An analysis of current software for nanopore metagenomic data

Current state of the art software

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

Our insight into DNA is controlled through 'sequencing'. Until recently, it was only possible to sequence DNA into short strings called 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 contents of the microbial community. We are interested to discover how well current bioinformatics software works with this new long read data and to try out some recent new developments for such analysis.

Introduction

Using various software, we want to analyse the FAST5 data to observe sequence similarities and discover what bacteria resides within the mine: Two data sets were collected from the mine expeditions [1]: **BP_v1** (Dec-2016) & **BP_v2** with improved protocol (Apr-2017).

- Numer of reads:- BP_v1: 1,770; BP_v2: 3,019.

Method

Tested the following software with the datasets on the IBERS cluster.

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

BP_v1 (A) has more short reads & BP_v2 (B) has more long reads - we limited to 10,000 base pairs due to the long reads being very low in quality.

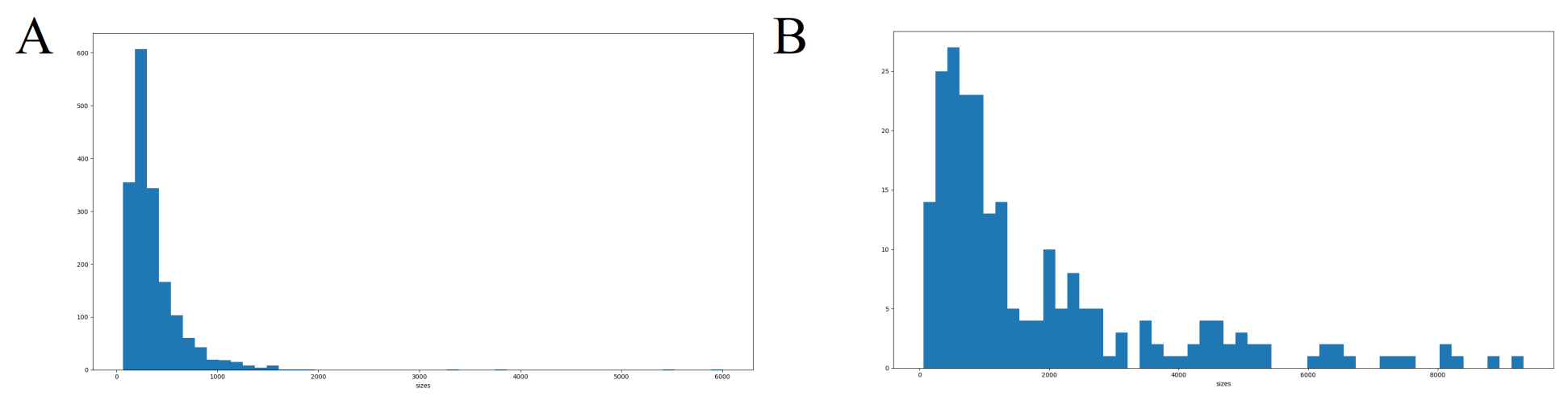

Figure 1: Histogram comparison of data-sets limited to 10,000 bp:- x-axis: read length (size); y-axis: cumulative frequency (count).

Quality

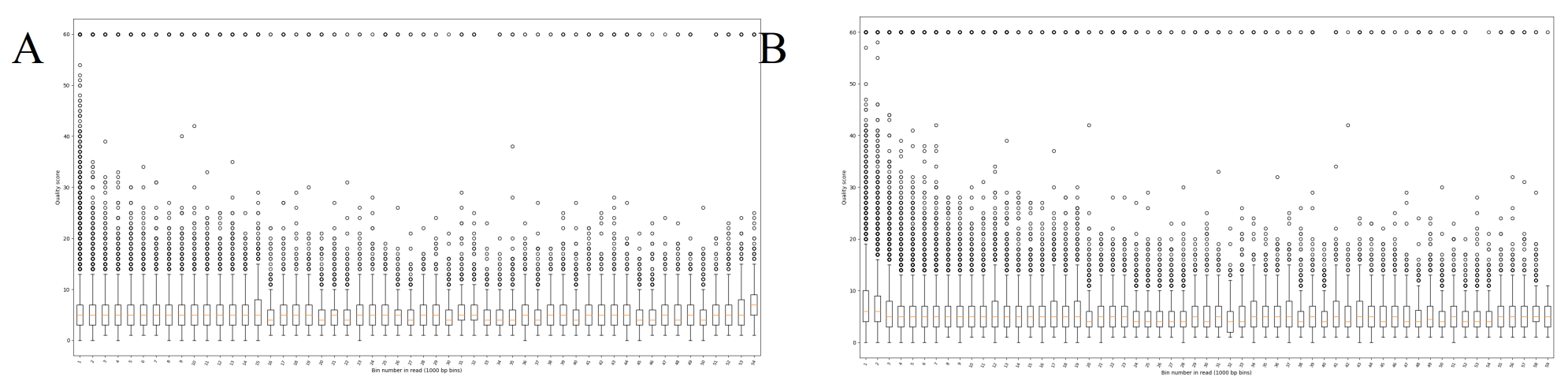

Figure 2: poretools' qualpos comparison - BP_v1 is (A) and BP_v2 is (B). The unusual high score of 60 could be an error with the software or a random outlier.

Analysing poretools qualdist (summary quality scores), we can conclude the data is poor - % is a specific symbol that relates to bad quality and both data sets were high in this symbol; BP_v1: 116,956 & BP_v2: 81,437.

Time

The research team left the mine at 50 minutes; BP_v2 ran throughout whilst BP_v1 was paused for 6 hours.
There are sections that show large vertical jumps (sudden production of base pairs): after studying, these are sections which were high in A & T (repetitive reads). BP_v2 data was affected during the return journey: the elevator trip back to the surface and the car journey (breaks, bumps, and going up/down a hill).

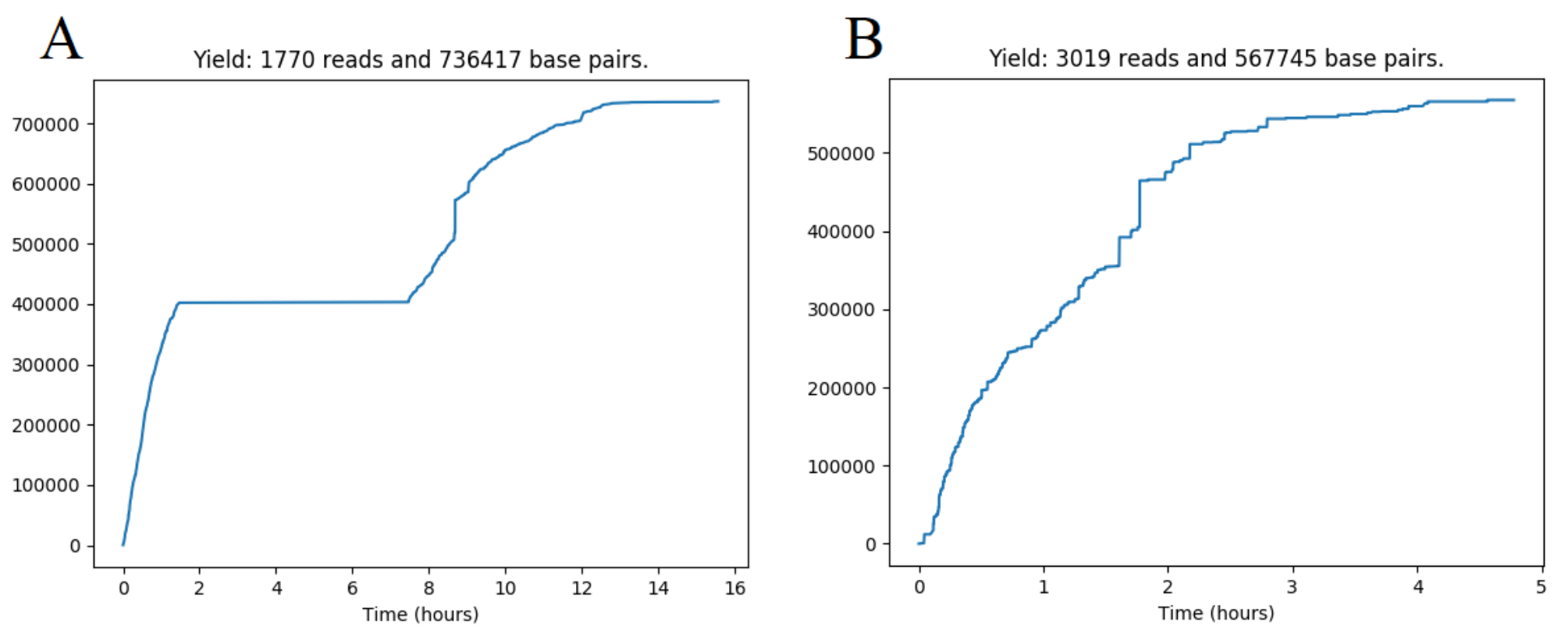

Figure 3: yield_plot in poretools. (A) is BP_v1 & (B) is BP_v2 - x-axis: time (hours); y-axis: total base pairs.

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

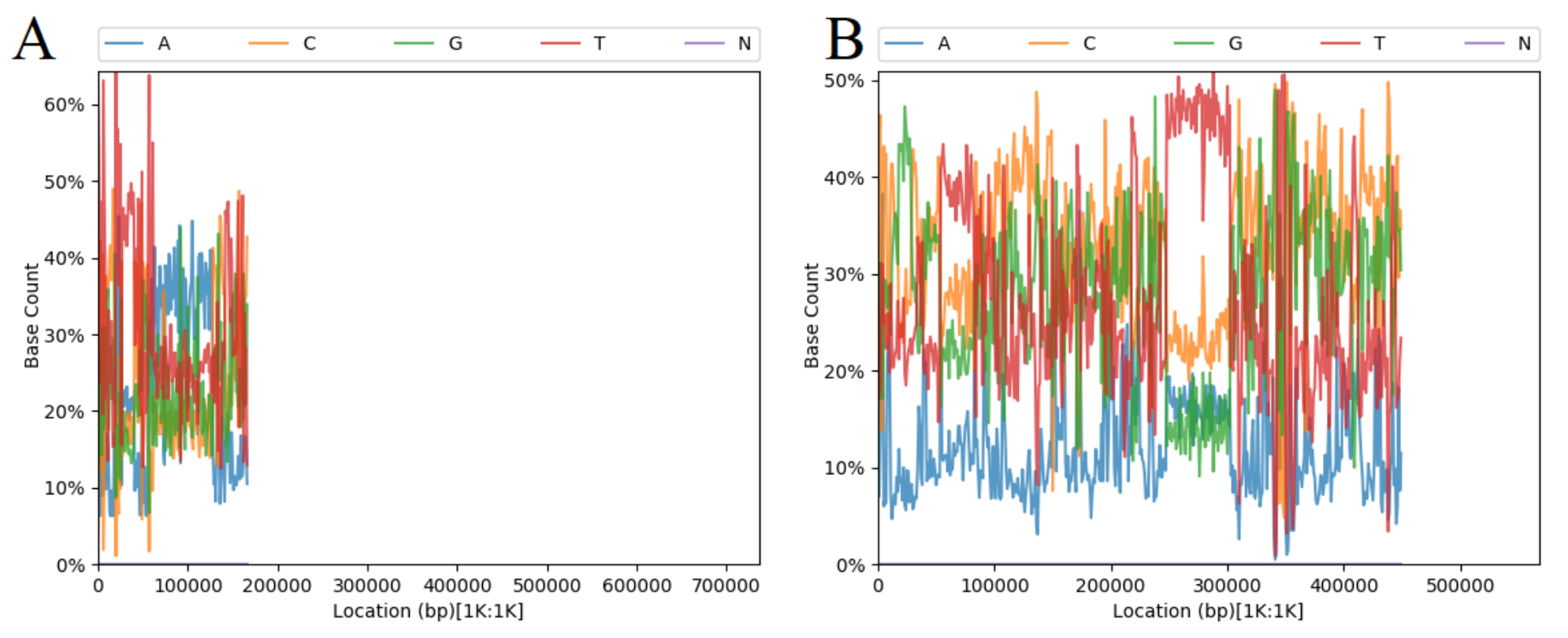

Figure 4: ACGT plot of NucleotideCounterStrategy content:- x-axis: the location along the genome; y-axis: percentage of content cover. (A) is BP_v1 & (B) is BP_v2. In cardinal order (no order): read as data position in file, (time order).

Both data sets have a high number of Ts with BP_v1 having an unusual high ratio of As. Goldilocks' g.query was used to find the reads with the most Ts in BP_v2 & most As in BP_v1 - we found that the T and A heavy reads were the longest reads in the data-sets. Moreover, we then looked into the top longest reads for both data sets, and performed quality checks on them with (poretools & FastQC).

FastQC

FastQC [4] was released Apr-2010 (most recent update Mar-2016) and provides basic statistics of the data-sets. FastQC's graphs weren't useful as the x-axis is uniform but stretched; whereas Goldilocks produces linear graphs: we want to observe read positions in linear form. Moreover, FastQC graphs do not display the whole data; they are, unknown, limited.

Read Length

BP_v1 has 1761 total sequences, whilst BP_v2 has 236; backing up our previous find: poretools' histograms show that BP_v1 has more short reads.

Quality

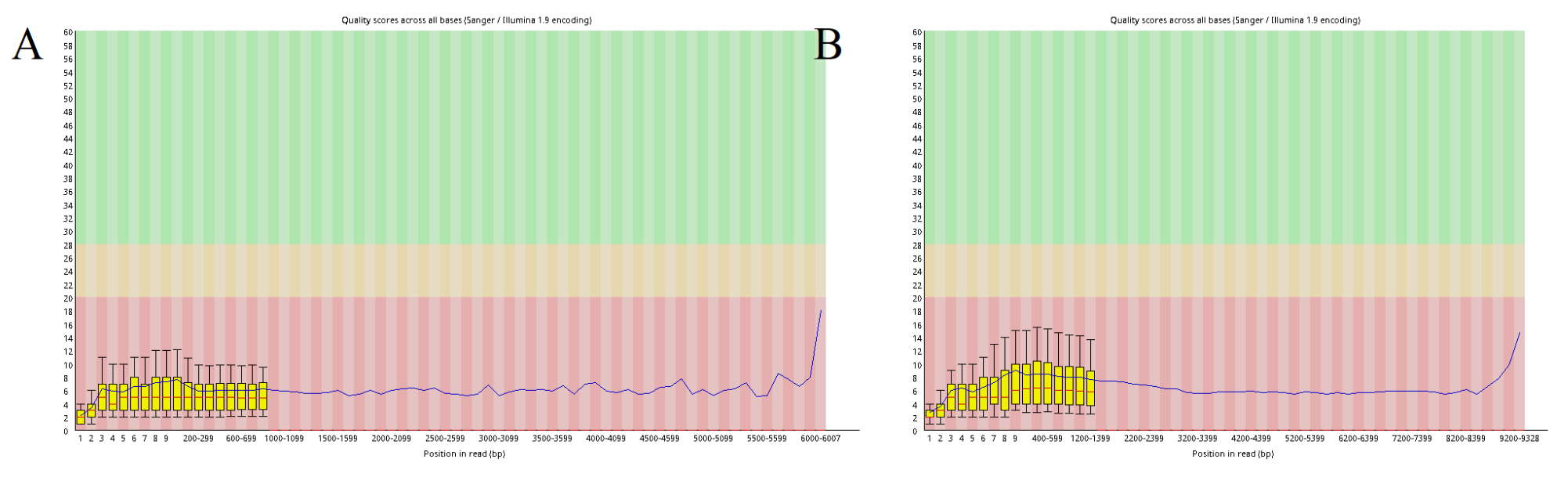

Figure 5: FastQC quality on data-sets limited to 10,000 bp:- x-axis: position in read; y-axis: quality score. (A) is BP_v1 & (B) is BP_v2.

The quality can be compared to poretools' qualpos box plots and we observe both are low.

ACGT

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

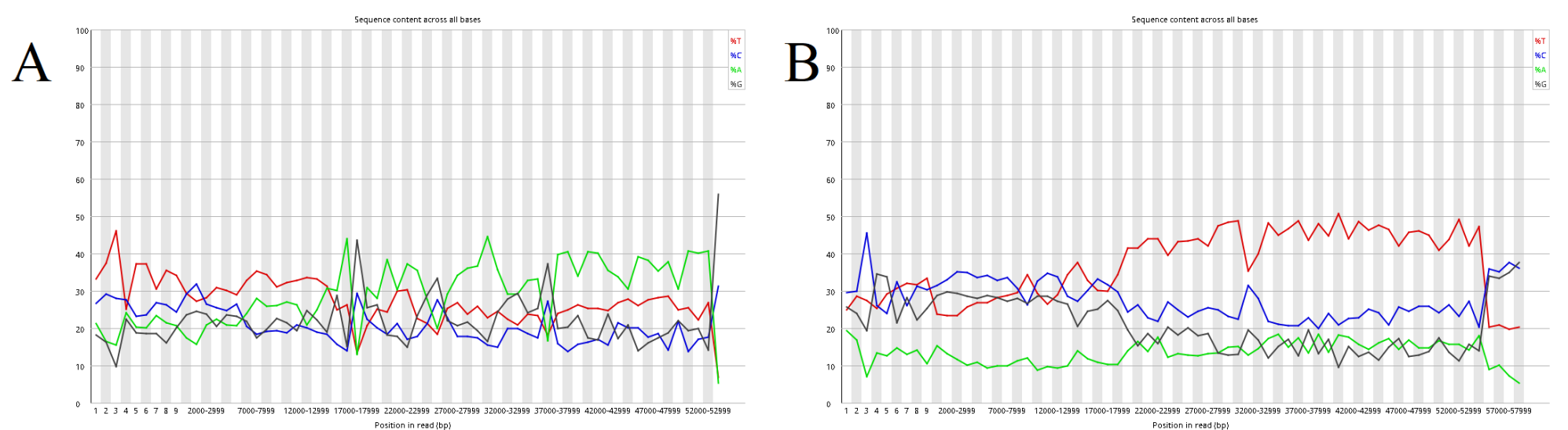

Figure 6: ACGT plotted on FastQC:- x-axis: position in read; y-axis: content score (ratio/percentage). (A) is BP_v1 & (B) is BP_v2.

When we plot ACGT with FastQC and compare to Goldilocks' ACGT plot, we can see visible differences. If we compare BP_v2 (B): Goldilocks' ACGT and FastQC's ACGT, you can see that T steadily increases for FastQC while it varies in detail Goldilocks. This is due to the scaling performed by FastQC.

BLAST

BLAST (*Basic Local Alignment Search Tool*) is an algorithm for comparing DNA sequence similarity. We used blastn with the NCBI database. 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; bit-scores were as high as 440; and percentage of identical matches varied due to longer reads, for the top 5 highest alignment scores, the average result was 74.941%.
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

The more we see similarities in results from the different software, the more the can rely on them in future: we can use the trusted software as a comparison when new software is released. On the other hand, where there are differences it is difficult to know whether or not which software is more accurate - plus we need remember that if software is faster than another, we shouldn't assume it's precise - some of our BLAST jobs took many hours, another piece of software we tried couldn't run as it didn't have enough RAM on the SCRATCH space (server). From the research, some software works well with metagenomic data - its features can be used to make some observations; we have come to understand how nanopore quality is quite low: with papers showing phred scores of 10.53 [5] (1 in 10 error). When running the DNA through the nanopore, we can see observe 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.