications that can be useful for the nanopore sequence analyses. For instance, NanoSim-H is a software to simulate the nanopore reads [39]. These simulated reads can be used to test performance of other analytical tools developed specifically for the ONT reads. NanoDJ [40] integrates many tools together enabling tasks such as base calling, sequence quality assessment, and sequence assembly in a single environment. Tandem-genotypes is an interesting software that finds changes in length of tandem repeats, from "long" DNA reads aligned to a genome [41]. It simply aligns long reads against a reference genome using last-split [14] and then compares tandem repeats annotated in the reference genome with the aligned long read, one read at a time. Then, it summarizes the results for each tandem repeat locus annotated in the reference genome.

The final piece of software that we would like to mention is RUBRIC [42]. It implements ONT's molecule-by-molecule real-time selective sequencing (Read Until) for real-time sequencing. RUBRIC compares in real time a molecule that is actively being sequenced in a nanopore to evaluate if it should be sequenced to its completeness. If not, the current in that particular pore is reversed and consequently the molecule that is currently being sequenced is rejected from the nanopore. This approach enables decision making after only 150 nt were sequenced. The method might be especially useful for pathogen detection when pathogen/host DNA ratio is unfavorable. Unfortunately, the whole method requires high computing power and the

Table 5 Miscellaneous tools useful in the nanopore sequence analyses

Tool	Description	Availability
DeepMod	Detection of DNA base modifications by deep recurrent neural network	https://github.com/WGLab/DeepMod
EPI2ME	A set of real-time analytical tools, including species identification and reads quality control	Only to ONT customers
FLAIR	Full-length alternative isoform analysis of RNA	https://github.com/BrooksLabUCSC/flair
Flappie	Base calling of 5mC in CpG context	https://github.com/nanoporetech/flappie
IGV	Integrative genomics viewer	https://software.broadinstitute.org/software/igv/home
LoReAn	Annotation pipeline designed for eukaryotic genomes using long and short reads	https://github.com/lfaino/LoReAn
Mandalorion	Analysis Pipeline to analyze Nanopore RNAseq data	https://github.com/rvolden/Mandalorion-Episode-II
MEGAN-LR	Part of MEGAN (metagenomic tool) designed for long reads.	http://ab.inf.uni-tuebingen.de/data/software/megan6/download/
MetaG	A metagenomics pipeline suitable for long-sequencing technologies	http://bioinformatics.uni-muenster.de/tools/metag/index.hbi
NanoDJ	A Jupyter notebook integration of tools for simplified manipulation and assembly of DNA sequences	https://github.com/genomicsITER/NanoDJ
Nanopolish	Software package for signal-level analysis of Oxford Nanopore-sequencing data, including methylation analysis	https://github.com/jts/nanopolish
NanoSim-H	A simulator of Oxford Nanopore reads	https://pypi.org/project/NanoSim-H/
pinfish	A collection of tools helping to make sense of long transcriptomics data	https://github.com/nanoporetech/pinfish
Pychopper	Tool developed by ONT to identify full-length cDNA reads.	https://github.com/nanoporetech/pychopper
RUBRIC	It enables Real-time Selective Sequencing	https://github.com/harrisonedwards/RUBRIC
SignalAlign	Methylation detection using hidden Markov models	https://github.com/ArtRand/signalAlign
SQANTI	A pipeline for the characterization of isoforms obtained by full-length transcript sequencing	https://bitbucket.org/ConesaLab/sqanti/src/master/
Tandem-genotypes	A software to find changes in length of tandem repeats, from "long" DNA reads aligned to a genome	https://github.com/mcfrith/tandem-genotypes

authors used two computers in parallel, one for sequencing and the other one to run RUBRIC, in order to run the system smoothly and in real time [42]. Recently ONT has been more active developing analytical tools and sharing them with the nanopore community via the GitHub site: https://github.com/nanoporetech. We encourage all readers to check this site for ever growing list of useful tools.

Conclusions

Nanopore sequencing technology promises to democratize nucleic acid sequencing. However, as a sequence is only a raw material in gaining biological knowledge, to do so, we need analytical tools. Unfortunately, most of the software developed for interpretation of the nanopore sequences require relatively high bioinformatics skills, which most biologists lack. In the growing number of tools available for the nanopore sequence analysis, the NanoPipe [43] seems to be an exception with a simple web interface and a clear, easy to understand output files. Nevertheless, "one swallow does not make a spring" [44] and we need more software for easy data analysis and interpretation to make sequencing fully democratized.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

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

References

1. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323:133–8.
2. Kasianowicz JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. Proc Natl Acad Sci. 1996;93:13770–3.
3. Leggett RM, Clark MD. A world of opportunities with nanopore sequencing. J Exp Bot. 2017;68:5419–29.
4. Loman NJ, Quinlan AR. Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics. 2014;30:3399–401.

5. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018;19:90.
6. Boza V, Brejova B, Vinar T. DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. PLoS ONE. 2017;12:e0178751.
7. Teng HT, Cao MD, Hall MB, Duarte T, Wang S, Coin LJM. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. Gigascience. 2018;7:giy037.
8. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford nanopore sequencing. Genome Biology. 2019;20:129.
9. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. 2017;14:407.
10. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48:443–53.
11. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147:195–7.
12. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
13. Shang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, Shen B. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. Biomed Res Int. 2014;2014: 309650.
14. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21: 487–93.
15. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26:589–95.
16. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.
17. Staden R. A strategy of DNA sequencing employing computer programs. Nucleic Acids Res. 1979;6:2601–10.
18. Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. Genome Res. 2008;18:802–9.
19. Simpson JT, Durbin R. Efficient construction of an assembly string graph using the FM-index. Bioinformatics 2010;26:i367–i73.
20. Gremme G, Steinbiss S, Kurtz S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE ACM Trans Comput Biol Bioinform. 2013;10:645–56.
21. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci USA. 2001;98:9748–53.
22. Pevzner PA, Tang H, Tesler G. De novo repeat classification and fragment assembly. Genome Res. 2004;14:1786–96.
23. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27: 722–36.
24. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, et al. A whole-genome assembly of Drosophila. Science. 2000;287:2196–204.
25. Ma ZW, Hu JC. Complete genome sequence of a marine-sediment-derived bacterial strain Bacillus velezensis SH-B74, a cyclic lipopeptides producer and a biopesticide. 3 Biotech. 2019; 9:162.
26. Brejova B, Lichancova H, Brazdovic F, Hegedusova E, Jakubkova MF, Hodorova V, et al. Genome sequence of the opportunistic human pathogen Magnusiomyces capitatus. Curr Genet. 2019;65:539–60.
27. Karageorgiou C, Gamez-Visairas V, Tarrio R, Rodriguez-Trelles F. Long-read based assembly and synteny analysis of a reference Drosophila subobscura genome reveals signatures of structural evolution driven by inversions recombination-suppression effects. BMC Genomics. 2019;20:223.
28. Wang MJ, Tu LL, Yuan DJ, Zhu D, Shen C, Li JY, et al. Reference genome sequences of two cultivated allotetraploid cottons, Gossypium hirsutum and Gossypium barbadense. Nat Genet. 2019;51:224.
29. Xiao YS, Xiao ZZ, Ma DY, Liu J, Li J. Genome sequence of the barred knifejaw Oplegnathus fasciatus (Temminck & Schlegel, 1844): the first chromosome-level draft genome in the family Oplegnathidae. Gigascience. 2019;8:giz013.
30. Lin Y, Yuan J, Kolmogorov M, Shen MW, Chaisson M, Pevzner PA. Assembly of long error-prone reads using de Bruijn graphs. P Natl Acad Sci USA. 2016;113:E8396–E405.
31. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530:228–32.
32. Zeng Y, Chen T. DNA methylation reprogramming during mammalian development. Genes. 2019;10:257.
33. Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, et al. Mapping DNA methylation with high-throughput nanopore sequencing. Nat Methods. 2017;14:411.
34. Liu Q, Fang L, Yu G, Wang D, Xiao CL, Wang K. Detection of DNA base modifications by deep recurrent neural network on Oxford nanopore sequencing data. Nat Commun. 2019;10:2449.
35. Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, del Risco H, et al. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. 2018;28:396–411.
36. Tang AD, Soulette CM, Baren MJV, Hart K, Hrabeta-Robinson E, Wu CJ, et al. Full-length transcript characterization of *SF3B1* mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. bioRxiv. 2018:410183.
37. Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat Commun. 2017;8:16027.
38. Cook DE, Valle-Inclan JE, Pajoro A, Rovenich H, Thomma BPHJ, Faino L. Long-Read Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA Sequencing. Plant Physiol. 2019;179:38–54.
39. Yang C, Chu J, Warren RL, Birol I. NanoSim: nanopore sequence read simulator based on statistical characterization. Gigascience. 2017;6:gix010.
40. Rodríguez-Pérez H, Hernández-Beeftink T, Lorenzo-Salazar JM, Roda-García JL, Pérez-González CJ, Colebrook M, et al. NanoDJ: a dockerized jupyter notebook for interactive Oxford Nanopore MinION sequence manipulation and genome assembly. BMC Bioinformatics. 2019;20:234.
41. Mitsuhashi S, Frith MC, Mizuguchi T, Miyatake S, Toyota T, Adachi H, et al. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. Genome Biol. 2019; 20:58.
42. Edwards HS, Krishnakumar R, Sinha A, Bird SW, Patel KD, Bartsch MS. ReAl-time Selective Sequencing with RUBRIC: read until with basecall and reference-informed criteria. BMC Bioinformatics. 2019;20:234.
43. Shabardina V, Kischka T, Manske F, Grundmann N, Frith MC, Suzuki Y, et al. NanoPipe-a web server for nanopore MinION sequencing data analysis. Gigascience. 2019;8:giy169.
44. Aristotle. The nicomachean ethics. Oxford; New York: Oxford University Press; 2009. xliii, p. 277.