bioRxiv preprint doi: https://doi.org/10.1101/094979; this version posted December 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

poRe GUIs for parallel and real-time processing of MinION sequence data

Robert Stewart[1] and Mick Watson[1,2,*]

[1]Department of Genetics and Genomics, The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, EH25 9RG,

[2]Edinburgh Genomics, The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, EH25 9RG

*To whom correspondence should be addressed.

Abstract

Motivation: Oxford Nanopore's MinION device has matured rapidly and is now capable of producing over one million reads and several gigabases of sequence data per run. The nature of the MinION output requires new tools that are easy to use by scientists with a range of computational skills and which enable quick and simple QC and data extraction from MinION runs.

Results: We have developed two GUIs for the R package poRe that allow parallel and real-time processing of MinION datasets. Both GUIs are capable of extracting sequence- and meta- data from large MinION datasets via a friendly point-and-click interface using commodity hardware.

Availability: The GUIs are packaged within poRe which is available on SourceForge: https://sourceforge.net/projects/rpore/files/.

Documentation is available on GitHub: https://github.com/mw55309/poRe_docs

Contact: mick.watson@roslin.ed.ac.uk

1 Introduction

Nanopore sequencing is the only sequencing technology that measures an actual single molecule of DNA, rather than incorporation events into a template strand [1,2]. Early access to Oxford Nanopore's MinION , a portable DNA sequencer approximately six inches in length, began in 2014. The MinION may be considered a mature platform, having been used to sequence bacterial genomes [3,4]; resolve repeats in the human genome [5]; study cDNA structure [6,7]; detect base modifications [8–10]; detect antibiotic resistance [11]; perform real-time enrichment ('read until')[12]; and provide surveillance in a human disease outbreak [13]. The latest chemistry release, R9.4, has seen the first high-coverage human genome data released (https://github.com/nanopore-wgs-consortium/NA12878; https://github.com/nanoporetech/ONT-HG1), with several MinION flowcells from the two projects producing over 4 gigabases (Gb) of sequence data.

The MinION has been designed to enable mobile, real-time sequencing. As soon as a sequencing library is placed onto the device, the MinION begins sequencing. Each channel/nanopore reports asynchronously, creating a single file per channel per read. These are created in HDF5, a compressed binary hierarchical data format (https://www.hdfgroup.org/). Depending on the sequencer and chemistry version, these HDF5 files include raw or event-level signal data, recorded as a DNA molecule passed through the pore. There are a range of base-calling options, including cloud-based Metrichor, local MinKNOW base-calling and open-source alternatives [14,15], that will convert the signal data into DNA sequences.

With 512 pores and a sequencing speed of 250 bases-per-second, each MinION flowcell has the capacity to produce several million reads in a 48-hour run. As each read presents as two files (one raw, one base-called) MinION runs represent huge challenges for researchers without sufficient computational skills. Tools exist, such as poRe[16] and poretools [17], to assist with this, but many are command-line based, and there is a need for easy-to-use, GUI-based tools for MinION data QC and analysis.

bioRxiv preprint doi: https://doi.org/10.1101/094979; this version posted December 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

2 Methods

We have designed and built two graphical-user-interfaces (GUIs) for MinON data processing, organization and extraction. Both are built as Shiny apps and released as part of the package poRe [16].

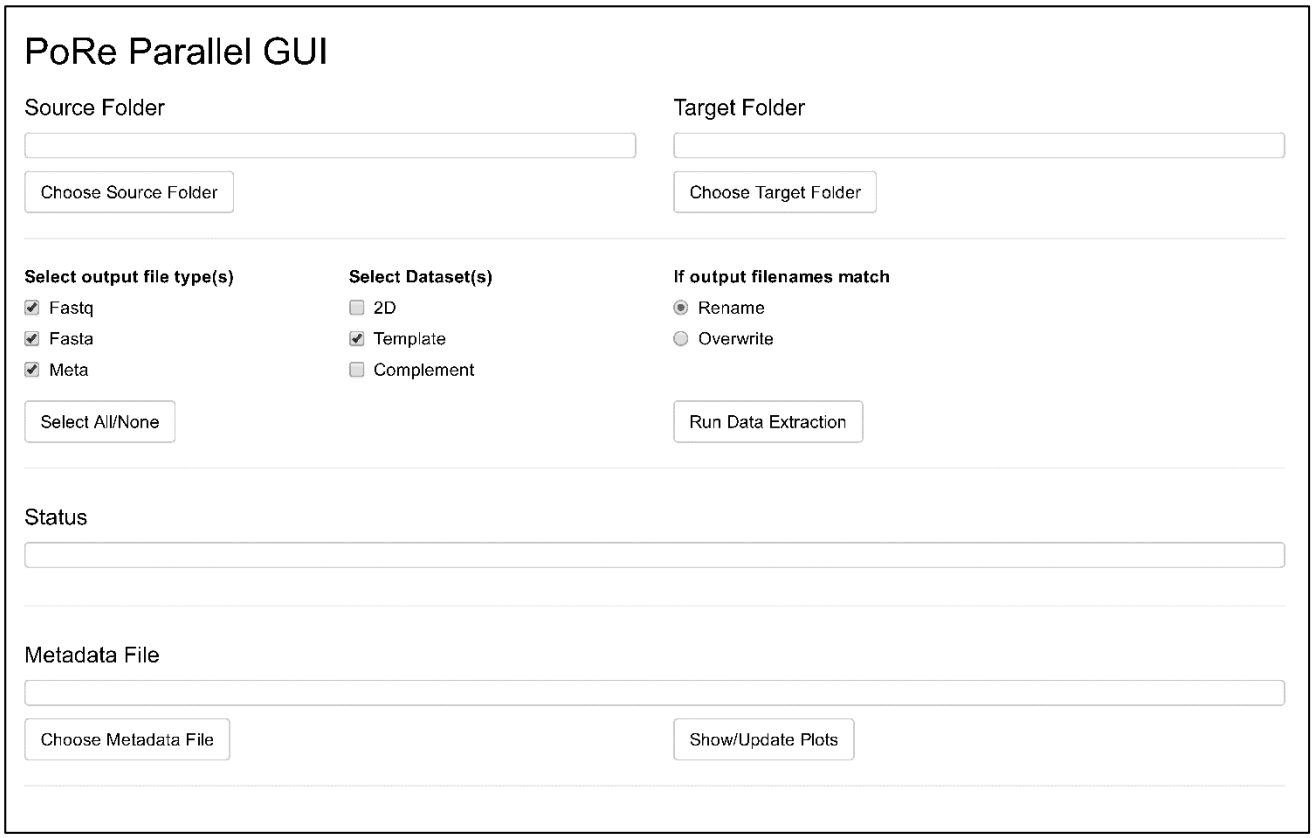

Figure 1. Screenshot of the pore parallel GUI, which as a Shiny App will open in the user's browser

The poRe real-time GUI is designed to extract data (FASTQ, FASTA and metadata) during a run, or during base-calling. A source and destination folder are required. The software then monitors the source folder for new FAST5 files; as FAST5 files arrive in the folder, they are processed, data are extracted and output to the destination folder. The poRe real-time GUI saves researchers a huge amount of time as data can be extracted while the MinION is running. The poRe real-time GUI has built in parallelization using R's built in parallel package, and is accessed by running the command pore_rt().

The pore parallel GUI is designed to extract data from runs that have already finished. Again, the software expects a source and destination folder; in addition, the user can select which data to extract, and the number of cores to use. The software then extracts FASTQ, FASTA and metadata from all files in the source folder into files in the destination folder; using the number of cores specified by the user, via the parallel package. The poRe parallel GUI is accessed via the pore_parallel() command.

3 Results

The poRe parallel GUI was able to simultaneously extract FASTQ, FASTA and metadata from 209,819 FAST5 files downloaded from the "cliveome" project in just 37 minutes on our 16-core Linux server, at a rate of approx. 90 FAST5 files per second.

Funding

This work was supported by The Biotechnology and Biological Sciences Research Council (BBSRC) including institute strategic support to The Roslin Institute (BB/M020037/1, BB/J004243/1, BB/J004235/1, BBS/E/D/20310000).

Conflict of Interest: the authors have received free flowcells and reagents from Oxford Nanopore as part of the MAP. Mick Watson has attended Oxford Nanopore events and had his travel paid for by ONT.

bioRxiv preprint doi: https://doi.org/10.1101/094979; this version posted December 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

1. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17,** 333–351 (2016).
2. Loman, N. J. & Watson, M. Successful test launch for nanopore sequencing. *Nat. Methods* **12,** 303–304 (2015).
3. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12,** 733–5 (2015).
4. Risse, J. *et al.* A single chromosome assembly of Bacteroides fragilis strain BE1 from Illumina and MinION nanopore sequencing data. *Gigascience* **4,** 60 (2015).
5. Jain, M. *et al.* Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* **12,** 351–356 (2015).
6. Hargreaves, A. D. & Mulley, J. F. Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing. *PeerJ* **3,** e1441 (2015).
7. Bolisetty, M. T., Rajadinakaran, G. & Graveley, B. R. Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biol.* **16,** 204 (2015).
8. Rand, A. C. *et al. Cytosine Variant Calling with High-throughput Nanopore Sequencing*. *bioRxiv* (Cold Spring Harbor Labs Journals, 2016). doi:10.1101/047134
9. Karlsson, E. *et al.* Scaffolding of a bacterial genome using MinION nanopore sequencing. *Sci. Rep.* **5,** 11996 (2015).
10. Stoiber, M. H. *et al.* De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv* (2016).
11. Ashton, P. M. *et al.* MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat. Biotechnol.* **33,** 296–300 (2014).
12. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nat. Methods* **13,** 751–754 (2016).
13. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530,** 228–32 (2016).
14. David, M., Dursi, L. J., Yao, D., Boutros, P. C. & Simpson, J. T. Nanocall: An Open Source Basecaller for Oxford Nanopore Sequencing Data. *Bioinformatics* btw569 (2016). doi:10.1093/bioinformatics/btw569
15. Boža, V., Brejová, B. & Vinař, T. DeepNano: Deep Recurrent Neural Networks for Base Calling in MinION Nanopore Reads. (2016).
16. Watson, M. *et al.* poRe: an R package for the visualization and analysis of nanopore sequencing data. *Bioinformatics* **31,** 114–5 (2015).
17. Loman, N. J. & Quinlan, A. R. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* **30,** 3399–401 (2014).

PoRe Parallel GUI

bioRxiv preprint doi: https://doi.org/10.1101/094879; this version posted December 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Source Folder

Choose Source Folder

Target Folder

Choose Target Folder

Select output file type(s)

- [x] Fastq
- [x] Fasta
- [x] Meta

Select All/None

Select Dataset(s)

- [ ] 2D
- [x] Template
- [ ] Complement

If output filenames match

- (•) Rename
- ( ) Overwrite

Run Data Extraction

Status

Metadata File

Choose Metadata File

Show/Update Plots