ONTdeCIPHER: An amplicon-based nanopore sequencing pipeline for tracking pathogen variants

Emira Cherif^{1*}, Fatou Seck Thiam¹⁺, Mohammad Salma^{2,3+}, Georgina Rivera-Ingraham^{1,4}, Fabienne Justy¹, Theo Deremarque¹, Damien Breugnot¹, Jean-Claude Doudou⁴, Rodolphe Elie Gozlan¹, Marine Combe¹

¹ISEM, Univ Montpellier, CNRS, IRD, Montpellier, France

²Institut de Génétique Moléculaire de Montpellier, Université de Montpellier, CNRS, Montpellier, France.

³Université de Paris, Laboratory of Excellence GR-Ex, Paris, France.

⁴IRD, Centre IRD de Cayenne, Guyane Française

*To whom correspondence should be addressed.

+The authors wish it to be known that, in their opinion, the second and third authors should be regarded as Joint Second Authors.

Abstract

Motivation: Amplicon-based nanopore sequencing is increasingly used for molecular surveillance during epidemics (e.g. ZIKA, EBOLA) or pandemics (e.g. SARS-CoV-2). However, there is still a lack of versatile and easy-to-use tools that allow users with minimal bioinformatics skills to perform the main steps of downstream analysis, from quality testing to SNPs effect to phylogenetic analysis.

Results: Here, we present ONTdeCIPHER, an amplicon-based Oxford Nanopore Technology (ONT) sequencing pipeline to analyze the genetic diversity of SARS-CoV-2 and other pathogenes. Our pipeline integrates 13 bioinformatics tools. With a single command line and a simple configuration file, users can pre-process their data and obtain the sequencing statistics, reconstruct the consensus genome, identify variants and their effects for each viral isolate, infer lineage and, finally perform multi-sequence alignments and phylogenetic analyses.

Availability and implementation: ONTdeCIPHER is available at <https://github.com/emiracherif/ONTdeCIPHER>

Contact: emira.cherif@ird.fr

Supplementary information: Supplementary data are available at ...

1 Introduction

The SARS-CoV-2 pandemic resulted in over 231 million cases and 4.74 million deaths worldwide in less than two years (Dong et al., 2020). Since the first spark of the epidemic, global public health actors and international research teams have made considerable efforts that have resulted in unparalleled knowledge, from the identification of the viral genome to its biological mechanisms of infection and spread. A striking example is the first SARS-CoV-2 genome sequence published less than a month after the first viral infection report in December 2019, followed by thousands of genome sequences identified and published. Genome sequencing has enabled rapid identification of the virus, development of diagnostic tests and monitoring of the virus' evolution (WHO, 2021). Nanopore sequencing technology combined with ARTIC's Multiplex PCR-based protocol (Quick et al., 2017) is widely used to rapidly and reliably generate large amounts of raw sequencing data for molecular surveillance of the SARS-CoV-2 pandemic and the ZIKA, Ebola and Nipah outbreaks. However, there is still a lack of a versatile, practical tool that can perform the main steps of downstream analysis, from quality testing to SNPs effect to phylogenetic analysis, on a laptop or an HPC (high performance computing cluster) without the need for advanced bioinformatics skills, to meet the demand of recurrent analysis of hundreds of samples. In addition, few tools are specific to nanopore sequencing based on SARS-CoV-2 amplicons allowing genotyping and genome reconstruction (e.g. porCOV (Brandt et al., 2021), viralrecon (Patel et al., 2020) and interARTIC (Ferguson et al., 2021)). Here, we present ONTdeCIPHER, an amplicon-based Oxford Nanopore Technology (ONT) sequencing pipeline to analyze and assess the genetic diversity of SARS-CoV-2. ONTdeCIPHER integrates and orchestrates 13 bioinformatics tools combining ONT dedicated tools, SARS-CoV-2 specific tools and well-established generic tools. It is a stand-alone pipeline compatible with Ubuntu, mac and HPC distributions, including an

easy-to-use configuration file. With a single command line and raw sequencing data as input, the users can pre-process their data and obtain sequencing quality, depth and coverage statistics. It then reconstructs consensus genomic sequences, identifies variants and their potential associated effects for each viral isolate, detects structural variants (SVs), infers lineage and finally performs multi-sequence alignments and phylogenetic analyses.

2 Features

While building on the main features of the ARTIC pipeline (<https://github.com/artic-network/fieldbioinformatics>), the ONTdeCIPHER pipeline incorporates additional useful features such as variant calling, variant annotation, lineage inference, multiple alignments and phylogenetic tree construction (Fig.1).

Pre-processing and quality control

ONTdeCIPHER performs a pre-filtering step based on read length, followed by reads loss estimation using guppyplex *via* the ARTIC pipeline and SeqKit (Shen et al., 2016) respectively. The user can choose the maximum and minimum read length, with the default value of 400-700 bp (ARTIC primers V3). pycoQC is used by ONTdeCIPHER to calculate sequencing metrics (read length distribution, read quality scores, etc.). Then, using the ARTIC pipeline implemented in ONTdeCIPHER, the filtered reads are mapped to the Wuhan reference genome (Accession: NC_045512.2) using minimap2 (Li, 2018). Finally, amplicon dropouts during the multiplex PCR step and sequencing depth and coverage for each sample are estimated by ONTdeCIPHER using MultiQC (Ewels et al., 2016), plotCoverage (Ramírez et al., 2014) and bamCoverage (Ramírez et al., 2014), respectively.

Genome reconstruction and genomic analysis

For each sample, ONTdeCIPHER reconstructs a reference-based consensus genome *via* the Medaka workflow and infers the viral lineage using the Pangolin Command-line tool (Rambaut et al., 2020). ONTdeCIPHER performs two variant calling steps for SNV (Simple

Nucleotide Variation) and SV (Structural Variation) genotyping using Medaka and Sniffles (Sedlazeck et al., 2018). SNP variants are annotated using SnpEff (Cingolani et al., 2012). Finally, ONTdeCIPHER performs multiple alignments followed by phylogenetic tree construction for consensus genomic sequences obtained using the MAFFT command-line (Kato et al., 2019), RAxML (Stamatakis, 2014) and ETE3 (Huerta-Cepas et al., 2016).

3 Implementation

ONTdeCIPHER is an open-source pipeline coded in python3 compatible with Ubuntu distributions and macOS. It uses the Snakemake (Köster et al., 2012) features to perform reproducible and scalable data analysis (see Supplementary information Fig.S1) and seamlessly adapts to server, cluster, grid and cloud environments. ONTdeCIPHER's dependencies and environments are installed and managed using the Conda package and environment management system. Currently, ONTdeCIPHER can be run locally or on an HPC.

ONTdeCIPHER includes 13 tools (see Supplementary information Table S1) validated by the bioinformatics community to perform key downstream analyses on raw sequencing data from amplicon-based ONT sequencing. In addition, it comes with a python master script (*run_deCiPHER.py*) allowing the user to run the various pipelines' steps *via* a command line. ONTdeCIPHER requires two configuration files to run; 1) one containing the path to raw sequencing data (fastq and fast5 directories), the path to the sequencing_summary.txt file (optional), the minimum and maximum read lengths, the RaxML output name prefix (optional) and the version of the reference genome used, 2) the second is a tabulated file that contains the barcodes and the associated sample.

The main option of the master script is (--step). The user can select one of the following values: *pycoQC*, *pip_core*, *m_r_p* or *all*. *PycoQC* will calculate sequencing metrics and generate interactive QC plots, *pip_core* will run ONTdeCIPHER core pipeline (ARTIC

pipeline, seqKit, DeepTools, Sniffles and snpEff), *m_r_p* will run mafft, raxml, ete3 and Pangolin while *all* will run all steps.

Output sub-directories are generated for each enabled step following a specific architecture (see the github repository for details). ONTdeCIPHER provides several visualization files (HTML, pdf, bedgraph, bigwig) to visualize and describe the results.

4 Applications

ONTdeCIPHER was tested on a subset of public clinical data of SARS-CoV-2 (Bioproject PRJNA675364) from Bull et al (2020) (see github repository test_data) and the SVs found mainly deletions, were comparable (see Supplementary information Table S2). ONTdeCIPHER was also tested on data from wastewater samples collected in French Guiana (South America) to assess the spatio-temporal dynamics and prevalence of the SARS-CoV-2 viral variants (see Supplementary information Fig. S2). ONTdeCIPHER was compared to the closest available tool, porCOV (Brandt et al., 2021) (see Supplementary information Table S3).

5 Conclusion and perspectives

ONTdeCIPHER uses a plethora of well-established tools to provide a practical pipeline for amplicon-based sequencing downstream analyses. ONTdeCIPHER aims to provide an efficient tool for large-scale monitoring of SARS-CoV-2 viral variants dynamics, but also of broader surveillance of a multitude of emerging environmental-based pathogens.

Acknowledgements

We acknowledge Dr. Stéphane Calmant, director of the IRD Centre at Cayenne, Dr. Mathieu Chouteau, researcher at CNRS Guyane, and Vincent Goujon, director of CNRS Guyane, for their logistical support and the availability of the laboratory facilities at Cayenne. We thank our local collaborators in French Guiana for their help to access wastewater samples (the Cayenne Prefecture, CTG, ARS, CACL, the town hall of Kourou, Sinnamary, Saint-Laurent, Apatou, Saint-Georges). We also acknowledge Marie-Ka Tilak for her precious advice,

experience and discussions about viral RNA extractions and ONT library preparation and sequencing.

Funding

This work was funded by the ANR (ANR-20-COV5-000) in French Guiana and the Institut de Recherche pour le Développement (IRD). EC received a postdoctoral fellowship from IRD, GRI a postdoctoral fellowship from ANR, FST a student fellowship from ANR and DB a student fellowship from IRD.

References

- Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, 20(5), 533-534.
- World Health Organization. (2021). Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health, 8 January 2021.
- Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., ... & Loman, N. J. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature protocols*, 12(6), 1261-1276.
- Brandt, C., Krautwurst, S., Spott, R., Lohde, M., Jundzill, M., Marquet, M., & Hölzer, M. (2021). poreCov-an easy to use, fast, and robust workflow for SARS-CoV-2 genome reconstruction via nanopore sequencing. *Frontiers in Genetics*. 12, 1397. <https://doi.org/10.3389/fgene.2021.711437>
- Patel, H. et al. nf-core/viralrecon: nf-core/viralrecon v1.1.0—Steel Pangolin. (Zenodo, 2020). <https://doi.org/10.5281/zenodo.3905178>.
- Ferguson, J. M., Gamaarachchi, H., Nguyen, T., Gollon, A., Tong, S., Aquilina-Reid, C., ... & Deveson, I. W. (2021). InterARTIC: an interactive web application for whole-genome nanopore sequencing analysis of SARS-CoV-2 and other viruses. *bioRxiv*.
- Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19), 2520-2522.
- Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS one*, 11(10), e0163962.
- Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094-3100.
- Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), 3047-3048.

- Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research*, 42(W1), W187-W191.
- Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., ... & Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature microbiology*, 5(11), 1403-1407.
- Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature methods*, 15(6), 461-468.
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., ... & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, 6(2), 80-92.
- Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics*, 20(4), 1160-1166.
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312-1313.
- Bull, R. A., Adikari, T. N., Ferguson, J. M., Hammond, J. M., Stevanovski, I., Beukers, A. G., ... & Deveson, I. W. (2020). Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nature communications*, 11(1), 1-8.
- Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*, 33(6), 1635-1638.

Fig. 1.

Overall ONTdeCIPHER pipeline. The diagram shows the ONTdeCIPHER features with the associated tool. All quality control features have the same color (dark yellow). Only the main results, files and figures are shown.