Ubiquitous Genomics: Hackathon2

Maya Anand
mva2112@columbia.edu

Cheyenne Parsley
cep2141@columbia.edu

Robert Piccone
rap2186@columbia.edu

Daniel Speyer
dls2192@columbia.edu

November 20, 2015

Problem 1

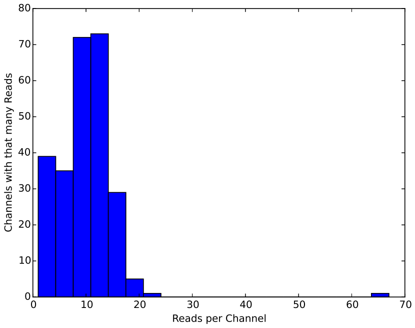
Number of 2D reads classified as failed: 258
Number of 2D reads classified as passed: 1082

Problem 2

255 channels had at least one read, and 216 had at least five. This compares with 412 "active" channels during initialization, and 618 immediately after loading fuel

The average channel had 9.9 reads. Channel 53 had 67 reads, which was the most.

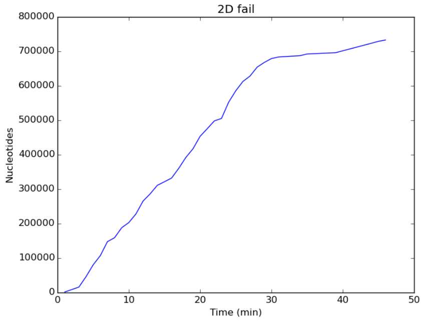
Just for fun, here's a histogram of reads per channel



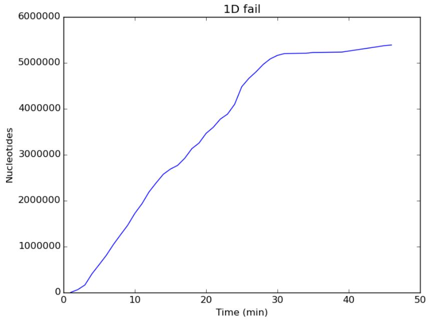
Problem 3

Failed Reads

The following plots show the length distribution of 2D and 1D reads for fails.



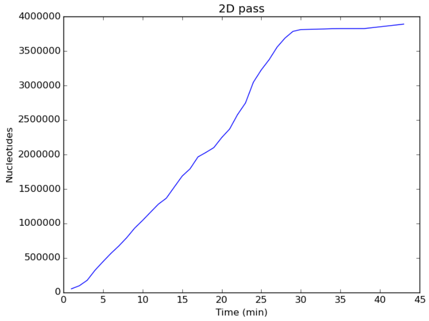
2D Reads



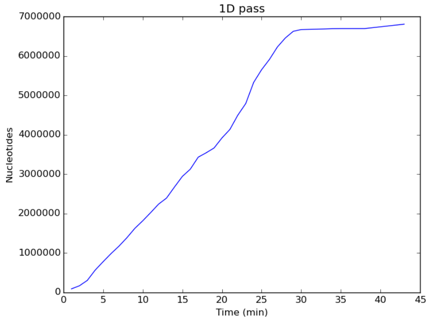
1D Reads

Passed Reads

The following plots show the length distribution of 2D and 1D reads for passes.



2D Reads

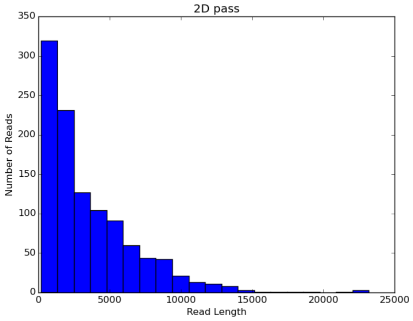


1D Reads

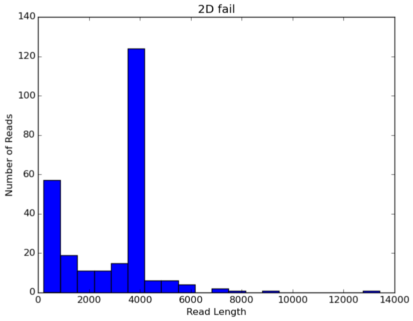
Problem 4

2D reads

The following histograms show the length distribution of 2D reads for passes and fails.



Passed Reads



Failed Reads

Problem 5

LONGEST PASSED 2D READ
From file: MINION02_Hackathon2_group4_TeamAWESOME_4029_1_ch9_file8_strand.fast5
Number of nucleotides: 23196

LONGEST FAILED 2D READ
From file: MINION02_Hackathon2_group4_TeamAWESOME_4029_1_ch360_file3_strand.fast5
Number of nucleotides: 13419

Problem 6

Total # of aligned reads: 851
Total # of unaligned reads: 231

Total # of reads: 1082

Problem 7

As with hackathon1, only some of the reads could be aligned and of those only portions of them. The usual concerns about selection bias apply. Furthermore, finding the reference sequence for alignments to the complement strand proved difficult, so we offer here only the reads which aligned to the template strand.

This table shows count of nucleotides from those alignments. Rows indicate the nucleotide in the reference genome, columns in the read returned by MinION.

	A	C	G	T	-
A	170885	3697	4060	1611	11374
C	2084	116750	2586	2262	6074
G	2515	2554	114255	1837	6446
T	1741	3952	3316	171694	10925
-	10330	11696	12358	9871	0

Problem 8

To reduce the number of errors in the reads, we could try several methods:

First, we could have replicated the DNA with PCR using random primers to have more copies of fewer fragments. By sequencing multiple fragments of the same sequence we could have some redundancy and use that to cross-check our data. Since this involves the materials preperation, it is too late to do it now.

Something we can do is look at the quality scores and only pay attention to high ones. This could lead to throwing out poor data, and improving the overall accuracy of the reads. Though this sounds like a solid strategy in theory, it should be noted that in the last hackathon, we found that the quality scores were not very good. Also, if there were certain sequences that are specifically difficult to sequence for whatever reason, this method would not include them, introducing selection bias.

Also, we can improve accuracy by only believing polymorphisms that are known to be in >1% of the population. Otherwise, we must believe it is a sequencing error, and not a variation.