BMC Bioinformatics

SOFTWARE Open Access

NanoDJ: a Dockerized Jupyter notebook for interactive Oxford Nanopore MinION sequence manipulation and genome assembly

Héctor Rodríguez-Pérez[1†], Tamara Hernández-Beeftink[1†], José M. Lorenzo-Salazar[2], José L. Roda-García[3], Carlos J. Pérez-González[4], Marcos Colebrook[3*] and Carlos Flores[1,2,5*]

Abstract

Background: The Oxford Nanopore Technologies (ONT) MinION portable sequencer makes it possible to use cutting-edge genomic technologies in the field and the academic classroom.

Results: We present NanoDJ, a Jupyter notebook integration of tools for simplified manipulation and assembly of DNA sequences produced by ONT devices. It integrates basecalling, read trimming and quality control, simulation and plotting routines with a variety of widely used aligners and assemblers, including procedures for hybrid assembly.

Conclusions: With the use of Jupyter-facilitated access to self-explanatory contents of applications and the interactive visualization of results, as well as by its distribution into a Docker software container, NanoDJ is aimed to simplify and make more reproducible ONT DNA sequence analysis. The NanoDJ package code, documentation and installation instructions are freely available at https://github.com/genomicsITER/NanoDJ.

Keywords: Genome analysis, Nanopore sequencing, Jupyter, Docker

Background

It has never been before so easy and affordable to access and utilize genetic variation of any organism and purpose. This has been motivated by the continuous development of high-throughput DNA sequencing technologies, most commonly known as Next Generation Sequencing (NGS). A key improvement is the possibility of obtaining long single molecule sequences with the fast and cost-efficiency technology released by Oxford Nanopore Technologies (ONT) and the marketing in 2014 of the MinION, a portable, pocket-size, nanopore-based NGS platform [1]. Since then, several algorithms and software tools have flourished specifically for ONT sequence data. Despite its size, it provides multi-kilobase reads with a throughput comparable to other benchtop sequencers in the market (1–10 Gbases by 2017), therefore still necessitating of efficient and integrated bioinformatics tools to facilitate the widespread use of the technology.

While MinION has shown promise in distinct applications [2], because of the low cost, laptop operability, and the USB-powered compact design of MinION, cutting-edge NGS technology is not any more necessarily linked to the established idea of a large machine with high cost that must be located in centralized sequencing centers or in a laboratory bench. As a consequence, the utility of MinION in field experiments to move from sample-to-answers on site have been demonstrated with infectious disease studies [3, 4], off-Earth genome sequencing [5], and species identification in extreme environments [6–8], among others. Leveraging of MinION capabilities in the academic classroom is a natural extension of these field studies to facilitate

* Correspondence: mcolesan@ull.edu.es; cflores@ull.edu.es
†Héctor Rodríguez-Pérez and Tamara Hernández-Beeftink contributed equally to this work.
[3]Departamento de Ingeniería Informática y de Sistemas, Universidad de La Laguna, Santa Cruz de Tenerife, Spain
[1]Research Unit, Hospital Universitario Nuestra Señora de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain
Full list of author information is available at the end of the article

education of genomics in undergraduate and graduate students [9].

To date, there is no specific software solution aimed to facilitate ONT sequence analyses by integrating capabilities for data manipulation, sequence comparison and assembly in field experiments or for educational purposes to help facilitate learning of genomics [9]. We have developed NanoDJ, an interactive collection of Jupyter notebooks to integrate a variety of software, advanced computer code, and plain contextual explanations. In addition, NanoDJ is distributed as a Docker software container to simplify installation of dependencies and improve the reproducibility of results.

Implementation

NanoDJ is distributed as a Docker container built underneath Jupyter notebooks, which is increasingly popular in life sciences to significantly facilitate the interactive exploration of data [10], and has been recently integrated in the widely used Galaxy portal [11]. The Docker container allows NanoDJ to run in an isolated, self-contained package, that can be executed seamlessly across a wide range of computing platforms [12], having a negligible impact on the execution performance [13]. NanoDJ integrates diverse applications (Additional file 1: Table S1) organized into 12 notebooks grouped on three sections (Fig. 1; Table 1). Main results are presented as embedded objects. In addition, one of the notebooks was conceived for educational purposes by setting a particularly simple problem and the inclusion of low-level explanations. To facilitate the use of the educational notebook and bypassing the installation of Docker and NanoDJ, a lightweight version of this notebook and small sets of ONT reads can be utilized from a web-browser using Binder (https://mybinder.org) in the NanoDJ GitHub repository. In addition, as part of the CyVerse project (https://www.cyverse.org/), NanoDJ has been incorporated into VICE, a visual and interactive computing environment that facilitates training of ONT data analysis. We illustrate the versatility of NanoDJ in distinct scenarios by providing results from four case studies (Additional file 1: Text S1).

Input, basecalling, and simulations

Input data can be a list of FAST5 files from previous basecalled runs (e.g. a Metrichor output) or event-level signal data to be basecalled using the latest ONT caller. The user can also simulate reads with NanoSim and pre-computed model parameters. This possibility is important in different scenarios as to help designing an experiment, or to bypass technical difficulties in academic setups [9].

Summary, quality control and filtering

Either for a simulated or an empirical run, the user will obtain summary data and plots informing of read length distribution, GC content vs. length, and read length vs. quality score (when available). If barcodes were used in the experiment, Porechop can be used for demultiplexing, barcode trimming and to filter out reads.

Genome assembly and comparison

Depending on the application, sequence data can be aligned against reference sequences or used for genome assembly using diverse methods. Alignment is performed either against one (BWA and Rebaler) or multiple (BLAST) reference sequences, providing the generation of BAM files for downstream applications (e.g., variant identification) or information of species composition. Alternatively, the user may opt for a de novo assembly. NanoDJ allows the use of some of the best-performing algorithms (Canu, Flye, and Miniasm), or to combine ONT reads with others obtained with second-generation NGS platforms for a hybrid assembly (Unicycler and MaSuRCA). The latter provides more

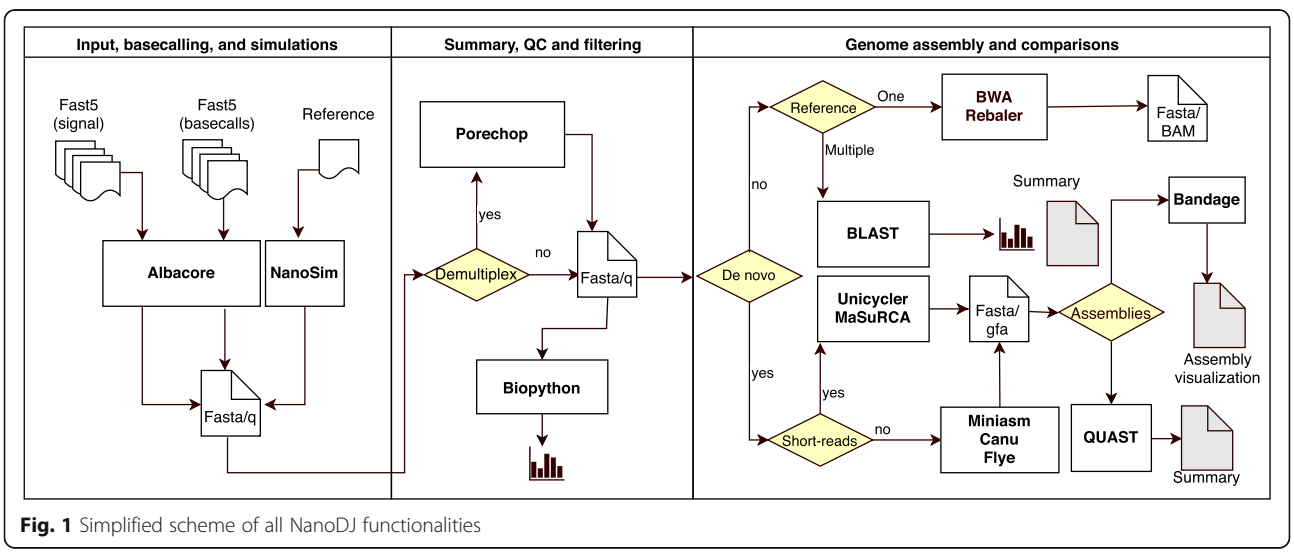

Fig. 1 Simplified scheme of all NanoDJ functionalities

Table 1 Summary of NanoDJ notebooks

Name	Functionality
0.0_QualityControl.ipynb	Evaluate the quality control and sequence handling
1.0_Basecalling.ipynb	Translates the events or the raw electrical signal from an ONT sequencer (FAST5 format) to a DNA sequence to obtain a FASTA or a FASTQ file
1.1_Trim+Demux.ipynb	Perform sequence trimming and demultiplexing
2.0_DeNovo_Canu-Miniasm.ipynb	De novo assembly with Canu or Miniasm, and polish with Racon and Pilon
3.0_DeNovo_Canu +polish.ipynb	Nanopolish modules to improve the Canu assembly
4.0_DeNovo_Flye.ipynb	De novo assembly with Flye software
5.0_DeNovo_Hybrid.ipynb	Perform de novo assembly of Nanopore reads in conjunction with Illumina reads using MaSuRCA and/or Unicycler software
6.0_AssemblyCompare.ipynb	Compare distinct assembly results based on QUAST software
7.0_SimulateReads.ipynb	Obtain simulated reads made with Nanosim software and the Nanosim-h fork with precomputed models
8.0_Alignment.ipynb	Reference-based assembly using either BWA, BLAST or Rebaler software
9.0_AssemblyGraph.ipynb	Assembly graph visualization
Educational.ipynb	Performs basecalling (with Albacore), quality control steps, and a BLAST-based classification of the reads (for educational purposes)

effective assemblies and reduced error rate compared to assemblies based only on ONT reads [14]. NanoDJ includes the possibility of contig correction (Racon, Nanopolish, and Pilon). Assemblies can be evaluated with the embedded version of QUAST, and represented with Bandage.

Limitations and future directions

For non-expert users, it would have been better if NanoDJ was envisaged as an on-line application to facilitate its use. However, our main objective was to integrate major tools for the analysis of ONT sequences in an interactive software environment to facilitate learning the basics behind ONT sequence analysis while providing a useful tool for professionals. Providing it as a Dockerized solution simply bolsters the focus on the use of the tool, reducing the burden of installing all dependencies by the user. At the moment, NanoDJ is set for the analysis of small genomes and targeted NGS studies, although focusing on primary and secondary analysis of DNA sequences. The integration of tools for variant identification and tertiary analysis (annotation of variants or sequence elements, interpretation, etc.) [15, 16], as well as for epigenetics [17] and direct RNA sequencing [18] will be the focus of further developments of NanoDJ.

Conclusions

We present NanoDJ as an integrated Jupyter-based toolbox distributed as a Docker software container to facilitate ONT sequence analysis. NanoDJ is best suited for the analyses of small genomes and targeted NGS studies. We anticipate that the Jupyter notebook-based structure will simplify further developments in other applications.

Availability and requirements

Project name: NanoDJ

Project home page: https://github.com/genomicsITER/NanoDJ

Operating system(s): Windows, Linux, Mac OS

Programming language: Bash/Python

Other requirements: Docker installation

License: GPL

Any restrictions to use by non-academics: None

Additional file

Additional file 1: Table S1. Applications integrated in NanoDJ. **Text S1.** Testing on case study datasets. **Table S2.** Datasets for illustrative uses of NanoDJ. **Table S3.** Comparison of de novo assemblies using different inputs or with an assembly corrector. **Table S4.** Comparison of three de novo assemblers in a high-coverage ONT dataset. **Table S5.** Comparison of results from two hybrid de novo assemblers. **Figure S1.** Human mitochondrial DNA variant representation against the reference sequence. **Table S6.** Source of mitochondrial DNA genomes, simulations and classification results. (DOCX 1544 kb)

Acknowledgements

Not applicable

Funding

This research was funded by the Instituto de Salud Carlos III (grants PI14/00844 and PI17/00610), the Spanish Ministry of Science, Innovation and Universities (grant RTC-2017-6471-1; MINECO/AEI/FEDER, UE), the Spanish Ministry of Economy and Competitiveness (grant MTM2016–74877-P), which were co-financed by the European Regional Development Funds 'A way of making Europe' from the European Union, Area Tenerife 2030 from Cabildo de Tenerife (CGIEU0000219140), and by the agreement OA17/008 with Instituto Tecnológico y de Energías Renovables (ITER) to strengthen scientific and technological education, training, research, development and innovation in Genomics, Personalized Medicine and Biotechnology. The founding entities had no role in the design of the study, analysis, interpretation of data or in manuscript writing.

Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files. Raw reads from MinION and Illumina are available from the SRA database (accession number(s) PRJNA451111, PRJNA451107).

Authors' contributions

HRP scripted and tested the software, and contributed to data analysis; THB was involved in data analysis and interpretation; JLS was involved in data analysis; JRG revised and tested the software and revised the manuscript; CPG was involved in visualization, data analysis and revised the manuscript; MC conceived the project, revised and tested the software, and revised the manuscript; CF conceived the project, designed the software, interpreted the data, and critically revised the manuscript. All authors have read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Research Unit, Hospital Universitario Nuestra Señora de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain. [2]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain. [3]Departamento de Ingeniería Informática y de Sistemas, Universidad de La Laguna, Santa Cruz de Tenerife, Spain. [4]Departamento de Matemáticas, Estadística e Investigación Operativa, Universidad de La Laguna, Santa Cruz de Tenerife, Spain. [5]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain.

Received: 22 June 2018 Accepted: 29 April 2019
Published online: 09 May 2019

References
1. Brown CG, Clarke J. Nanopore development at Oxford Nanopore. Nat Biotechnol. 2016;34:810–1.
2. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 2016;17:239.
3. Quick J, Loman NJ, Duraffour S, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530:228–32.
4. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, Kraemer MUG, Hill SC, Black A, da Costa AC, Franco LC, Silva SP, Wu C-H, Raghwani J, Cauchemez S, du Plessis L, Verotti MP, de Oliveira WK, Carmo EH, Coelho GE, Santelli ACFS, Vinhal LC, Henriques CM, Simpson JT, Loose M, Andersen KG, Grubaugh ND, Somasekar S, Chiu CY, Muñoz-Medina JE, Gonzalez-Bonilla CR, Arias CF, Lewis-Ximenez LL, Baylis SA, Chieppe AO, Aguiar SF, Fernandes CA, Lemos PS, Nascimento BLS, Monteiro HAO, Siqueira IC, de Queiroz MG, de Souza TR, Bezerra JF, Lemos MR, Pereira GF, Loudal D, Moura LC, Dhalia R, França RF, Magalhães T, Marques ET Jr, Jaenisch T, Wallau GL, de Lima MC, Nascimento V, de Cerqueira EM, de Lima MM, Mascarenhas DL, Neto JPM, Levin AS, Tozetto-Mendoza TR, Fonseca SN, Mendes-Correa MC, Milagres FP, Segurado A, Holmes EC, Rambaut A, Bedford T, Nunes MRT, Sabino EC, Alcantara LCJ, Loman NJ, Pybus OG. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017;546:406–10.
5. Castro-Wallace SL, Chiu CY, John KK, Stahl SE, Rubins KH, McIntyre ABR, Dworkin JP, Lupisella ML, Smith DJ, Botkin DJ, Stephenson TA, Juul S, Turner DJ, Izquierdo F, Federman S, Stryke D, Somasekar S, Alexander N, Yu G, Mason CE, Burton AS. Nanopore DNA sequencing and genome assembly on the international Space Station. Sci Rep. 2017;7:18022.
6. Johnson SS, Zaikova E, Goerlitz DS, Bai Y, Tighe SW. Real-time DNA sequencing in the Antarctic dry valleys using the Oxford Nanopore sequencer. J Biomol Tech. 2017;28(1):2–7.
7. Pomerantz A, Peñafiel N, Arteaga A, Bustamante L, Pichardo F, Coloma LA, Barrio-Amoros CL, Salazar-Valenzuela D, Prost S. Real-time DNA barcoding in a remote rainforest using nanopore sequencing. Gigascience. 2018;7(4):giy033.
8. Menegon M, Cantaloni C, Rodriguez-Prieto A, Centomo C, Abdelfattah A, Rossato M, Bernardi M, Xumerle L, Loader S, Delledonne M. On site DNA barcoding by nanopore sequencing. PLoS One. 2017;12:e0184741.
9. Zaaijer S, Columbia University Ubiquitous genomics 2015 class, Erlich Y: using mobile sequencers in an academic classroom. Elife. 2016, 5:e14258.
10. Almugbel R, Hung LH, Hu J, Almutairy A, Ortogero N, Tamta Y, Yeung KY. Reproducible Bioconductor workflows using browser-based interactive notebooks and containers. J Am Med Inform Assoc. 2018;25:4–12.
11. Grüning BA, Rasche E, Rebolledo-Jaramillo B, Eberhard C, Houwaart T, Chilton J, Coraor N, Backofen R, Taylor J, Nekrutenko A. Jupyter and galaxy: easing entry barriers into complex data analyses for biomedical researchers. PLoS Comput Biol. 2017;13:e1005425.
12. Boettiger C. An introduction to Docker for reproducible research. Oper Syst Rev. 2015;49:71–9.
13. Di Tommaso P, Palumbo E, Chatzou M, Prieto P, Heuer ML, Notredame C. The impact of Docker containers on the performance of genomic pipelines. PeerJ. 2015;3:e1273.
14. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb Genom. 2017;3:e000132.
15. Cook DE, Valle-Inclan JE, Pajoro A, Rovenich H, Thomma B, Faino L. Long-read annotation: automated eukaryotic genome annotation based on long-read cDNA sequencing. Plant Physiol. 2019;179:38–54.
16. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15:461–8.
17. Stoiber MH, Quick JF, Egan R, Lee JE, Celniker SE, Neely RK, Loman NJ, Pennacchio LA, Brown JO. De novo identification of DNA modifications enabled by genome-guided Nanopore signal processing. bioRxiv. . https://doi.org/10.1101/094672.
18. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, Jordan M, Ciccone J, Serra S, Keenan J, Martin S, McNeill L, Wallace EJ, Jayasinghe L, Wright C, Blasco J, Young S, Brocklebank D, Juul S, Clarke J, Heron AJ, Turner DJ. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods. 2018;15:201–6.