Toolkit to explore and analyze nanopore sequencing data on a Hadoop framework

Asunción Gallego[1], Joaquin Tarraga[1], Vicente Arnau[2], Ignacio Medina[3] and Joaquin Dopazo[1,4,5]

[1] Computational Genomics Department,Centro de Investigación Príncipe Felipe(CIPF)

[2] Departamento de informática,ETSE,Universidad de Valencia

[3] HPC Service, University Information Services, University of Cambridge

[4] Bioinformatics of Rare Diseases(BIER),CIBER de Enfermedades Raras(CIBERER)

[5] Functional Genomics Node,(INB) at CIPF

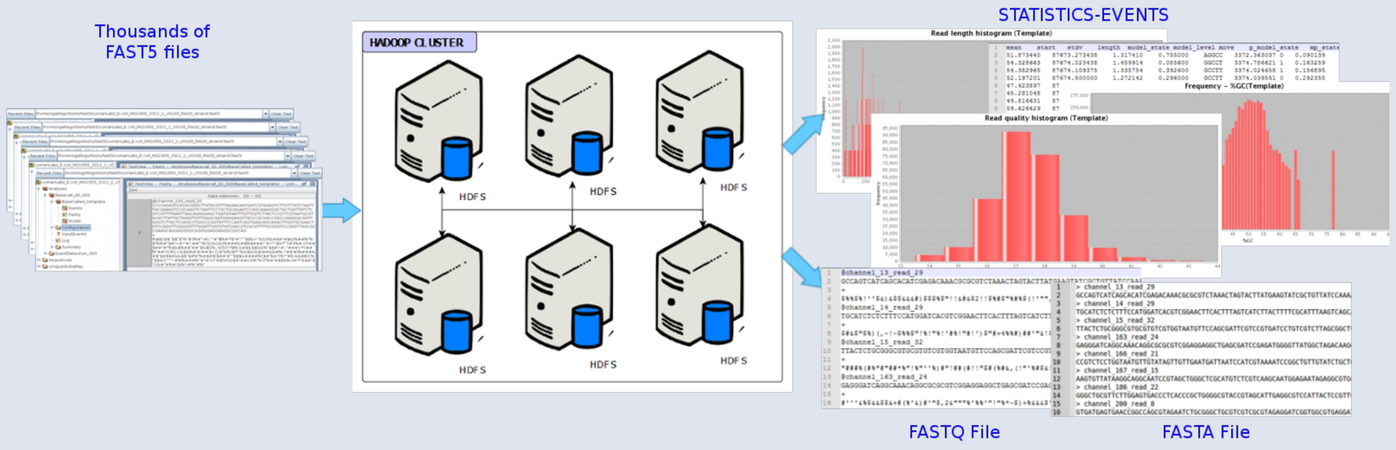
INTRODUCTION

The use of nanopore technologies is expected to spread in the future because they are portable and can sequence long fragments of DNA molecules without prior amplification. The first nanopore sequencer available, the MinION™ from Oxford Nanopore Technologies, is a USB-connected, portable device that allows real-time DNA analysis. In addition, other new instruments are expected to be released soon, which promise to outperform the current short-read technologies in terms of throughput. Despite the flood of data expected from this technology, the data analysis solutions currently available are only designed to manage small projects and are not scalable.

RESULTS

Here we present HPG Pore[1], a toolkit for exploring and analysing nanopore sequencing data. HPG Pore can run on both individual computers and in the Hadoop distributed computing framework, which allows easy scale-up to manage the large amounts of data expected to result from extensive use of nanopore technologies in the future.

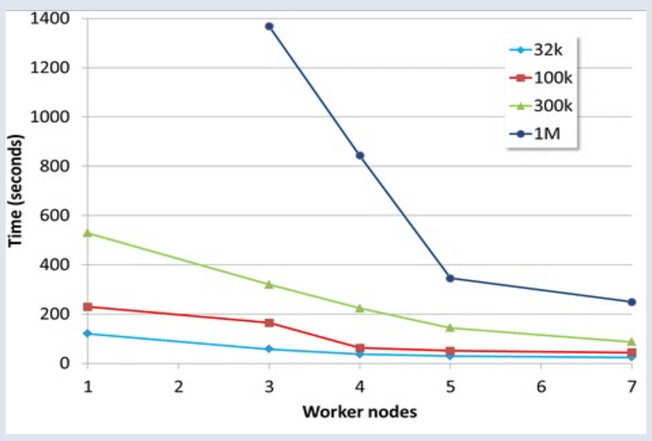
NANOPORE SEQUENCE ANALYSIS



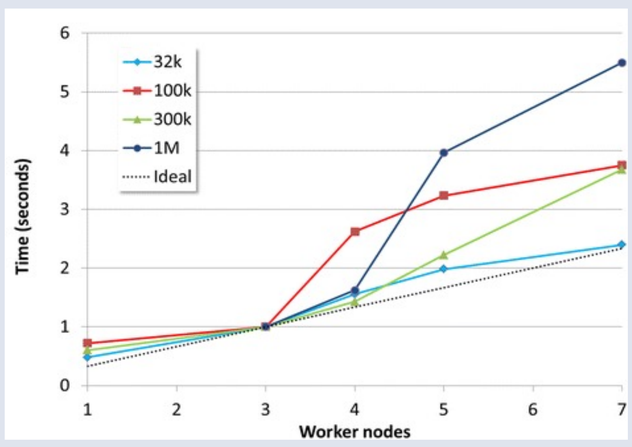
HPG PORE'S COMMANDS

Command HPG Pore	Local	Hadoop	Description
stats	yes	yes	Compute statistics and generates plots and histograms
events	yes	yes	Extract raw data of the electronic signal measured for a given MinION read
signal	yes	yes	Plot the electronic signal measured over time
Fasta	yes	yes	Extract the sequences in format FASTA
Fastq	yes	yes	Extract the sequences in format FASTQ
import	no	yes	Store the individual FAST5 files into a single Hadoop MapFile in the Hadoop Distributed File System(HDFS)
export	no	yes	Store the MinION reads to FAST5 local files from the Hadoop MapFile
fast5names	no	yes	Extract the filenames of the imported FAST5 files

HPG PORE'S RUNTIMES

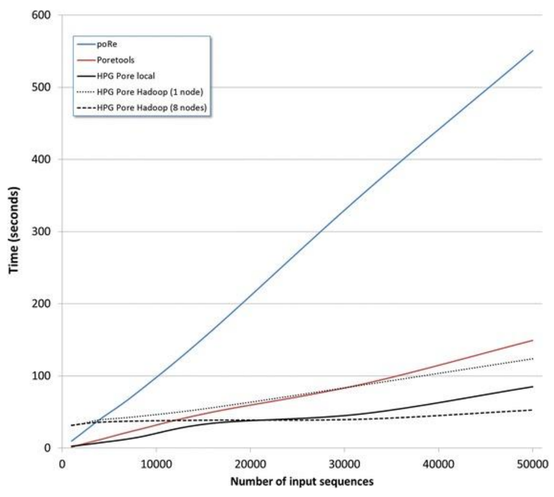


Runtimes (left panel) and increase in speed (right panel) as the number of nodes increase in the Hadoop system in two different scenarios: FAST5 file containing 32,000 (blue line), 100,000 (red line), 300,000 (green line) and 1 million (dark blue line) sequences. Dotted line in the lower panel represents the ideal speed-up according to the number of nodes used. Speed-ups have been calculated using 3 nodes as the starting point given that the 1 million reads could not be calculated for1 only one node.



FEATURES, RUNTIMES AND SCALABILITY ON THREE NANOPORE TOOLS HPG PORE, PORE AND PORETOOLS

Feature	HPG Pore	PoRe	Poretools
Extract Fastq	✓	✓	✓
Extract Fasta	✓	✓	✓
Organise fast5 into run folders	-	✓	-
Create tar files of runs	-	-	✓
Organise the results into run folders	✓	-	-
Plot yield	✓	✓	✓
Plot signal	✓	✓	✓
Extract run stats	✓	✓	✓
Read length histogram	✓	✓	✓
Read length (max., avg., min)	✓	✓	✓
Mean read quality	✓	-	-
Nucleotides content: count and %	✓	-	✓
%GC	✓	-	-
Plot Frequency- %GC	✓	-	-
Plot per base sequence content	✓	-	-
Read quality histogram	✓	-	-
Reads per channel histogram	✓	✓	✓
Nucleotides per channel histogram	✓	✓	-



Our study shows that runtimes in poRe, Poretools and HPG Pore (running locally) are approximately linearly dependent on the number of sequences in the FAST5 file, with a trend towards an increased slope for high numbers of sequences. HPG Pore runs the fastest, followed by Poretools, while poRe presents remarkably slower execution times.

CONCLUSIONS

HPG Pore allows for virtually unlimited sequencing data scalability, thus guaranteeing its continued management in near future scenarios. HPG Pore is available in GitHub at http://github.com/opencb/hpg-pore.

REFERENCES

[1] Tarraga J, Gallego A, Arnau V, Medina I, Dopazo J. 2016 HPG pore: an efficient and scalable framework for nanopore sequencing data. BMC Bioinformatics. 17(1):107

PRINCIPE FELIPE CENTRO DE INVESTIGACION
UNIVERSITY OF CAMBRIDGE
UNIÓN EUROPEA FONDO EUROPEO DE DESARROLLO REGIONAL "Una manera de hacer Europa"