NanoOK: flexible, multi-reference software for pre- and post-alignment analysis of nanopore sequencing data, quality and error profiles

Richard M. Leggett, Darren Heavens, Mario Caccamo, Matthew D. Clark, Robert P. Davey
The Genome Analysis Centre, UK

Abstract

The recent launch of the Oxford Nanopore Technologies MinION® Access Program (MAP) resulted in the rapid development of a number of open source tools [1, 2, 3] aimed at extracting reads and yield information from the HDF5 format files produced by the platform. However, all tools so far published only facilitate the production of FASTA/Q files, with only [3] providing assistance with alignment. In particular, none of the tools provides alignment-based analysis and error profiling of Nanopore reads, something that is critical in order to understand the applicability of the application to a new problem area and is often performed ad hoc. NanoOK has been written to address this gap. Here, we present a description of the tool as well as a comparative analysis of error profiles across runs performed on the R7.3 version of the MinION® flowcells and chemistry, using real world samples.

Reports

Process the HDF5 files outputted by the MinION® basecaller
- Extract FASTA/Q files (NanoOK or 3rd party tool, e.g. [1])
- Align to references
- Calculate QC and error metrics
- Consolidate information into a PDF report

Features

NanoOK produces in-depth data on a variety of key metrics (see Figures (a) to (e)):
- number of reads aligning
- quality of alignment summaries (reads aligning, coverage, strand bias, perfect sequence length)
- coverage plots (template, complement and 2D reads)
- longest perfect k-mer analysis
- rigorous error analysis (location, overall identify, substitutions, indels, plots of size distribution)
- error motifs (preceeding n-mers before observed errors)
- filter on pass, fail or pass+fail

NanoOK also supports multiple concurrent references, enabling analysis of metagenomic samples and pooled libraries.

Technical Details

• Fully open source, developed in Java
• R for graph plotting
• LaTeX for reporting
• Simple command line interface and execution
• Includes in-built parser for LAST alignments
• Extensible – parser classes can be written easily by extending existing code

Availability

All source code is available in the TGAC GitHub repository:
https://github.com/TGAC/NanoOK

Documentation:
https://documentation.tgac.ac.uk/display/NanoOK

Issue Tracker:
https://tracker.tgac.ac.uk/browse/NANOOK

References


Acknowledgements

NanoOK is funded by TGAC's Biotechnology and Biological Sciences Research Council (BBSRC) core strategic grant.