Quality Control

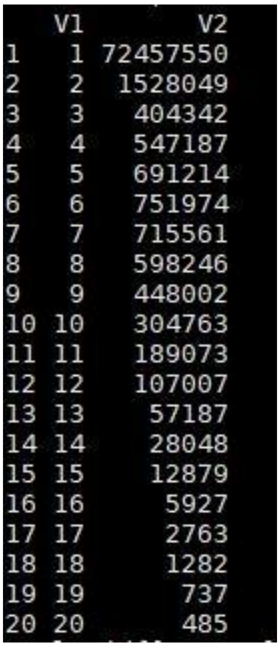
Illumina PE 101x2 reads are all 17 quality. Although 17 is rather low and could be considered trimmable, it is practically not possible for this data. All reads are 101 bp long so there is also no need for trimming depending on read length. There are 2475000 number of sequences with an overall GC% of 36%.

Pacbio long reads are between average of 7 and 10 quality, which is expected since long reads will usually have much lower quality than short reads. Trimming based on quality is also not practical on this data since the quality range is rather short, which means any trimming would significantly reduce the overall usability of the data. Reads are between 167 and 24651, with most of the reads being on the lower end of the spectrum. There are 33413 number of sequences with an overall GC% of 37%.

It is expected to get a quality assembly with these data sets with a high N50 since hybrid assemblers can use the long reads to combine short reads into contigs. But since the long reads are clustered around the low end read lengths, there might be higher than desired number of short contigs.

K-mer Analysis

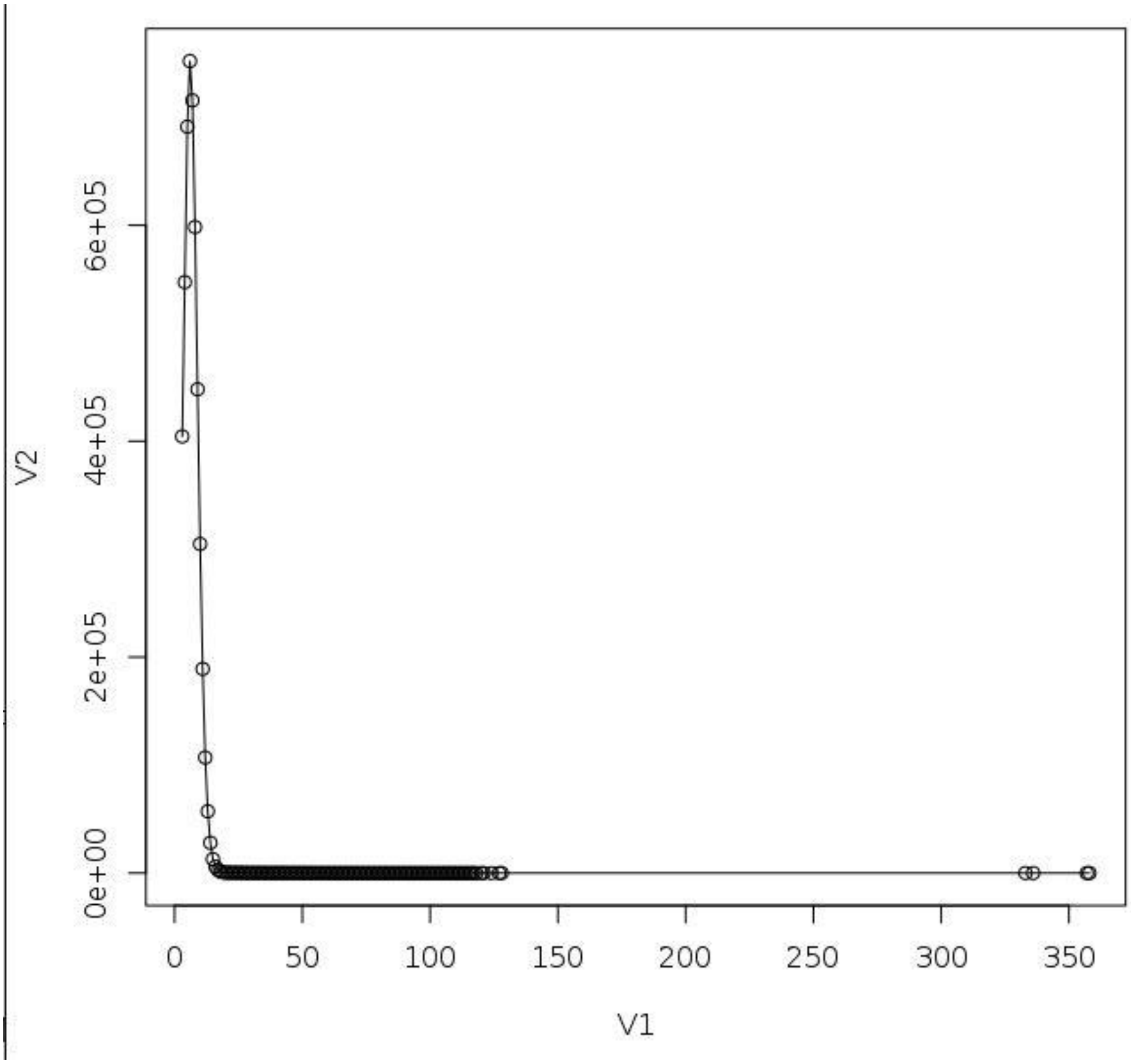
K-mer analysis is done with jellyfish count function on the short Illumina reads. K-mer size is selected as **58** since a k-mer size between 50%-66% of the read length is often optimal.



```
   V1       V2
1   1 72457550
2   2  1528049
3   3   404342
4   4   547187
5   5   691214
6   6   751974
7   7   715561
8   8   598246
9   9   448002
10 10   304763
11 11   189073
12 12   107007
13 13    57187
14 14    28048
15 15    12879
16 16     5927
17 17     2763
18 18     1282
19 19      737
20 20      485
```

First 20 k-mer counts

We can assume that the first two data points are leftover error kmers. Ignoring them, the below graph shows the points that we can regards as real kmers to use in calculations.



Total number of real kmers are 33386352. The peak is at position 6 with 751974. The genome size G can be estimated by total_kmers / peak position, which gives us **G = 5564392.**

General Information about the whole dataset

R (number of reads) = 2475000

L (length of reads) = 101

K (kmer size) = 58

N_base (bases sequenced) = 253948358

N_kmer (kmer count) = 108900000

C_base (base coverage) = 45.6

C_kmer (kmer coverage) = 19.5

De-Novo Assembly

Steps that are used in the pipeline to produce the final assembly fasta files:

-QC (which is covered previously)

-Error correction (using KmerFreq_AR and Corrector_AR)

-Hybrid Assembly (using DBG2OLC or MaSuRCA)

-Hybrid Assembly Polishing (using samtools and pilon)

Error Correction

KmerFreq_AR is used to create a kmer spectrum for both the Illumina short reads. Kmer size is selected as 15 because of the program's memory limitations.

Then Corrector_AR is used to correct the kmer spectrum that is produced previously with the same kmer size of 15.

These steps produce corrected reads for PE short reads. Before correction, the total sequence length for a PE was 253948358, after correction, it has been reduced to 223928427, which is around 12% reduction in length.

Hybrid Assembly

1) First attempt to assemble was with DBG2OLC. SparseAssembler function of DBG2OLC is used to build the de-bruijin graph assembly of the short reads. Genome size was selected as **5564392.** Node Coverage threshold as 1, edge coverage threshold as 1, and kmer size as **27**. Kmer size is rather low for 101 read lengths but it will be tested against the later attempts (which will be 50+)

Then Overlap Layout Consensus step is performed to integrate the contigs we produced into longer contigs using the long pacbio reads. Parameters that are used are, kmer size 17 (because memory restrictions), Adaptive threshold 0.0001, kmer coverage threshold 2, minimum overlap 20 and removing chimeras as true.

The script my_split_nrun_sparc.sh is used in the final step to put together the data and produce an assembled fasta file.

2) Second attempt followed the same steps with DBG2OLC, except selecting the kmer size as **50,** which is more conventional for a read length of 101 bp.

3) Third attempt used MaSuRCA hybrid assembler. MaSuRCA requires the read size of the short reads along with the standard deviation. Read length is 101 and standard deviation by estimate is 15. MaSuRCA produces an assembled fasta file.

Assembly Polishing

Polishing is used to improve the quality of the assembled fasta files. Short reads and the assembled fasta file is required to do the polishing step. Pilon will be used to polish the results, but first, samtools are used to convert the short reads into indexed alignments.bam file.

Then pilon is used with PE short reads as parameters and genome as the assembled fasta file. Result is the final fasta file which is polished and corrected.

Results

1st Attempt (DBG2OLC, Short read kmer = 27, Long read kmer = 17):

```
Total number of sequences:          53
Total length of sequences:          4838568 bp
Shortest sequence length :          1310 bp
Longest sequence length  :          367916 bp
Total number of Ns in sequences:    0
N50:    244710  (8 sequences)   (2481661 bp combined)
```

2nd Attempt (DBG2OLC, Short read kmer = 50, Long read kmer = 17)

```
Total number of sequences:          40
Total length of sequences:          4809030 bp
Shortest sequence length :          7501 bp
Longest sequence length  :          490640 bp
Total number of Ns in sequences:    0
N50:    210212  (8 sequences)   (2451467 bp combined)
```

3rd Attempt (MaSuRCA, Short read length = 101, std = 15):

```
Total number of sequences:          36
Total length of sequences:          4914209 bp
Shortest sequence length :          8540 bp
Longest sequence length  :          644526 bp
Total number of Ns in sequences:    0
N50:    245065  (7 sequences)   (2500838 bp combined)
```

Estimated genome size was 5564392, assembled genome size in the best(3rd) attempt is 4914209, which is respectable.

Difference between first two attempts can be explained by the kmer size, lower kmer size will give more matches with the consensus layout alignment with the long reads, which will lead to a higher N50. But this does not mean that the quality of the result is better.

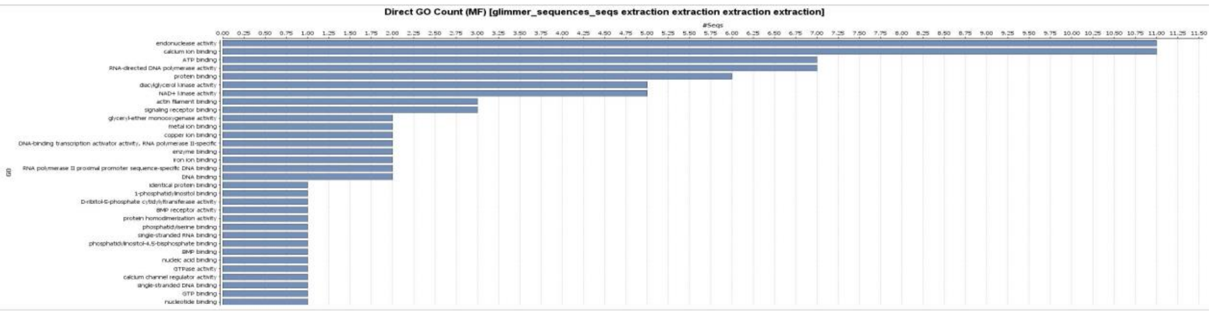
Although 1st and 3rd attempt N50 numbers are very close, 3rd attempt (MaSuRCA) result should be the more reliable and stable result with more quality since the script is dependent on the read length more than the kmer size.

in-silico Gene Prediction using Augustus

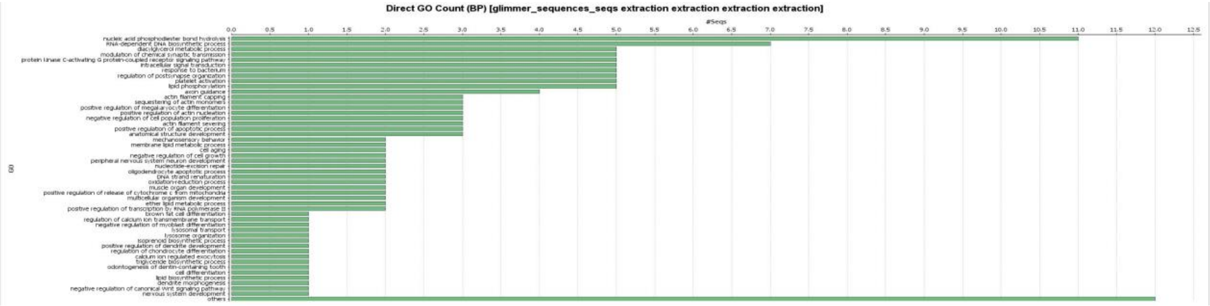
Gene prediction is done by using Blast2Go with prokaryotic geneFinding function in Augustus.

Sequences are blasted, mapped and annotated (Interpro also).

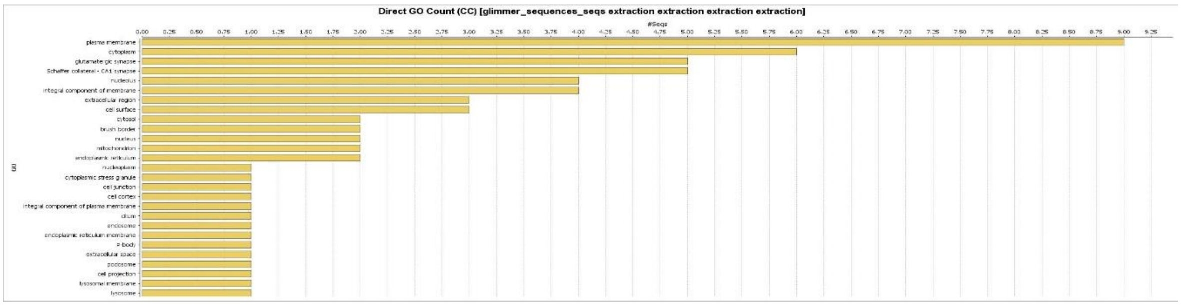
Results that had more than 99% sim mean, in three different categories are as follows:



Molecular Function Results



Biological Processes Results



Cellular Component Results