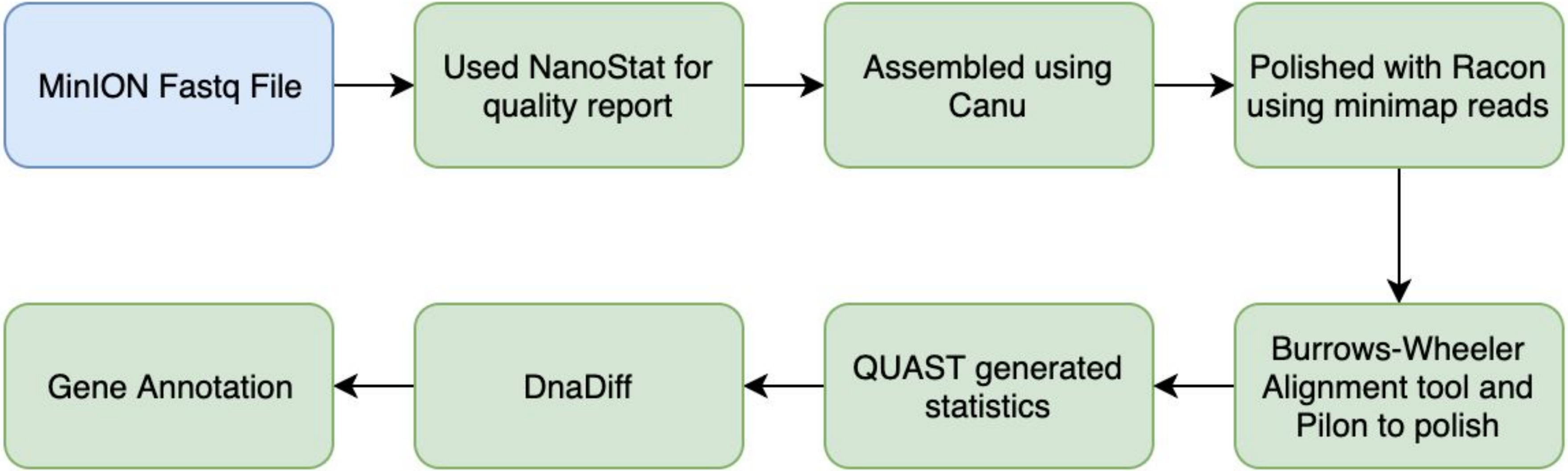
Arabidopsis thaliana, Thale Cress, Genome Assembly

By Jordan Callahan

Purpose of Project

- Assemble complete genome using only minION data
- Establish high contiguity genome with racon and pilon polishing
- Observe the top hits for protein-encoding genes

Workflow



MinION Fastq File

Used NanoStat for quality report

Assembled using Canu

Polished with Racon using minimap reads

Burrows-Wheeler Alignment tool and Pilon to polish

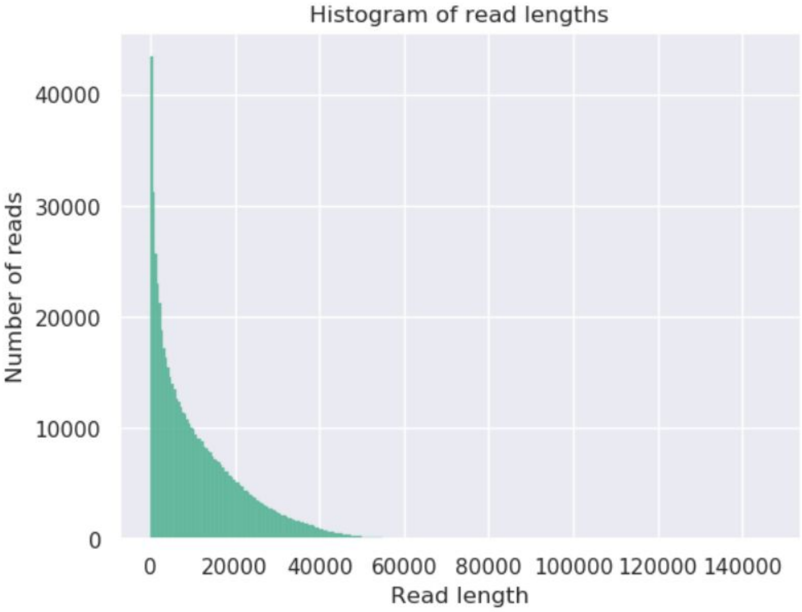
QUAST generated statistics

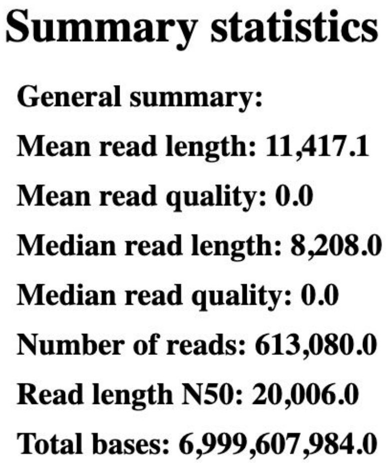
DnaDiff

Gene Annotation

Nano Plot

- 309947 reads of the original ~9 million did not pass quality score of 7





Summary statistics

General summary:

Mean read length: 11,417.1

Mean read quality: 0.0

Median read length: 8,208.0

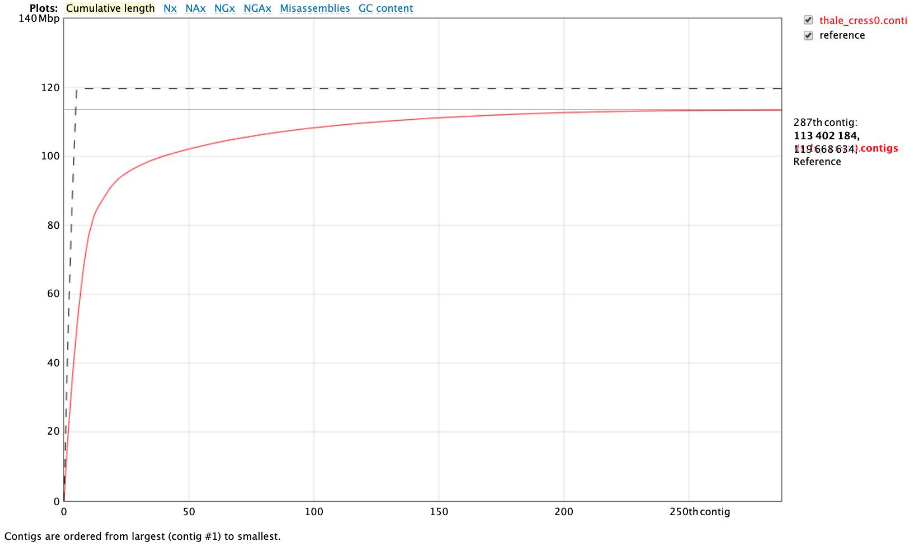
Median read quality: 0.0

Number of reads: 613,080.0

Read length N50: 20,006.0

Total bases: 6,999,607,984.0

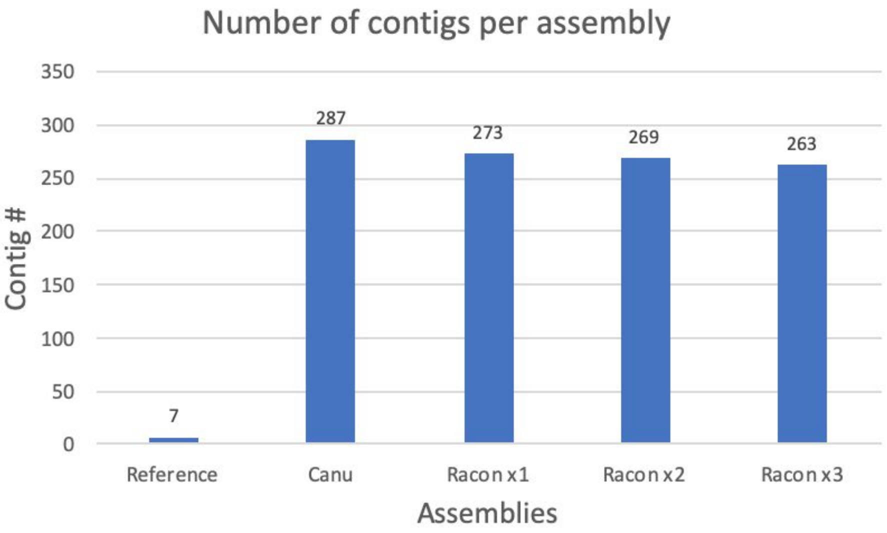
QUAST


Plots: Cumulative length Nx NAx NGx NGAx Misassemblies GC content
140 Mbp
120
100
80
60
40
20
0
0
50
100
150
200
250th contig

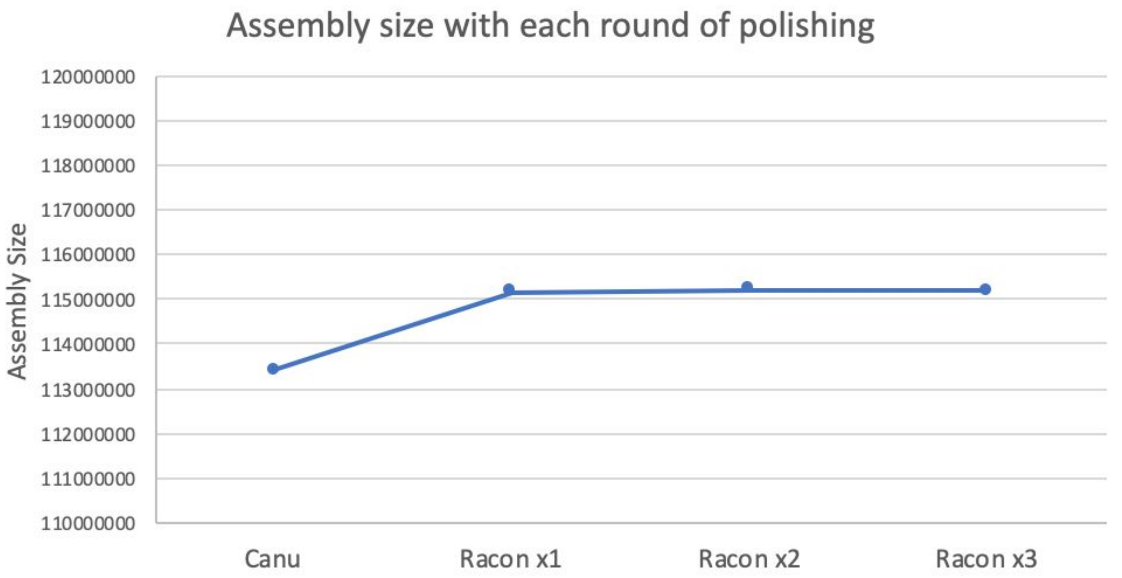
☑ thale_cress0.conti
☑ reference

287th contig:
113 402 184,
119 668 634) **contigs**
Reference

Contigs are ordered from largest (contig #1) to smallest.

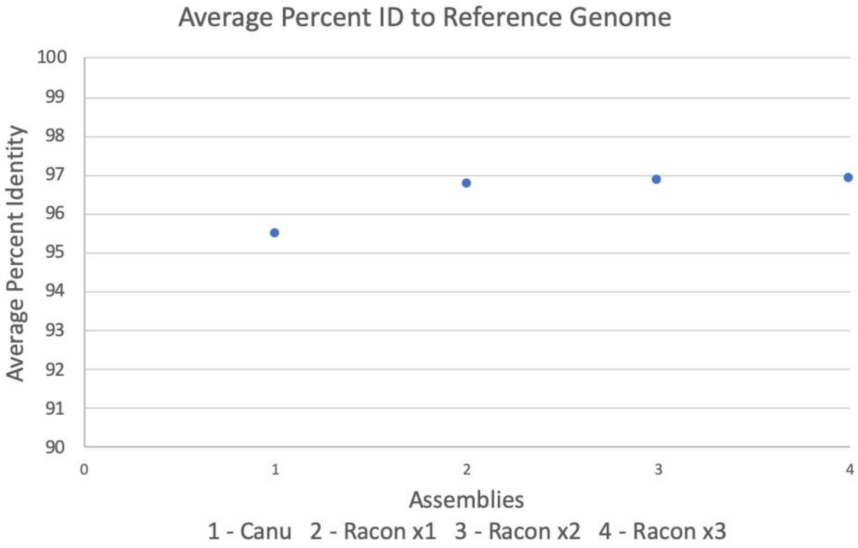


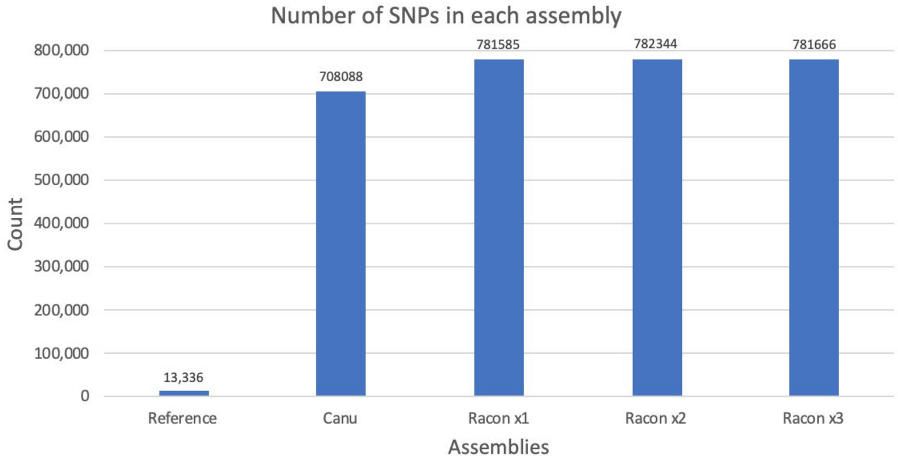
Assembly Size

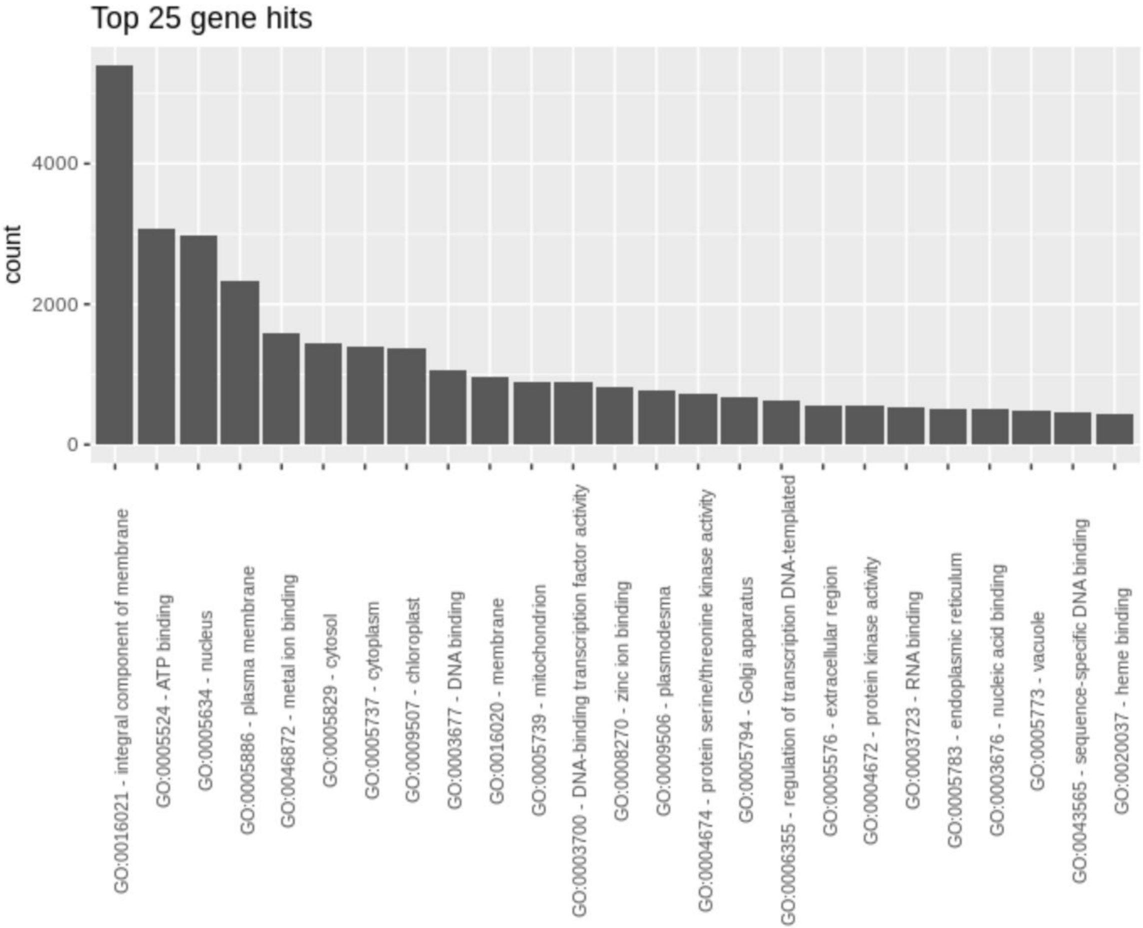


- Reference assembly size is 119,667,750

DnaDiff report







Gene Annotation

- Prodigal gene prediction to create a list of proteins
- GOFEAT website
- Read csv file into ggplot
- Top 25 gene hit

Concluding Statements

- A complete genome was successfully assembled and analyzed
- Workflow was altered from original experiment (Miniasm, blastn)
- From the paper, researchers were able to assemble into 62 contigs with an N50 of 12.3 Mb.

References

Michael, T.P., Jupe, F., Bemm, F., Motley, S.T., Sandoval, J.P., Lanz, C., Loudet, O., Weigel, D. and Ecker, J.R., 2018. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature communications*, *9*(1), p.541.