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

Citation for published version:
Stewart, R & Watson, M 2017, 'poRe GUIs for parallel and real-time processing of MinION sequence data' Bioinformatics, vol. 33, no. 14, pp. 2207-2208. DOI: 10.1093/bioinformatics/btx136

Digital Object Identifier (DOI):
10.1093/bioinformatics/btx136

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Bioinformatics

Publisher Rights Statement:
© The Author(s) 2017. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Sequence Analysis

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

Robert Stewart¹ and Mick Watson¹,²,*

¹Department of Genetics and Genomics, The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, EH25 9RG,
²Edinburgh Genomics, The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, EH25 9RG

*To whom correspondence should be addressed.
Associate Editor: Prof. Bonnie Berger

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

Supplementary information: Supplementary data are available at Bioinformatics online.

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 (Goodwin et al., 2016; Loman and Watson, 2015). 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 (Loman et al., 2015; Risse et al., 2015); resolve repeats in the human genome (Jain et al., 2015); study cDNA structure (Hargreaves and Mulley, 2015; Bolisetty et al., 2015); detect base modifications (Rand et al., 2016; Karlsson et al., 2015; Stoiber et al., 2016); detect antibiotic resistance (Ashton et al., 2014); perform real-time enrichment (‘read until’)(Loose et al., 2016); and provide surveillance in a human disease outbreak (Quick et al., 2016). 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 (David et al., 2016; Boža et al., 2016), that will convert the signal data into DNA sequences.

With 512 pores and a sequencing speed of several hundred 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(Watson et al., 2015) and porertools (Loman and Quintlan, 2014), 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.

2 Methods

We have designed and built two graphical-user-interfaces (GUIs) for MinION data processing, organization and extraction. Both are built as Shiny apps and released as part of the package poRe (Watson et al., 2015). At present the original poRe and the new GUI code are separate, but we envisage merging the functions over time. Both are available through the R package poRe. 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...
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.

### References


