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

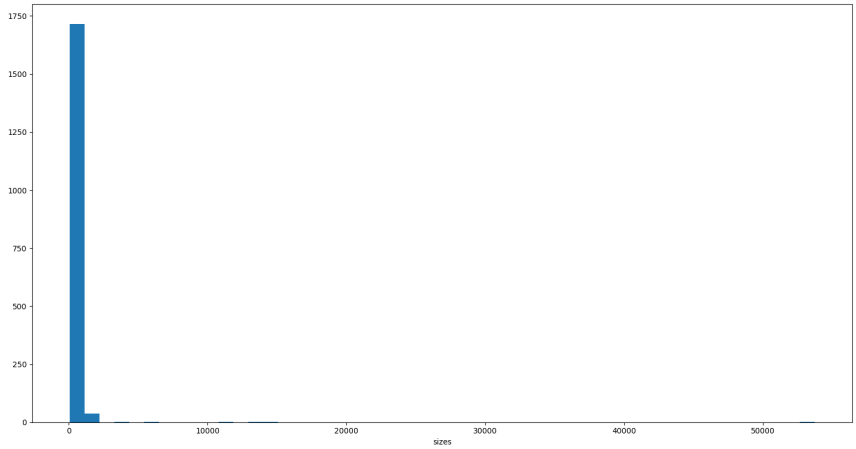
A wide variety of software is available today for the analysis of nanopore metagenomic data. DNA extracted from a coal mine located in South Wales, where no internet access is available, was run through a the MinION[1] and then analysed in a pre-printed paper[1] - results told us of the wide variety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

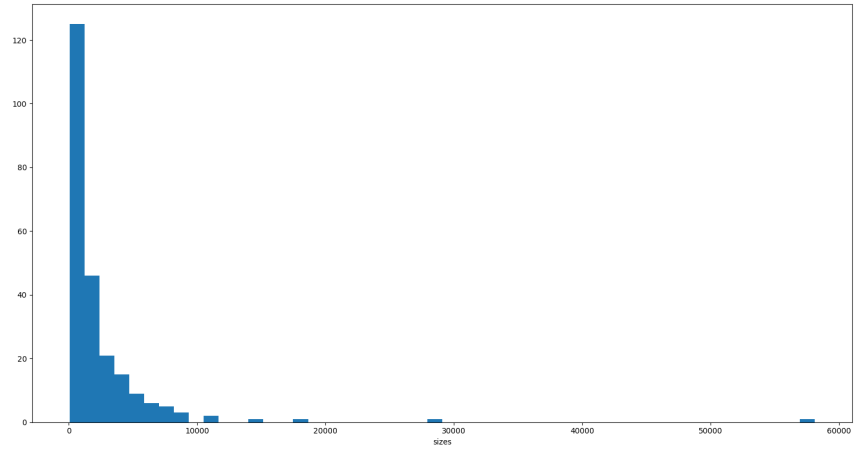
Moreover `BP_v2` was not quality filtered; it had MUX genes (QC reads), which were later removed for analysis - `BP_v1` has a total of 1,770 `fast5` files & `BP_v2` has a total of 3,019.

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
[2] http://onlinelibrary.wiley.com/doi/10.1111/j.1574-6941.2012.01325.x/full



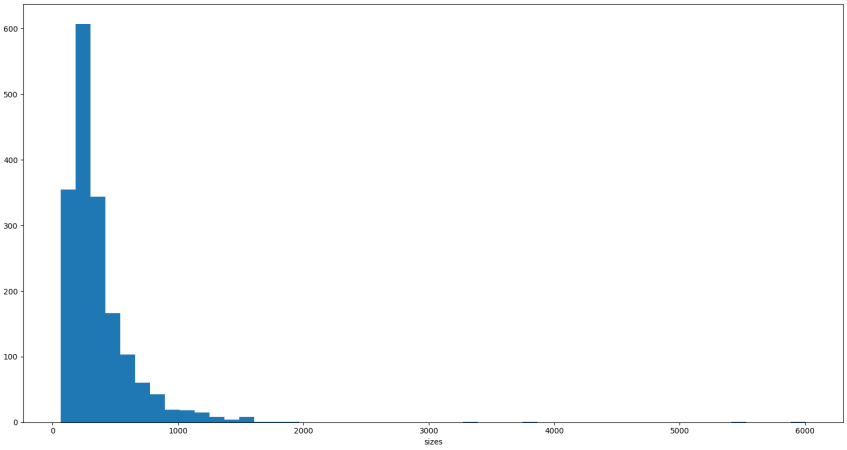
(a) `BP_v1`



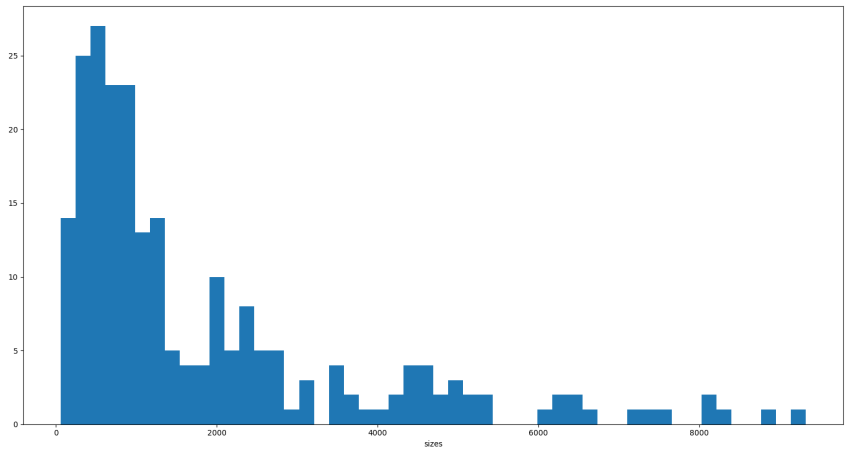
(b) `BP_v2`

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).



(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

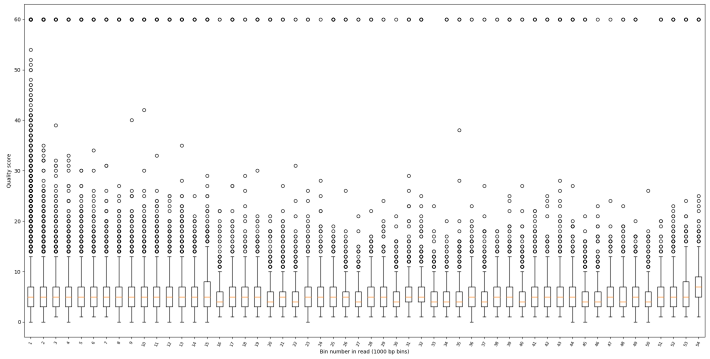
Further, in-depth analysis of DNA data-set will be conducted in this report, using a variety of software and methods/tools (looking into sequence similarity).

2.1. poretools

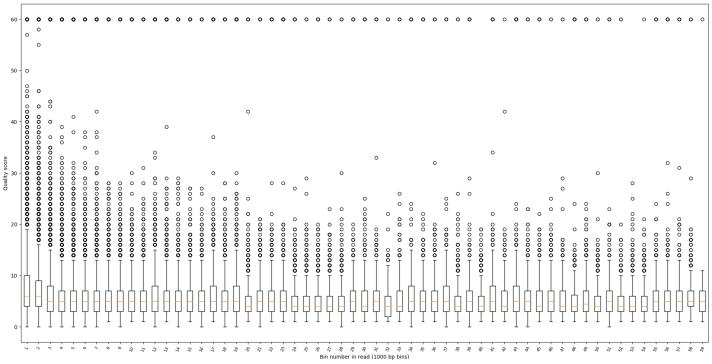
`poretools`[2] is a toolkit for analysing nanopore sequence data - histograms were produced to observe read length (as seen with Figures 1 & 2).

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]).



(a) `BP_v1`



(b) `BP_v2`

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

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

!	0	150	736417	0.000203688942542
"	1	15852	736417	0.0215258474478
#	2	78037	736417	0.105968493394
$	3	116139	736417	0.157708200653
%	4	116956	736417	0.158817626426
&	5	100021	736417	0.135821144813
'	6	78721	736417	0.106897314972
(	7	59902	736417	0.081342500241
)	8	43800	736417	0.0594771712223
*	9	31744	736417	0.043106011947
+	10	22487	736417	0.0305356883396
,	11	15933	736417	0.0216358394768
-	12	11292	736417	0.0153337035946
.	13	8562	736417	0.0116265648403
/	14	6180	736417	0.00839198443273
0	15	4576	736417	0.00621387067382
1	16	3380	736417	0.00458979083861
2	17	2684	736417	0.00364467414522
3	18	2082	736417	0.00282720252248
4	19	1594	736417	0.00216453449608
5	20	1319	736417	0.00179110476809
6	21	1038	736417	0.00140952748239
7	22	734	736417	0.000996717892173
8	23	606	736417	0.00082290332787
9	24	469	736417	0.000636867427015
:	25	358	736417	0.000486137609534
;	26	248	736417	0.000336765718336
<	27	225	736417	0.000305533413813
=	28	164	736417	0.000222699910513
>	29	128	736417	0.000173814564303
?	30	89	736417	0.000120855439242
@	31	77	736417	0.000104560323838
A	32	70	736417	9.5054839853e-05
B	33	49	736417	6.65383878971e-05
C	34	23	736417	3.12323045231e-05
D	35	32	736417	4.34536410756e-05
E	36	30	736417	4.07377885084e-05
F	37	16	736417	2.17268205378e-05
G	38	13	736417	1.7653041687e-05
H	39	9	736417	1.22213365525e-05
I	40	12	736417	1.62951154034e-05
J	41	11	736417	1.49371891198e-05
K	42	3	736417	4.07377885084e-06
L	43	1	736417	1.35792628361e-06
M	44	4	736417	5.43170513446e-06
N	45	2	736417	2.71585256723e-06
O	46	2	736417	2.71585256723e-06
P	47	2	736417	2.71585256723e-06
Q	48	1	736417	1.35792628361e-06
S	50	2	736417	2.71585256723e-06
T	51	1	736417	1.35792628361e-06
U	52	1	736417	1.35792628361e-06
W	54	1	736417	1.35792628361e-06
]	60	10585	736417	0.0143736497121

(a) `BP_v1`

!	0	67	567745	0.000118010726647
"	1	11421	567745	0.0201164255079
#	2	50787	567745	0.0894538921523
$	3	77712	567745	0.136878352077
%	4	81437	567745	0.143439396208
&	5	72117	567745	0.127023575725
'	6	58995	567745	0.103911086844
(	7	46189	567745	0.0813551858669
)	8	35772	567745	0.0630071599045
*	9	27351	567745	0.0481747967838
+	10	20815	567745	0.0366625861963
,	11	16396	567745	0.0288791623
-	12	12640	567745	0.0222635161912
.	13	9995	567745	0.0176047345199
/	14	8205	567745	0.0144519106289
0	15	6549	567745	0.0115351081912
1	16	5687	567745	0.0100168209319
2	17	4711	567745	0.00829773930198
3	18	4069	567745	0.00716694995112
4	19	3316	567745	0.00584065029194
5	20	2633	567745	0.0046376454218
6	21	1605	567745	0.00282697337713
7	22	854	567745	0.00150419642621
8	23	583	567745	0.00102686945724
9	24	423	567745	0.00074505279659
:	25	293	567745	0.000516076759813
;	26	233	567745	0.00041039551207
<	27	167	567745	0.000294146139552
=	28	123	567745	0.000216646557874
>	29	106	567745	0.00018670353768
?	30	68	567745	0.000119772080776
@	31	51	567745	8.98290605818e-05
A	32	56	567745	9.8635831227e-05
B	33	32	567745	5.63633321297e-05
C	34	18	567745	3.1704374323e-05
D	35	21	567745	3.69884367101e-05
E	36	13	567745	2.28976036777e-05
F	37	11	567745	1.93748954196e-05
G	38	15	567745	2.64203119358e-05
H	39	10	567745	1.76135412905e-05
I	40	12	567745	2.11362495487e-05
J	41	13	567745	2.28976036777e-05
K	42	6	567745	1.05681247743e-05
L	43	4	567745	7.04541651622e-06
M	44	2	567745	3.52270825811e-06
N	45	2	567745	3.52270825811e-06
O	46	3	567745	5.28406238716e-06
P	47	1	567745	1.76135412905e-06
S	50	1	567745	1.76135412905e-06
X	55	1	567745	1.76135412905e-06
Z	57	1	567745	1.76135412905e-06
[	58	1	567745	1.76135412905e-06
]	60	6152	567745	0.0108358506019

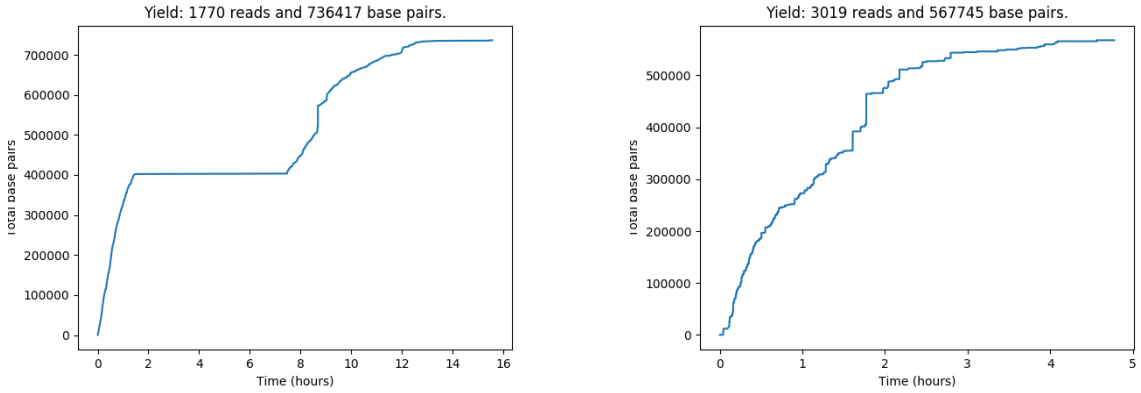
(b) `BP_v2`

Figure 4: `poretools qualdist` comparison

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.



(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 shoes 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 alot of `AT` sequences.

2.2. Goldilocks

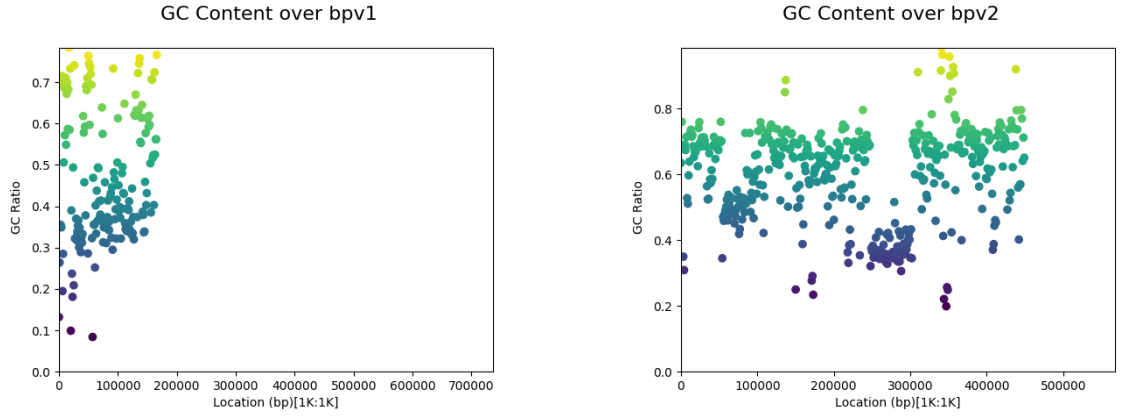
`Goldilocks`[3] was developed and used to "quickly locate *interesting* regions on the human genome that expressed a desired level of variability, which were *just right* for later variant calling and comparison."

`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.

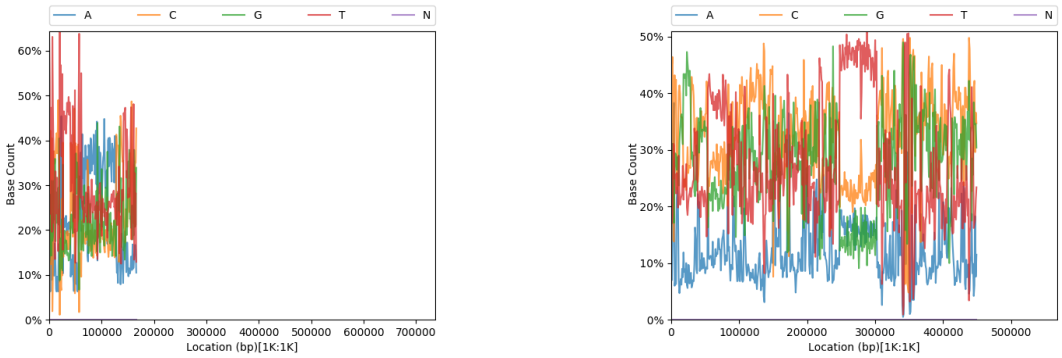


(a) `BP_v1` (b) `BP_v2`

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

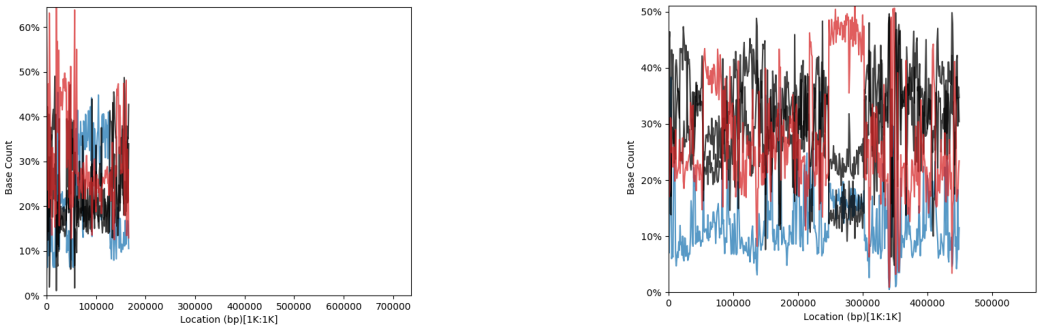
2.2.2. ACGT Content

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 `T`s. Moreover, `BP_v1` again has a high number of `T`s 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.



(a) `BP_v1` (b) `BP_v2`

Figure 7: `ACGT` Plot of `NucleotideCounterStrategy` content using `Goldilocks`:- `x-axis`: the location along the genome; `y-axis`: percentage of content cover



(a) `BP_v1` (b) `BP_v2`

Figure 8: `AT` Plot of `BP_v1` & `BP_v2`

`Goldilocks`'s `g.query` was used to find the reads with the most `T`s in `BP_v2` & most `A`s 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).

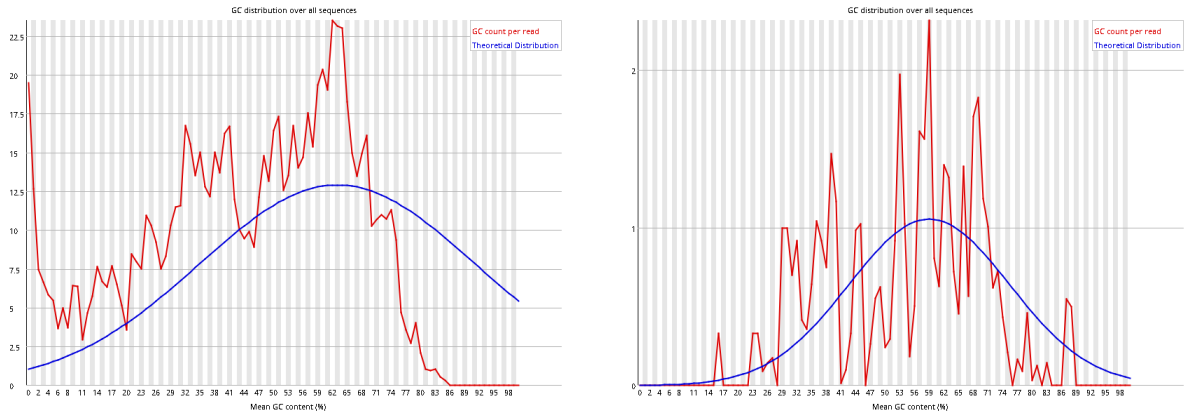
Note: `SAMtools`[4] was used to index `fasta` files so we could use them with `Goldilocks`, plus to convert files to `fastq` (we also used `poretools`).

2.3. FastQC

`FastQC`[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%.



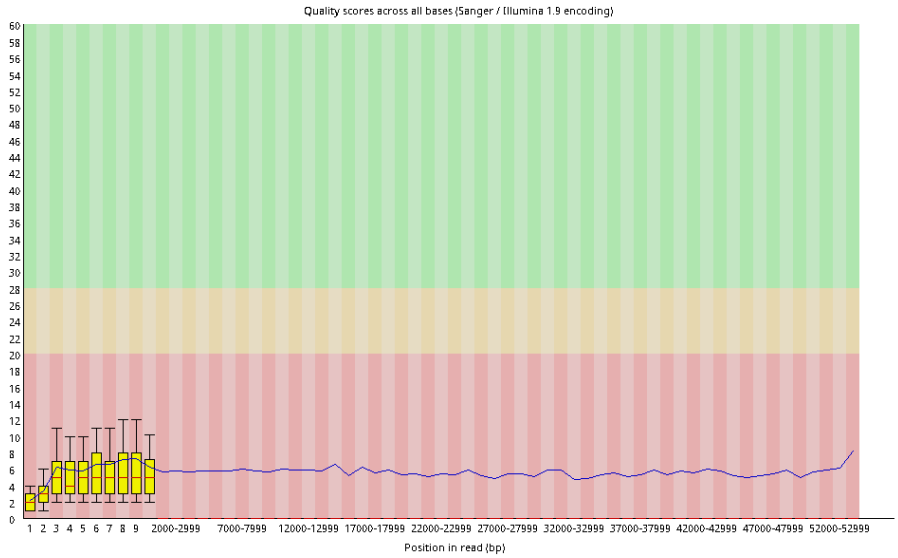
(a) `BP_v1` (b) `BP_v2`

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

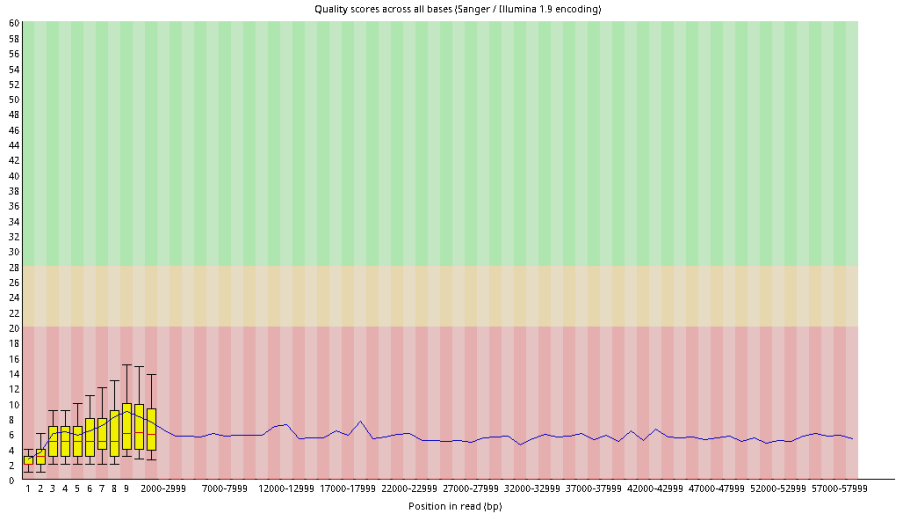
Comparing `GC` in `FastQC` & `Goldilocks` may be quite difficult here due to difference of graphs: `FastQC` creates line whilst `Goldilocks` creates scatter, as seen in Figure 6.

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.



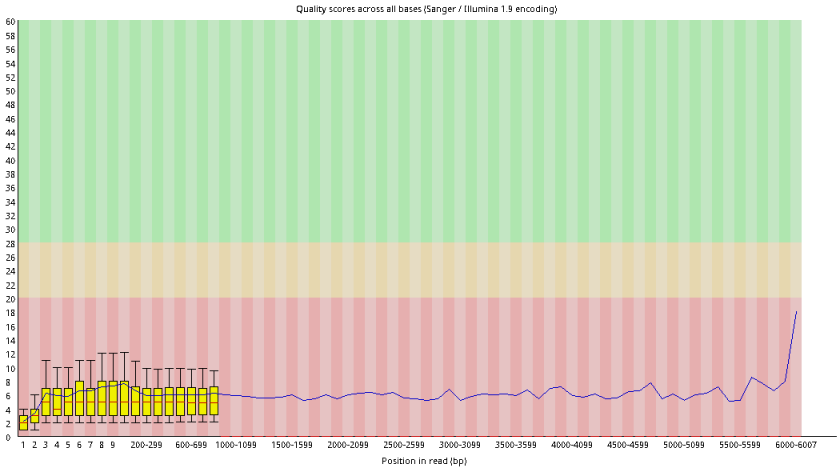
(a) `BP_v1`



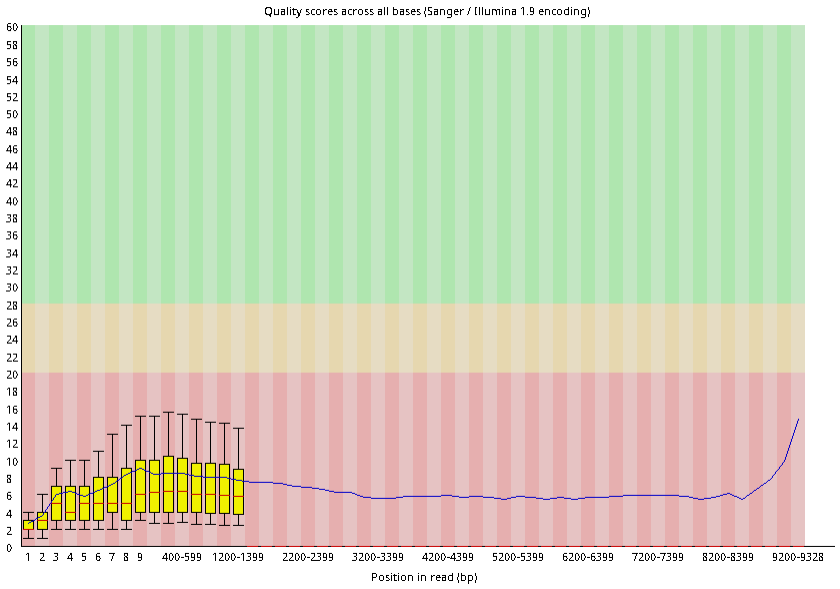
(b) `BP_v2`

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.



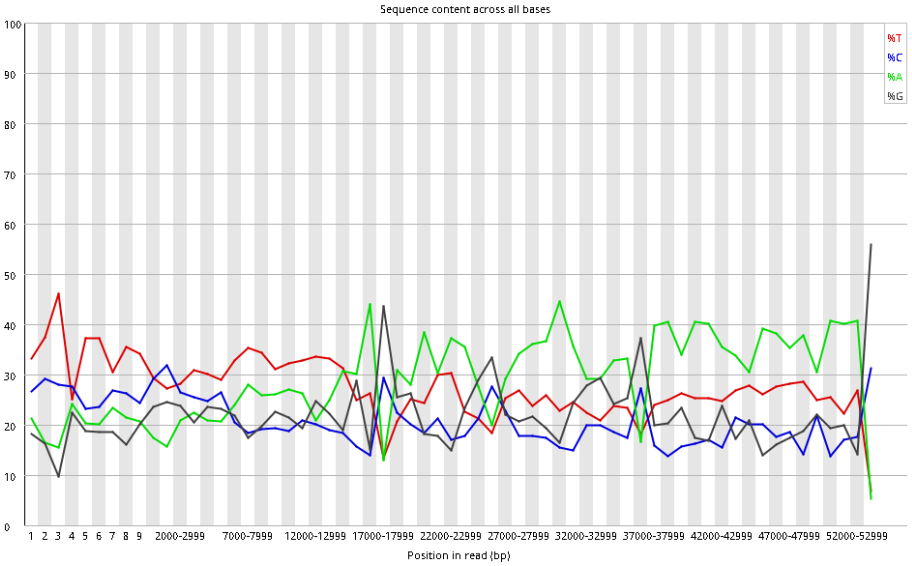
(a) `BP_v1`



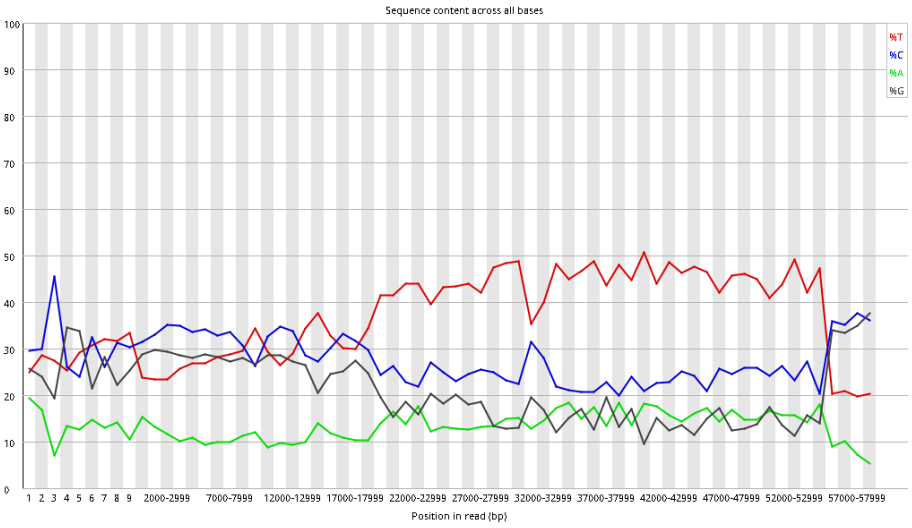
(b) `BP_v2`

Figure 11: Plot of whole data-sets limited to 10,000 in `FastQC` quality graph

2.3.3. *ACGT Content*



(a) `BP_v1`



(b) `BP_v2`

Figure 12: Plot of `ACGT` content in `FastQC` graph

A	C	G	T
green	blue	black	red

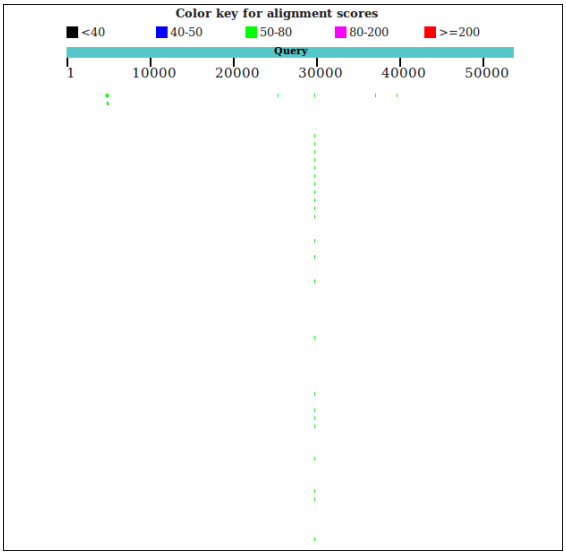
2.4. BLAST

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

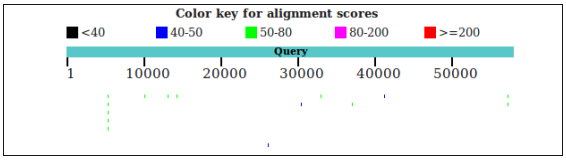
Note: links in this section may have expired.

Blasting longest reads

Further tests included using nucleotide BLAST[6] (blastn) with discontigious 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.



(a) `BP_v1`



(b) `BP_v2`

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

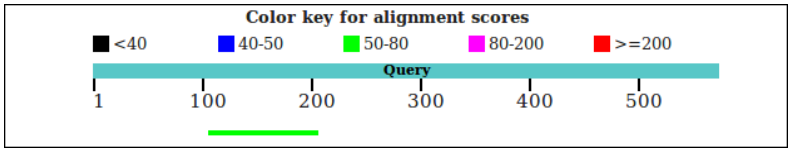
Blasting random reads

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_v2`[5] (see Figure 14) - we can observe this data read was bacteria (filamentous thermophilic bacteria[6]) - Figure 14b.
Figure 14a shows the results: there was a query cover of 17%.

`channel_349_bf5a3f39-3e35-4870-82ec-a7f9679d4000_qscore_9_read_score_-1.8`



(a) Blast Alignment Scores

Score	Expect	Identities	Gaps	Strand
53.6 bits(58)	0.006	73/101(72%)	1/101(0%)	Plus/Plus

```
Query  106      GTACTTTACCATCAGATTATACTCTCTTTACATCCGGAAAGCCCGGGGCGCTACGTGGAT  165
                 ||||||||||| | |||||||||| | |||| ||| | || || || | || ||||||
Sbjct  2574171  GTACTTTACCAAGAAATTATACTCGCATTACGTCCACACAGTCCCGGCAGGTATGTGGAT  2574230

Query  166      ACGACCGTCGGAGC-AGAGGACATGCCGCGAGCATCCTGGA  205
                 | || || || || | || || |||||  |||| |||||
Sbjct  2574231  GCAACGGTTGGCGCGGGGGGTCACGCCGCTGGCATTCTGGA  2574271
```

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

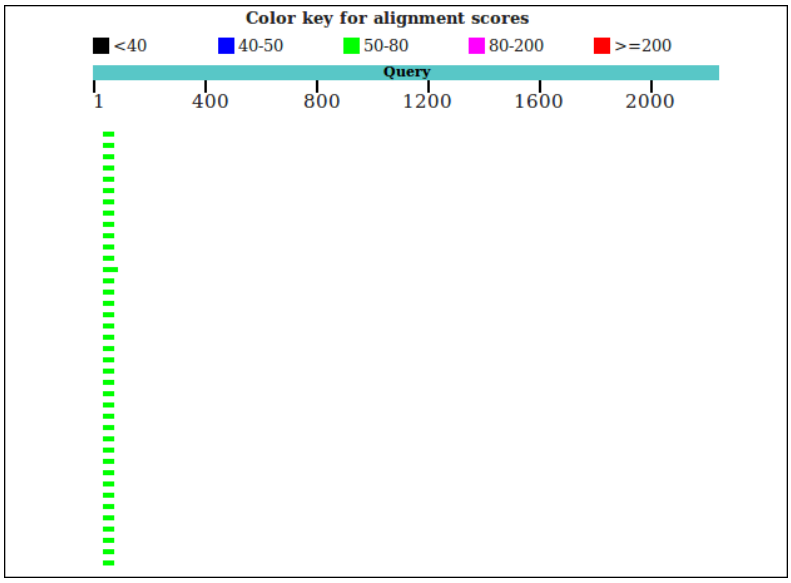
Figure 14: Blast of a random read from `BP_v2`

Anaerolinea thermophilia 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).

[5]https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Get&RID=V2PU8EH301R
[6]https://www.ncbi.nlm.nih.gov/pubmed/14657113

For further analysis, we extracted another random read from `BP_v2`[7] (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.

`channel_176_363eaef2-4772-408e-8641-5c1c111fd6bb_qscore_8.4_read_score_-1.6`



(a) Blast Alignment Scores

Score	Expect	Identities	Gaps
67.6 bits(36)	2e-06	43/46(93%)	2/46(4%)

```
Query  35  GTTTCGCATTTATCGTGAAACGCTTTCGCATTT--CGTGCGCCGCT  78
           ||||||||||||||||||||||||||||||| ||||||||||||
Sbjct  49  GTTTCGCATTTATCGTGAAACGCTTTCGCGTTTTTCGTGCGCCGCT  4
```

(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.

[7] https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Get&RID=V33NGY1B016

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].

`BP_v1` didn't produce any good results, unfortunately the highest result of alignment length was 49; highest bit-score was 67.6.
`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_qscore_4.5_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_qscore_4.6_read_score_-3.3` Capsicum annuum[11]: Sweet and chili peppers (plant)
- 39 - `channel_425_8d5bc381-5611-4884-ac32-8046adabe9d7_qscore_4.5_read_score_-3.4` Pygocentrus nattereri[12]: Red-bellied piranha (animal)

`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.
`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_qscore_9.5_read_score_-1.6` Neorhizobium galegae[13]: bacteria that forms nitrogen-fixing root nodules

[8] ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/
[9] https://www.ncbi.nlm.nih.gov/gquery/
[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[14]: a phototrophic genus of bacteria - organisms that carry out photosynthesis
  - Azorhizobium caulinodans[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[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[17]: recovered strains from cystic fibrosis (strains from human respiratory specimens)
- 330 - channel_185_cdc2062b-5bab-4330-8fdc-1ec4311566cc_qscore_8.3_read_score_-1.8
  Nitrosomonas[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

[14]https://www.ncbi.nlm.nih.gov/gquery/?term=CP007440.1
[15]https://www.ncbi.nlm.nih.gov/nuccore/AP009384.1
[16]https://www.ncbi.nlm.nih.gov/gquery/?term=CP001965.1
[17]https://www.ncbi.nlm.nih.gov/gquery/?term=CP016172.1
[18]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

[1] A. Edwards, A. Soares, S. Rassner, P. Green, J. Felix, and A. Mitchell. *Deep Sequencing: Intra-Terrestrial Metagenomics Illustrates The Potential Of Off-Grid Nanopore DNA Sequencing.* bioRxiv, May-2017. bioRxiv 133413 — http://www.biorxiv.org/content/early/2017/05/02/133413.

[2] N. Loman, and 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, and J. Randall. *Goldilocks: a tool for identifying genomic regions that are just right.* Bioinformatics Vol.32 2016. pages: 2047-2049.

[4] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. *The Sequence Alignment/Map format and SAMtools..* Bioinformatics Vol.25 2009. pages: 2987-2993.

[5] S. Andrews. *A quality control tool for high throughput sequence data.*, 2015. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[6] National Center for Biotechnology Information (NCBI). *BLAST, Basic Local Alignment Search Tool*, U.S. National Library of Medicine. Accessed Aug-2017. https://blast.ncbi.nlm.nih.gov/Blast.cgi