An analysis of current software for Nanopore metagenomic data

Current state of the art software

Samantha Pendleton — sap21@aber.ac.uk — @sap218
Department of Computer Science, Aberystwyth University, Wales

Our insight into DNA is enabled by 'sequencing'. Until recently it was only possible to sequence DNA into short strings, 'reads'. Nanopore is a new sequencing technology to produce significantly longer reads. Using nanopore sequencing, a single molecule of DNA can be sequenced without the need for time consuming amplification.
Metagenomics is the study of genetic material recovered from environmental samples. A research team from Aberystwyth University have sampled metagenomes from a coal mine in South Wales using the Nanopore MinION and given initial taxonomic (classification of organisms) summaries of the microbial community. We are interested to discover how well current bioinformatics software works for quality control of this new long read data.

Introduction

Using various software, we want to analyse the data to observe sequence similarities and discover what bacteria resides within the mine [1].

- **BP_v1** (Dec-2016), with 1,770 reads
- **BP_v2** (Apr-2017), with 3,019 reads & extraction was an improved protocol

poretools

Toolkit for analysing nanopore sequence data [2] - developed Aug-2014 though 77 issues as of Sep-2017 on Github - explaining errors/bugs.

Read Length

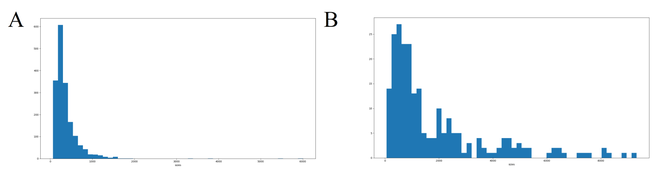

Figure 1: Histogram comparison of data-sets limited to 10,000 base pairs (due to long reads low in quality):- `x-axis`: read length (size); `y-axis`: cumulative frequency (count). `BP_v1` (A) has more short reads & `BP_v2` (B) has more long reads due to the improved extraction protocol.

Quality

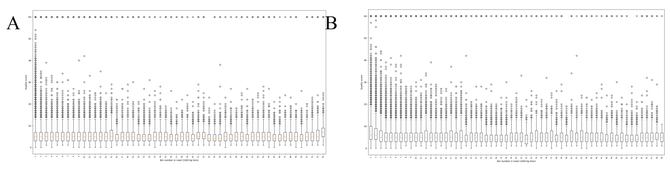

Figure 2: `poretools' qualpos` comparison - `BP_v1` (A) & `BP_v2` (B). The unusual high score of 60 could be an error with software or a random outlier - overall, the summary quality scores are low.

Time

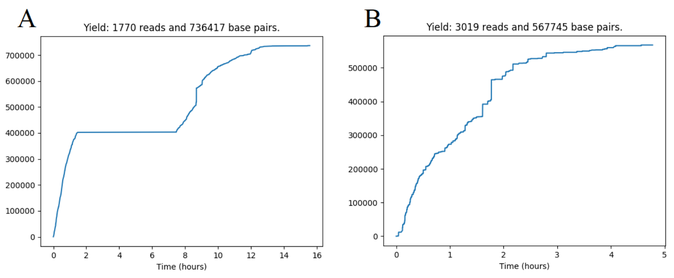

Figure 3: `yield_plot` in `poretools`: `BP_v1` (A) & `BP_v2` (B) - `x-axis`: time (hours); `y-axis`: total base pairs. Sections show large vertical jumps (sudden production of base pairs): these were high in `A` & `T` (repetitive reads).

The team left the mine at 50 minutes; `BP_v1` was paused for 6 hours. `BP_v2` data was affected during the return journey: the elevator trip back to the surface and the car (breaks, hill bumps).

Goldilocks

`Goldilocks` [3] was developed in Aug-2014, last updated Jul-2016 — to locate *"interesting* regions" on a genome that are "*just right*" for some user-provided criteria.

ACGT

The *'building blocks of DNA'*.

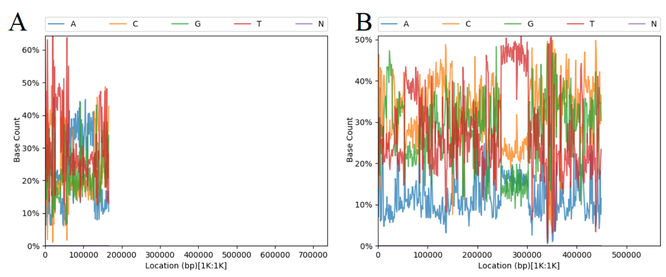

Figure 4: `ACGT` plot of `NucleotideCounterStrategy` content:- `x-axis`: the location along the genome; `y-axis`: percentage of content cover. (A) is `BP_v1` & (B) is `BP_v2`. In cardinal order (no order): read as data position in file, (time order).

Both data sets have a high number of `T`s. `Goldilocks' g.query` was used, and we found that the `T` and `A` heavy reads were the longest reads in the data sets, moreover these reads were very low quality.

FastQC

`FastQC` [4] was released Apr-2010 (recent update Mar-2016) and provides basic statistics of the data-sets.
`FastQC`'s graphs limit the data - the `x-axis` is scaled; whereas `Goldilocks` produces linear graphs.

Quality

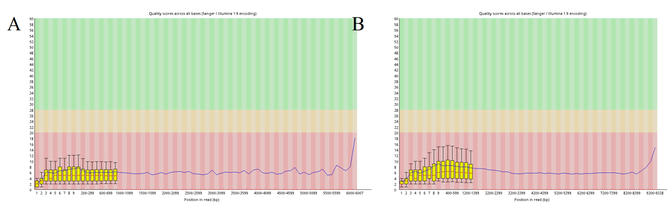

Figure 5: `FastQC` quality on data-sets limited to 10,000 bp:- `x-axis`: position in read; `y-axis`: quality score - `BP_v1` (A) & `BP_v2` (B).

ACGT

`GC` ratio of `BP_v1` is 51%, whilst `BP_v2` is 60%.

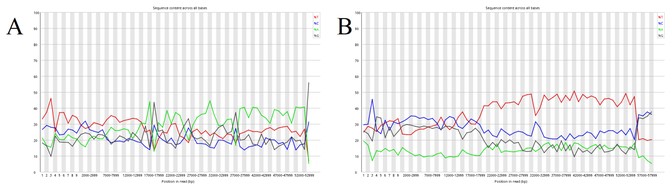

Figure 6: `ACGT` plotted on `FastQC`:- `x-axis`: position in read; `y-axis`: content score (ratio/percentage). (A) is `BP_v1` & (B) is `BP_v2`.

When we plot `ACGT` with `FastQC` and compare to `Goldilocks' ACGT` plot, we can see visible differences. If we compare `ACGT` plots of `BP_v2` we see that `T` steadily increases for `FastQC` while it varies in detail `Goldilocks` - this is due to the scaling performed by `FastQC` which makes it unsuitable for this purpose.

BLAST

`BLAST` (*Basic Local Alignment Search Tool*) is an algorithm for comparing DNA sequence similarity. I analysed the alignment lengths aiming for results within the hundreds, bit-score (high bit-scores mean better sequence similarity), and percentage match identity.

BP_v1 had poor results - unfortunately the highest result of alignment length was 49; highest bit-score was 67.6; though 100% identical matches but this is because the reads are short. **Species found**:

- Capsicum annuum: Sweet and chili peppers (plant)
- Pygocentrus nattereri: Red-bellied piranha (animal)

BP_v2 had better results ranging from 300 to 900 in alignment score with 74.941% average for the top 5 highest; bit-scores were as high as 440; and percentage of identical matches varied due to longer reads. **Bacteria found**:

- Neorhizobium galegae: bacteria that forms nitrogen-fixing root nodules
- Nitrosomonas: bacteria that oxidizes ammonia into nitrite as a metabolic process; found in nitrogen rich areas
- Rhodoplanes: bacteria organisms that carry out photosynthesis

Conclusion

I aimed to demonstrate more software, however the `BLAST` jobs alone took many hours and another piece of software couldn't run as it didn't have enough RAM on the IBERS compute cluster to scale to the input data.

We come away from this project understanding:

- Multiple bioinformatics tools are not yet suitable for long-reads
- Can acknowledge the early state of Nanopore sequencing itself - it is still biologically difficult to handle DNA & keep it intact for sequencing
- Nanopore quality is quite low even in good data: with papers showing phred scores of 10.53 [5] (1 in 10 error)
- When running the DNA through the Nanopore, we discover that if the sequencer isn't steady then results are affected greatly

Acknowledgements

Amanda Clare — afc@aber.ac.uk — @afcaber
Andre Soares — ans74@aber.ac.uk — @GeoMicroSoares
Sam Nicholls — msn@aber.ac.uk — @samstudio8

References

[1] A. Edwards, A. Soares, S. Rassner, P. Green, J. Felix, & A. Mitchell. *Deep Sequencing: Intra-Terrestrial Metagenomics Illustrates The Potential Of Off-Grid Nanopore DNA Sequencing*, 2017. bioRxiv: 133413.

[2] N. Loman, & A. Quinlan. *poretools: a toolkit for working with nanopore sequencing data from Oxford Nanopore*, Bioinformatics Vol.30 2014. pages: 3399-3401.

[3] S. Nicholls, A. Clare, & J. Randall. *Goldilocks: a tool for identifying genomic regions that are just right*, Bioinformatics Vol.32 2016. pages: 2047-2049.

[4] S. Andrews. *A quality control tool for high throughput sequence data.*, 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[5] S. Nicholls, W. Aubrey, A. Edwards, K. de. Grave, S. Huws, L. Schietgat, A. Soares, C. Creevey, & A. Clare. *Computational haplotype recovery and long-read validation identifies novel isoforms of industrially relevant enzymes from natural microbial communities*, 2018. bioRxiv: 223404.