An analysis of current state of the art software on nanopore metagenomic data

Samantha Pendleton
Department of Computer Science
Aberystwyth University
sap21@aber.ac.uk

Abstract
Nanopore is new technology for DNA samples to produce long read DNA sequences. A research team from Aberystwyth University have sampled metagenomes using the Nanopore MinION from a coal mine, 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.

Contents
1 Introduction 2
2 Method 5
   2.1 poretools ................................................................. 5
   2.1.1 Quality ................................................................. 5
   2.1.2 Time ................................................................. 7
   2.2 Goldilocks ............................................................... 8
   2.2.1 GC ................................................................. 8
   2.2.2 ACGT Content ....................................................... 9
   2.3 FastQC ............................................................ 10
   2.3.1 GC ............................................................. 10
   2.3.2 Quality ......................................................... 11
1. Introduction

A wide variety of software is available today for the analysis of nanopore metagenomic data. DNA extracted from a coal mine located in South Wales, where no internet access is available, was run through a the MinION\(^1\) and then analysed in a pre-printed paper\([1]\) - results told us of the wide variety of bacteria which reside in the subsurface.

Two data sets were concluded from the mine expeditions (same mine in South Wales) - extracted 4 samples for BP\(_v1\) generated December 2016. 8 samples extracted for BP\(_v2\), generated April 2017.

Version 2 is better quality: the researcher had optimised DNA extraction to improve recovery yields from this kind of environments by using an improved protocol\(^2\).

Moreover BP\(_v2\) was not quality filtered; it had MUX genes (QC reads), which were later removed for analysis - BP\(_v1\) has a total of 1,770 fast5 files & BP\(_v2\) has a total of 3,019.

As seen in Figure 1, you can observe how BP\(_v2\) (Figure 1b) generated consistently longer reads compared to BP\(_v1\) (Figure 1a) - BP\(_v1\) has more short reads:-

\(^1\)https://nanoporetech.com/products/minion

Figure 1: Histogram Comparison of both data-sets: x-axis: read length (size); y-axis: cumulative frequency (count)
After some quality tests, which will be explained later, it is observed that the longest reads were low in quality and most likely errors, so here are the histograms limited to 10,000 base pairs (bp) - as seen in Figure 2. We can see that BP_v2 (Figure 2b), even without the top 5 longest reads, it still results in longer reads compared to BP_v1 (Figure 2a).

![Histogram Comparison](image)

(a) BP_v1

(b) BP_v2

Figure 2: Histogram Comparison with both limited to 10,000 bp: x-axis: read length (size); y-axis: cumulative frequency (count)
2. Method

Further, in-depth analysis of DNA data-set will be conducted in this report, using a variety of software and methods/tools (looking into sequence similarity).

2.1. poretools

poretools[2] is a toolkit for analysing nanopore sequence data - histograms were produced to observe read length (as seen with Figures 1 & 2).

2.1.1. Quality

The quality overall averaged low - using poretools’ qualpos, these box plots would originally be useful however there’s an unusual number of plots on the 60: poretools was developed in 2014 with hardly any updates: release history\(^3\) v0.6.0 (29-Aug-2016), so this error could be a result of a software that’s not had issues fixed (as of 15-Sep-2017 77 issues\(^4\)).

![poretools comparison](image)

(a) BP_v1

(b) BP_v2

Figure 3: poretools qualpos comparison: x-axis: bin number in read; y-axis: quality score

---

3https://github.com/arq5x/poretools/blob/master/docs/content/history.rst

4https://github.com/arq5x/poretools/issues
Using poretools’ qualdist (summary quality scores), analysing the lack of count for letters: we can conclude that both data-sets are poor.

![Figure 4: poretools qualdist comparison](image)

Note: Letters are what we are aiming for: these are good quality scores. % is a specific symbol that relates to bad quality and as we can see, both data sets are high in this symbol; count of % symbol in BP_v1: 116,956 & BP_v2: 81,437.

We can see that BP_v1 has a higher count of bad symbols giving more evidence that BP_v1 is lower quality than BP_v2.
2.1.2. Time

Using *poretools’ time* and *yield_plot* (Figure 5 - plots are base pairs over time) features, we were able to see what reads were taken and when, plus produce time graphs that we can use to observe the base pairs collected throughout the time span.

![Graphs showing time evolution of base pairs](image)

(a) BP.v1  
(b) BP.v2

Figure 5: Time graph using *poretools’ yield_plot*

Note: research team left the mine after 50 minutes for both data sets.

**BP.v1** differs from **BP.v2** - Figure 5a shows that there is an increase of base pairs as time increases, as we would expect. However, **BP.v1** was paused for 6 hours after 50 minutes, then continued once a researcher arrived home. We can see at 9 hours there was a sudden long peak: reduced/no yield in DNA - this could potentially be that long A heavy read, which coincidental is the longest read of the data set.

In **BP.v2** - Figure 5b - the top 5 longest reads were taken after 50 minutes - which can conclude why there was data error. After leaving the mine (50+ mins), the team continued to run the analysis: a researcher from the team stated how 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).

However, results showed that the top 5 T heavy reads of **BP.v2** were not the same as the top 5 longest reads. Moreover, the top 5 T heavy reads were majority after the first hour too, despite one read that was actually within the 50 minutes; after analysing this read we see that despite high in T, it actually has bacteria and fungus, but also plant and animal.

The results are quite unique: both data-sets seem to have collected 400,000 base pairs within two hours - **BP.v1** collected 736,417 base pairs over 10 hours (16 hours total w/ a 6 hour pause) and **BP.v2** collected 567,745 in 5 hours, yet is better quality. Moreover, **BP.v1**’s base pairs were a lot of AT sequences.
2.2. Goldilocks

Goldilocks[3] was developed and used to "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."

Goldilocks produces Linear graphs and the reads are cardinal order (no specific order): read as data position in file, which is most likely time order.

2.2.1. GC

GC in DNA analysis tells us about the stability: DNA with low GC content is less stable. Majority of researchers look for higher ratio in GC compared to AT.

Figure 6 shows the comparison of BP_v1 & BP_v2. As seen in Figure 6a, BP_v1 is quite scattered so further explanation will be concluded below. Furthermore, Figure 6b shows us that BP_v2 has mixed GC content with an unusual section of low GC content (300,000) but due to being better quality, we can analyse this data set.

![Figure 6: GC ratio using Goldilocks: x-axis: the location along the genome; y-axis: the value of the censused region](image-url)

(a) BP_v1
(b) BP_v2

Figure 6: GC ratio using Goldilocks: x-axis: the location along the genome; y-axis: the value of the censused region
2.2.2. ACGT Content

Using Goldilocks’ NucleotideCounterStrategy, we produced graphs in order to display the ACGT ratio/percentage - see Figure 7. Both data sets have a high number of Ts. Moreover, BP_v1 again has a high number of Ts throughout however an unusual A count (Figure 7a). If you observe Figure 8, the GC content colour has been changed (to black) so the AT is a more more visible.

![Figure 7: ACGT Plot of NucleotideCounterStrategy content using Goldilocks:- x-axis: the location along the genome; y-axis: percentage of content cover](image)

![Figure 8: AT Plot of BP_v1 & BP_v2](image)

Goldilocks’s g.query was used to find the reads with the most Ts in BP_v2 & most As in BP_v1 to which we then used the following command `head -n10 BP_v2.fasta.fai` in a Linux terminal to observe the top longest reads for both data sets, we did a quality check on both: the longest reads were poor in quality (to be explained further).

Note: SAMtools[4] was used to index fasta files so we could use them with Goldilocks, plus to convert files to fastq (we also used poretools).
2.3. FastQC

FastQC\cite{5} is quite the useful tool: to start it provides basic statistics of the data-sets; BP\_v1 total sequences = 1761, whilst BP\_v2 total sequences = 236 - this backs up the statement plus poretools’ histograms that BP\_v1 has more short reads.

2.3.1. GC

The GC ratio of BP\_v1 is 51\%, whilst BP\_v2 is 60\%.

![Figure 9](image.png)

Figure 9: Plot of GC content in FastQC- x-axis: mean content (%); y-axis: GC count per read

Comparing GC in FastQC \& Goldilocks may be quite difficult here due to difference of graphs: FastQC creates line whilst Goldilocks creates scatter, as seen in Figure 6.
2.3.2. Quality

Quality tests were conducted using FastQC. As you can see in Figure 10 both BP_v1 & BP_v2 are as a whole quite poor in quality. Comparing this tool with poretools’ qualpos, I think this tool is useful as the box plots are presented much clearer plus the colour scheme is very useful: we can see the data subsides in the red area - however the $x$-axis is uniform but stretched, at first were misunderstood as logarithmic graphs; in these instances we prefer Goldilocks due to their graphs being linear.

Figure 10: Plot both data-sets reads in FastQC quality graph
Figure 11 shows us FastQC on both data-sets limited to 10,000 - BP.v2’s highest values on the box plots almost reach the medium/decent quality range.

Figure 11: Plot of whole data-sets limited to 10,000 in FastQC quality graph
2.3.3. ACGT Content

Figure 12: Plot of ACGT content in FastQC graph

(a) BP_v1

(b) BP_v2

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>color</td>
<td>green</td>
<td>blue</td>
<td>black</td>
<td>red</td>
</tr>
</tbody>
</table>
2.4. **BLAST**

BLAST for Basic Local Alignment Search Tool; an algorithm for comparing primary biological sequence information: the nucleotides of DNA sequences.

Note: links in this section may have expired.

**Blasting longest reads**

Further tests included using nucleotide BLAST[6] (blastn) with discontiguous MEGABLAST on the longest reads of both data sets to prove the quality was low: the query cover was 0% of tiny hits, under 100 base pairs (bp) - see Figure 13.

![Figure 13](image)

*Figure 13: Longest reads Blasted*

**BP_v1** (Figure 13a):
Species: Animal (Fish: Japanese rice fish & Zebrafish), Insects (Fruit flies).
Dictyostelium discoideum: species of soil-living amoeba - commonly referred to as slime mold.

**BP_v2** (Figure 13b):
Species: Animal (Mouse & Tapeworm), Plant (Tomato & Kiwi), Insects (Fruit flies), Human.
Parasites (including flatworms & malaria), Fungus
**Blasting random reads**

Due to the long reads being low in quality, further tests on the data sets are to be conducted: specifically blasting random, shorter, reads.

I extracted numerous shorter, random reads from BP_v1 data set and Blasted them: yet no results/similarities came up - the reads were mostly: TATATA content: very short and repetitive. As a whole, BP_v1 didn’t have anything useful when blasted.

After extracting a random read from BP_v2⁵ (see Figure 14) - we can observe this data read was bacteria (filamentous thermophilic bacteria⁶) - Figure 14b. Figure 14a shows the results: there was a query cover of 17%.

---

\[\text{channel}_349\text{bf5a3f39-3e35-4870-82ec-a7f9679d4000}_{-}\text{qscore}9{}_{-}\text{read_score}^{-1.8}\]

---

(b) Alignment: Anaerolinea thermophila UNI-1 DNA, complete genome

Figure 14: Blast of a random read from BP_v2

Anaerolinea thermophila is thermophilic bacteria which thrives in high temperatures around 41-122°C - consequently, we can see that this result is quite unusual: the mine was not significantly a high temperature (researcher stated mine was 15-20°C).

---

For further analysis, we extracted another random read from BP.v2⁷ (ID: 164 - Figure 15). Despite no actually good hit (as seen in Figure 15b), E coli was a reoccurring factor. Figure 15a shows the highest result of E coli: there was a query cover of 1% commonly.

![Blast Alignment Scores](https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Get&RID=V33NGY1B016)

(a) Blast Alignment Scores

(b) Alignment: *Escherichia coli* transposon Mu dl-R insertion site

> Figure 15: Blast of a random read from BP.v2

A transposon is expected; it is a sequence of DNA that can move to new positions within the genome: *jumping genes*. Furthermore, the alignment score seems to be 100: these are not genes however could be 10,000 matches of 1,000 genes.

---

Blasting whole data sets

Using AU IBERS cluster server, I was able to blast the data-sets as a whole, rather than just blasting long or random reads. The NCBI Blast website wasn’t successful (timeout due to the data-sets being large: CPU usage limit was exceeded) so I had to blast locally with the nt nucleotide collection/database\(^8\) (nr is the proteins collection).

So with the results, we are looking for long alignment lengths: we aimed for lengths within hundreds plus looked at bit-score: the higher the bit-score, the better the sequence similarity. We also looked at the percentage of identical matches; the shorter the read the higher the percentage.
We obtained a bunch of taxon IDs that we searched in NCBI\(^9\).

\(\text{BP}_v1\) didn’t produce any good results, unfortunately the highest result of alignment length was 49; highest bit-score was 67.6. \(\text{BP}_v1\) mostly had 100% of identical matches but this is because the reads are majority short.

- 49 - channel\(_{46}\)c4208b23-0984-4f08-adef-a196248bd603\(\text{qscore}\)4.5\(\text{read}\)score-3
  Anaeromyxobacter dehalogenans 2CP-C\(^{10}\); bacteria strain that efficiently reduces metals such as ferric iron, Fe(III), and oxidized uranium.
- 46 - channel\(_{425}\)9f125c87-0c2b-4ba9-a015-53897a1b3b77\(\text{qscore}\)4.6\(\text{read}\)score-3.3
  Capsicum annuum\(^{11}\): Sweet and chili peppers (plant)
- 39 - channel\(_{425}\)8d5bc381-5611-4884-ac32-8046adabe9d7\(\text{qscore}\)4.5\(\text{read}\)score-3.4
  Pygocentrus nattereri\(^{12}\): Red-bellied piranha (animal)

\(\text{BP}_v2\)’s results were much better: despite multiple low scores too, we had better results ranging from 300 to 900 in alignment score - bit-scores were as high as 440. \(\text{BP}_v2\) percentage of identical matches varied due to longer reads, for the top 5 highest alignment scores, the average result was 74.941%.

- 833 - channel\(_{470}\)c75b822d-29ef-45aa-a065-675768cbf973\(\text{qscore}\)9.5\(\text{read}\)score-1.6
  Neorhizobium galegae\(^{13}\): bacteria that forms nitrogen-fixing root nodules

\(^{10}\)https://www.ncbi.nlm.nih.gov/gquery/?term=CP000251.1
\(^{11}\)https://www.ncbi.nlm.nih.gov/gquery/?term=XM_016684803.1
\(^{12}\)https://www.ncbi.nlm.nih.gov/gquery/?term=XM_017706955.1
\(^{13}\)https://www.ncbi.nlm.nih.gov/nuccore/HG938353
• 674 - channel_427_cc4005f9-1542-4808-93e2-2eef88ba5a41_qscore_8.8_read_score_-1.8
  – Rhodoplanes\textsuperscript{14}: a phototrophic genus of bacteria - organisms that carry out photosynthesis
  – Azorhizobium caulinodans\textsuperscript{15}: species of bacteria that forms a nitrogen-fixing symbiosis with plants of the genus Sesbania (flowering plants in pea family)

• 652 - channel_43_592b8e8f-b9b2-46f7-831f-528f42467191_qscore_8_read_score_-1.9
  Sideroxydans lithotrophicus\textsuperscript{16}: autotrophic iron-oxidizing Gram-negative bacterium isolated from iron contaminated groundwater

• 548 - channel_470_c75b822d-29ef-45aa-a065-675768cbf973_qscore_9.5_read_score_-1.6
  Bordetella flabilis\textsuperscript{17}: recovered strains from cystic fibrosis (strains from human respiratory specimens)

• 330 - channel_185_cdc2062b-5bab-4330-8fde-1ec431156cc_qscore_8.3_read_score_-1.8
  Nitrosomonas\textsuperscript{18}: bacteria that oxidizes ammonia into nitrite as a metabolic process; found in soil, freshwater, and on building surfaces, especially in areas that contains high levels of nitrogen compounds

\textsuperscript{14}\url{https://www.ncbi.nlm.nih.gov/gquery/?term=CP007440.1}
\textsuperscript{15}\url{https://www.ncbi.nlm.nih.gov/nuccore/AP009384.1}
\textsuperscript{16}\url{https://www.ncbi.nlm.nih.gov/gquery/?term=CP001965.1}
\textsuperscript{17}\url{https://www.ncbi.nlm.nih.gov/gquery/?term=CP016172.1}
\textsuperscript{18}\url{https://www.ncbi.nlm.nih.gov/gquery/?term=CP002876.1}
3. Conclusion

Overall, poretools is handy but not up to date and has issues: specifically the box plots. Though it’s histograms and time graphs are very useful to observe.

Goldilocks, if provided more features, would be a very practical tool, specifically altering the ACGT graphs is very useful but can obviously be improved.

FastQC offers a variety, however not suitable for our data due to graph truncation.

Blast is very functional and useful, but slow.

Though the aim of this was the observe how software ran nanopore data-sets, with long reads, and unfortunately the data-sets were short and FastQC didn’t seem to run well for nanopore sequence data. Also, despite continuously observing that the results were low in quality, perhaps this is how nanopore data-set conclude: many other teams who run quality checks on their data and come to the results of poor quality.

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 — afc@aber.ac.uk
Andre Soares — ans74@aber.ac.uk
Sam Nicholls — msn@aber.ac.uk
4. References

References


