An analysis of current software for nanopore metagenomic data

Samantha Pendleton

Aberystwyth University, Department of Computer Science

sap21@aber.ac.uk — @sap218

Abstract

Nanopore is new technology for DNA samples to produce long read DNA sequences. A research team from IBERS (*Institute of Biological, Environmental & Rural Sciences*) at Aberystwyth University have sampled metagenomes from a coal mine in South Wales using the Nanopore MinION, and given initial taxonomic summaries of the contents of the microbial communities. We are interested to discover how well current bioinformatics software works with long read data and to try out some recent new developments (from Aberystwyth University and elsewhere) for such analysis.

Introduction

- We will be looking into:
  - ACGT content
  - Quality
  - Time
- We will analyse the data to observe sequence similarities and find what bacteria resides within the mine.
- Two data sets were concluded from the mine expeditions [1]: **BP_v1** (Dec-2016) & **BP_v2** (Apr-2017),
  - BP_v2 extraction was an improved protocol.

Method

We downloaded/installed various software and ran them on the data-sets on the IBERS cluster server.

Software used to look into ACGT, quality, and time; plus sequence similarities:

- poretools
- Goldilocks
- FastQC
- BLAST

poretools

Toolkit for analysing nanopore sequence data [2]. poretools was developed in Aug-2014 though with 77 issues as of Sep-2017 on Github - which explains errors and 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 long reads being very low in quality.

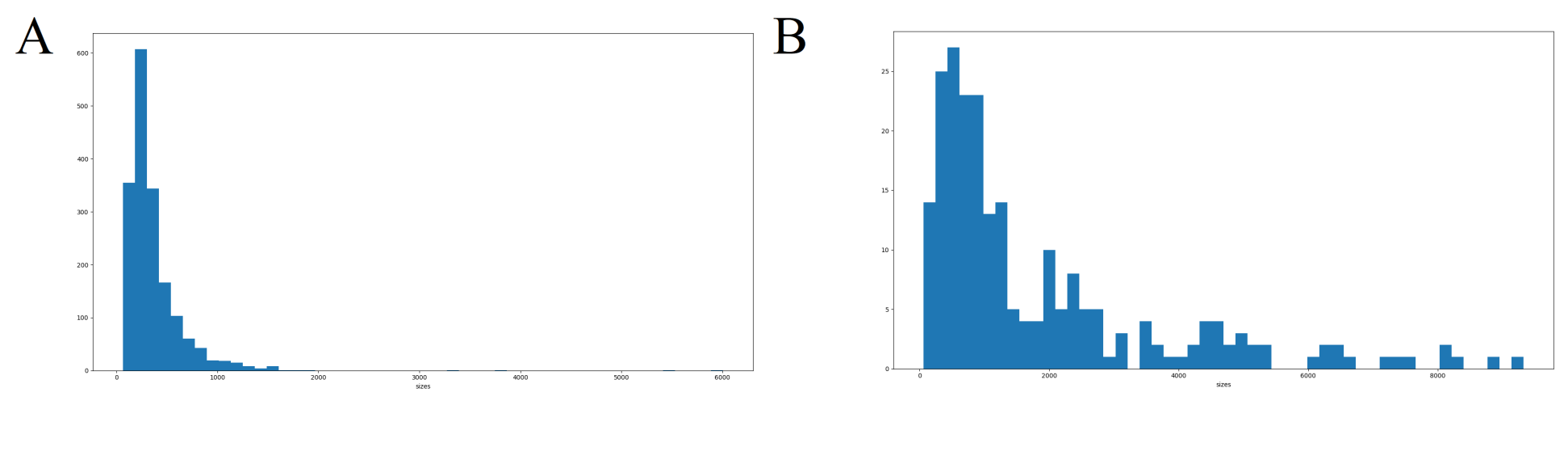

Figure 1: Histogram Comparison of both data-sets limited to 10,000 bp:- x-axis: read length (size); y-axis: cumulative frequency (count).

These histograms show the read lengths, they were recreated from the ones used in the IBERS pre-printed paper on bioRxiv (133413).

Quality

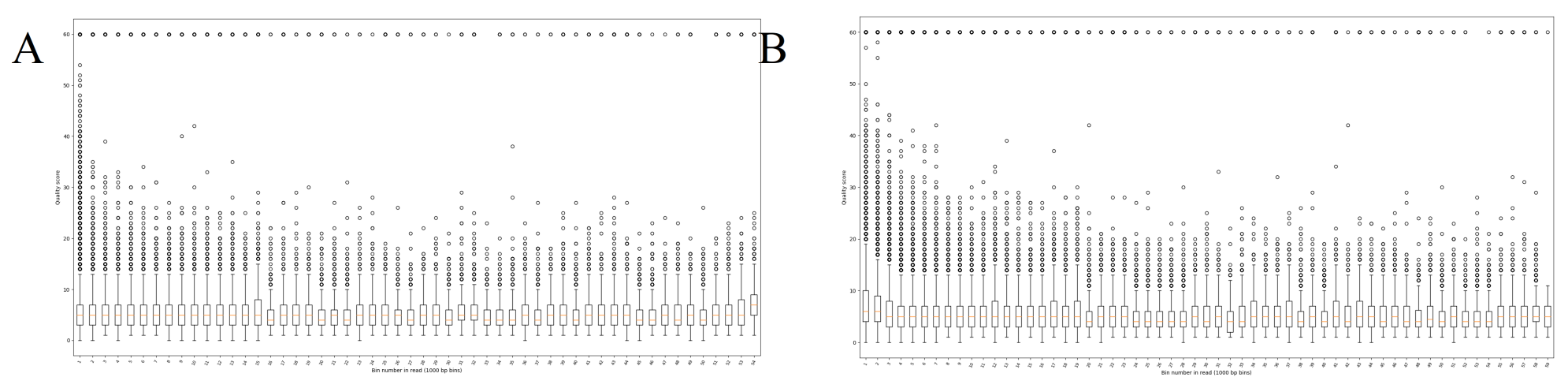

Figure 2: poretools' qualpos comparison - BP_v1 is (A) and BP_v2 is (B). The unusual high score of 60 could be an error with the software.

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 are high in this symbol; BP_v1: 116,956 & BP_v2: 81,437.

Time

The research team left the mine at 50 minutes: BP_v1 was paused for 6 hours then continued to run after - BP_v2 continued to run during the return journey.

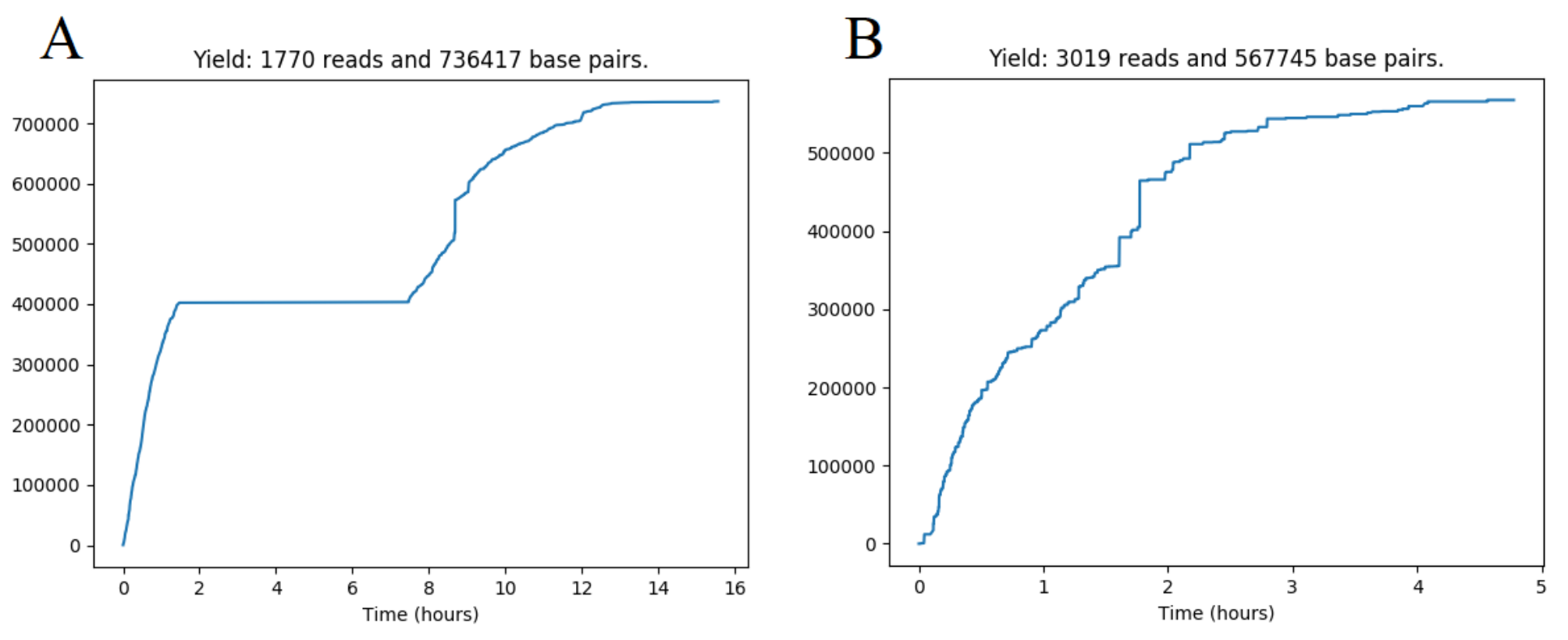

Figure 3: yield_plot in poretools. (A) is BP_v1 & (B) is BP_v2 - x-axis: time (hours); y-axis: total base pairs.

There are sections that show large vertical jumps (suddenly produces a lot of base pairs) in the graphs: after analysing, these are the sections which were high in A & T (repetitive reads). Specifically BP_v2 the DNA was affected after they left the mine: during the elevator trip back to the surface and the car journey (breaks, bumps, and going up/down a hill).

Goldilocks

Goldilocks [3] was developed in Aug-2014 with last update Jul-2016; quickly locate *interesting* regions on the human genome that expressed a desired level of variability, which were *just right* for later variant calling and comparison.

ACGT

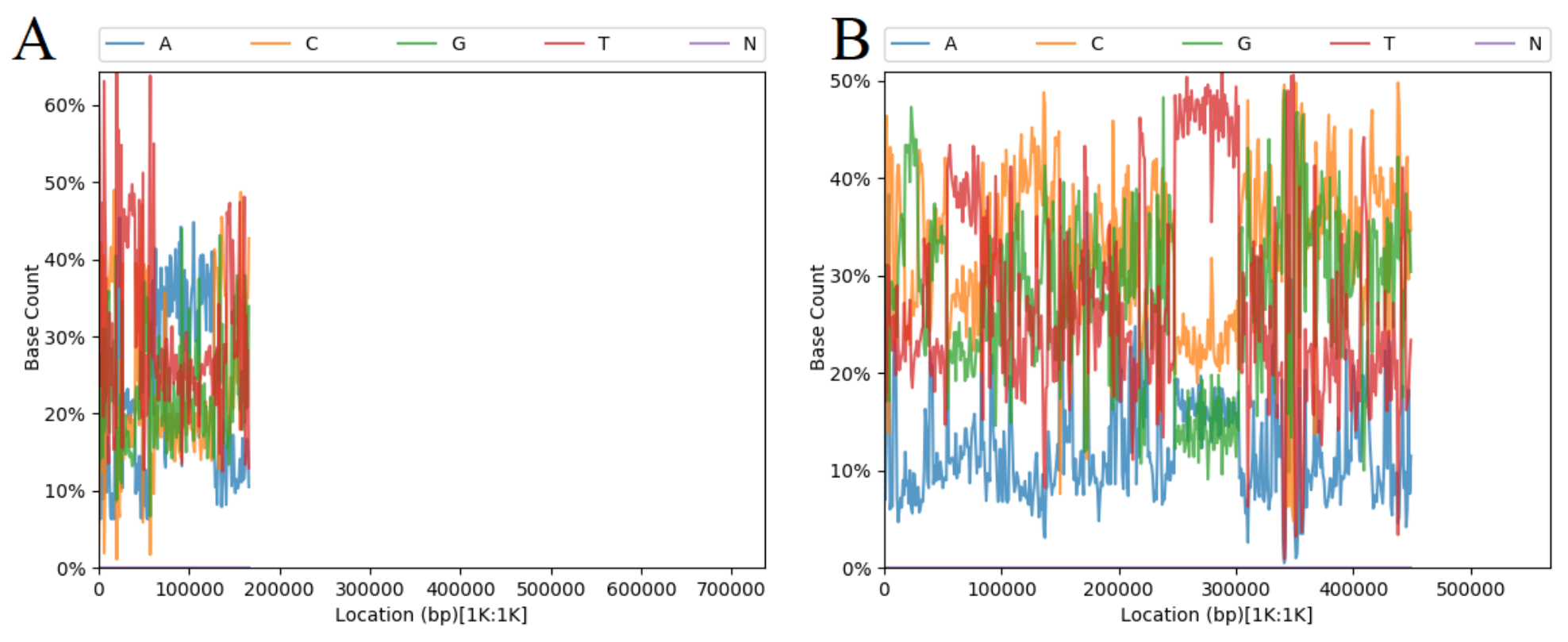

Figure 4: ACGT Plot of NucleotideCounterStrategy content using Goldilocks:- x-axis: the location along the genome; y-axis: percentage of content cover. (A) is BP_v1 & (B) is BP_v2. In cardinal order (no specific order): read as data position in file, which is most likely time order.

Both data sets have a high number of Ts with BP_v1 having an unusual high ratio of As.
Goldilocks's 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 datasets. Moreover, we then looked into the top longest reads for both data sets, and performed quality checks on them (poretools & FastQC).

FastQC

FastQC [4] was released Apr-2010 (most recent update Mar-2016) and provides basic statistics of the data-sets: BP_v1 total sequences = 1761, whilst BP_v2 = 236 — this backs up our previous find regarding read length: poretools' histograms show that BP_v1 has more short reads.

FastQC has produced graphs that are not useful as the x-axis is uniform but stretched; whereas Goldilocks produces linear graphs: we want to observe read positions in linear form. Plus, FastQC graphs do not display the whole data-sets, they are (unknown why) limited.

Quality

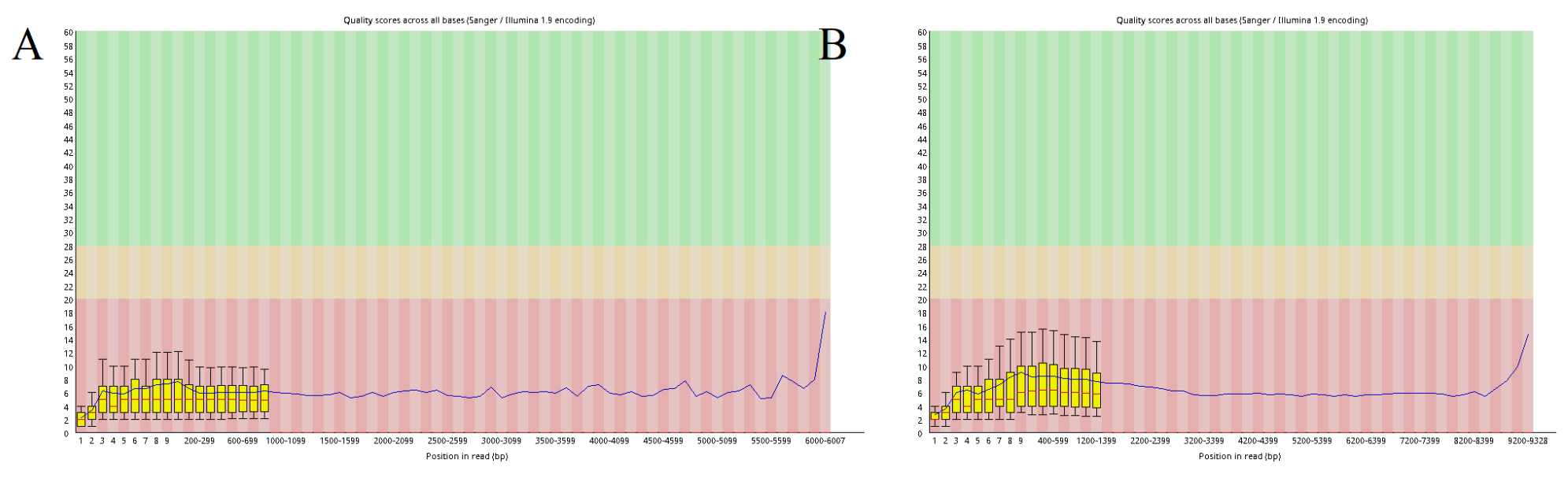

Figure 5: FastQC quality on both data-sets limited to 10,000 bp:- x-axis: position in read; y-axis: quality score. (A) is BP_v1 & (B) is BP_v2.

The quality can be compared to poretools' qualpos box plots and both are low. However some single reads of BP_v2 were plotted in the medium range.

ACGT

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

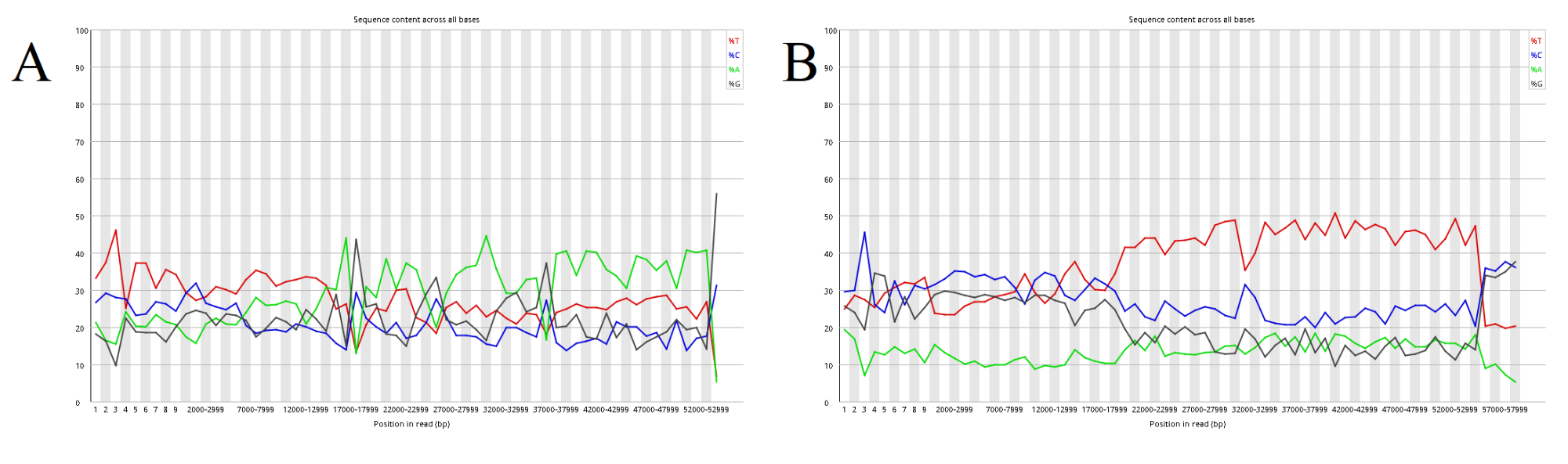

Figure 6: ACGT plotted on FastQC:- x-axis: position in read; y-axis: content score (ratio/percentage). (A) is BP_v1 & (B) is BP_v2.

When we plot ACGT with FastQC and compare to Goldilocks' ACGT plot, we can see visible differences: the results aren't similar. If we compare BP_v2: (B) from Goldilocks' ACGT and (B) from FastQC's ACGT, you can see that T steadily increases for FastQC whilst it varies more in Goldilocks: specifically peaks midway then lowers again.
This is due to the truncated x-axis and it's scaling.

BLAST

Blast [5] for *Basic Local Alignment Search Tool* is an algorithm for comparing sequence information: sequence similarities. We used BLASTn on both data-sets and searched the results (taxon IDs) in the NCBI database.

- alignment lengths: result within hundreds
- bit-score: high bit-scores = better sequence similarity
- percentage of identical matches

BP_v1 had poor results - unfortunately the highest result of alignment length was 49; highest bit-score was 67.6; and mostly 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 areas that contains high levels of nitrogen compounds
- 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 consider the premises that if a software is faster than another, is it more reliable/precise?

Acknowledgements

Amanda Clare[1] - afc@aber.ac.uk
Andre Soares[2] - ans74@aber.ac.uk
Sam Nicholls[3] - msn@aber.ac.uk

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.*, 2015. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[5] NCBI. *BLAST*, U.S. National Library of Medicine, 1990.

[1] Department of Computer Science
[2] IBERS (*Institute of Biological, Environmental & Rural Sciences*)
[3] Dept. Computer Science & IBERS