Genome Assembly Final Results

Team 1 Genome Assembly

Lawrence McKinney, Laura Mora,
Jessica Mulligan, Heather Patrick,
Devishi Kesar and
Cecilia (Hyeonjeong) Cheon

February 18, 2020

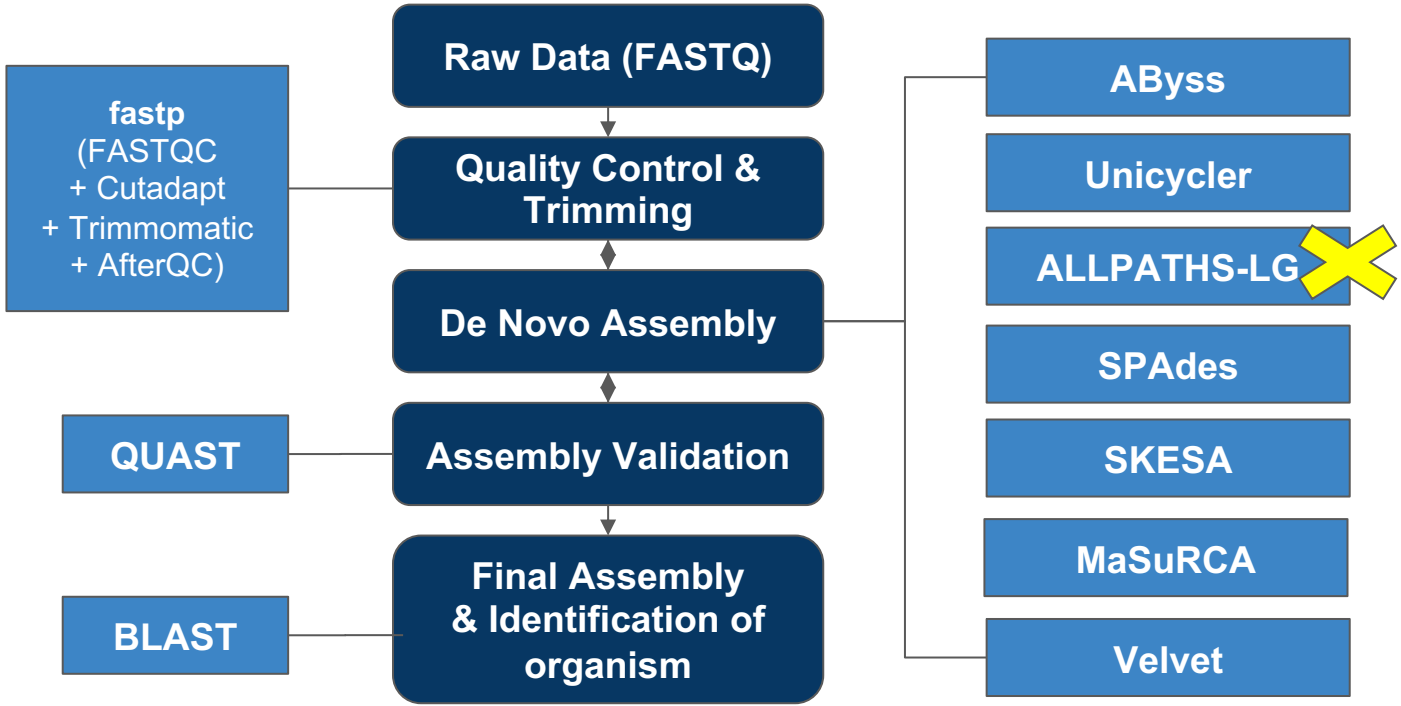
Outline

- Approach Overview
- Reads Pre-processing - Fastp
- Assembler Evaluation Criteria
- Evaluation Results
- Identification of Pathogen
- Revised Pipeline
- References

Approach Overview

LEGEND:

Removed due to the need for multiple libraries to enable higher assembly quality



Reads preprocessing - Fastp

- Used for quality control analysis as well as read trimming.
- fastp: includes most features of FASTQC + Cutadapt + Trimmomatic + AfterQC while running 2–5 times faster than any of them alone.
- After experiment with different parameter values for
  - Sliding window : 4, 5, 8, 10, 12
  - Minimum quality threshold for cutting [cut low quality bases for per read in its 5' and 3' by evaluating the mean quality from a sliding window] : 18, 20, 22, 25, 28
- Chosen sliding window : 10
- Chosen minimum quality threshold : 22

Using fastp with SW 8, MQ 28

SW: Sliding Window
MQ: Minimum Quality Threshold

Fig.3: Summary Quality Table
Before and After Filtering

Before filtering

total reads:	1.165192 M
total bases:	291.298000 M
Q20 bases:	263.803236 M (90.561293%)
Q30 bases:	243.422655 M (83.564822%)
GC content:	50.555203%

After filtering

total reads:	1.165192 M
total bases:	287.736400 M
Q20 bases:	262.268990 M (91.149048%)
Q30 bases:	242.273384 M (84.199769%)
GC content:	50.557496%

SW 8 too low; MQ 28 too high

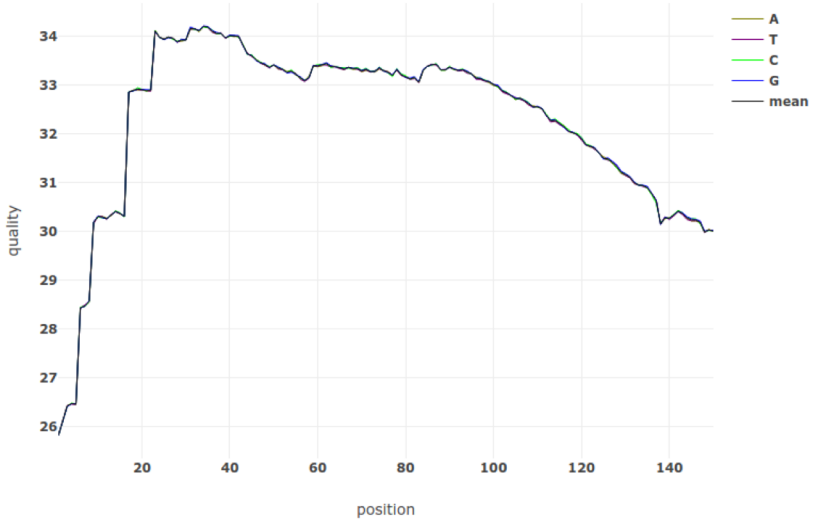


Fig.1: Before Filtering: read 2 quality

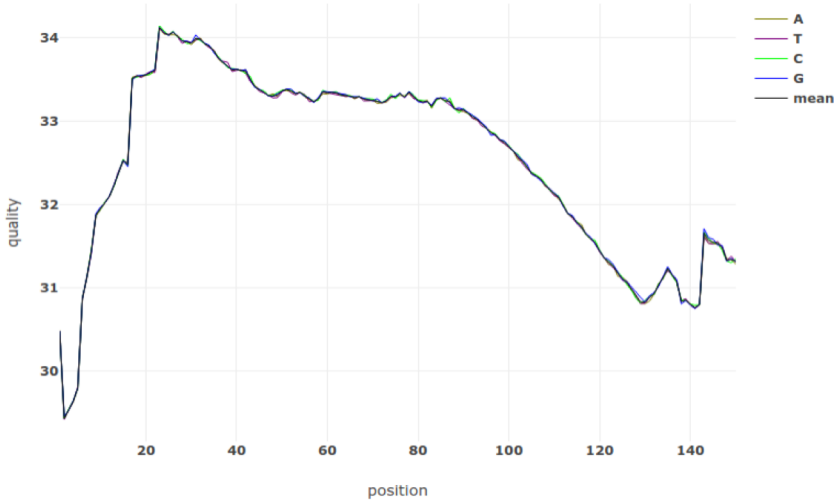


Fig.2: After Filtering: read 2 quality

Using fastp with SW 10, MQ 20

SW: Sliding Window
MQ: Minimum Quality Threshold

SW 10 good; MQ 20 not good

Fig. 3: Summary Quality Table Before and After Filtering

Before filtering

total reads:	1.597948 M
total bases:	239.692200 M
Q20 bases:	226.920460 M (94.671608%)
Q30 bases:	210.268582 M (87.724416%)
GC content:	50.558582%

After filtering

total reads:	1.597948 M
total bases:	232.458974 M
Q20 bases:	220.892243 M (95.024184%)
Q30 bases:	206.983619 M (89.040924%)
GC content:	50.559314%

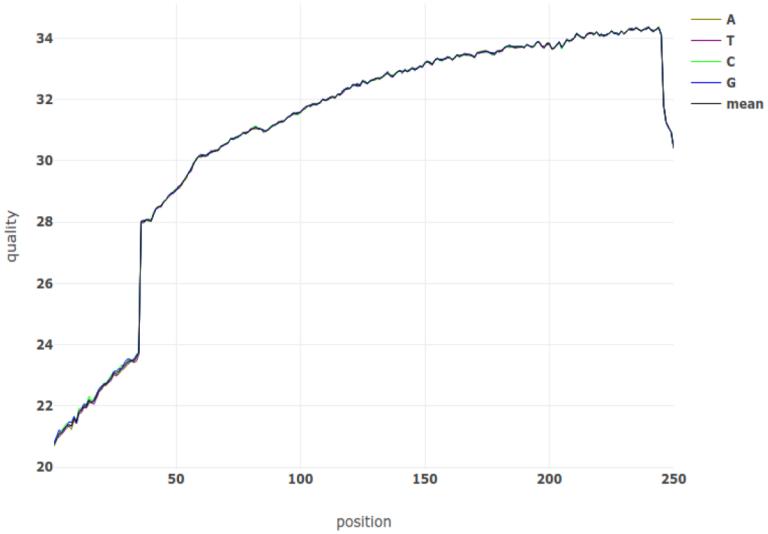


Fig.1: Before Filtering: read 2 quality

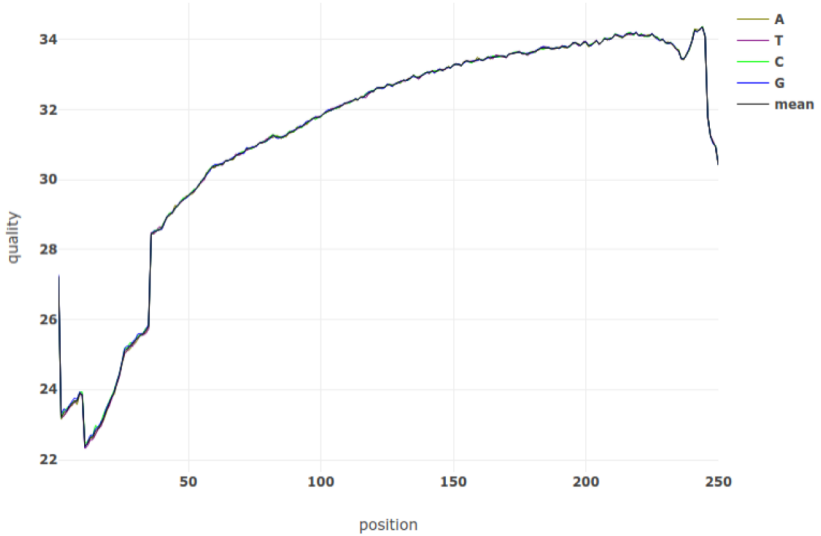


Fig. 2: After Filtering: read 2 quality

Using fastp with SW 10, MQ 22

SW: Sliding Window
MQ: Minimum Quality Threshold

SW 10 good; MQ 22 good

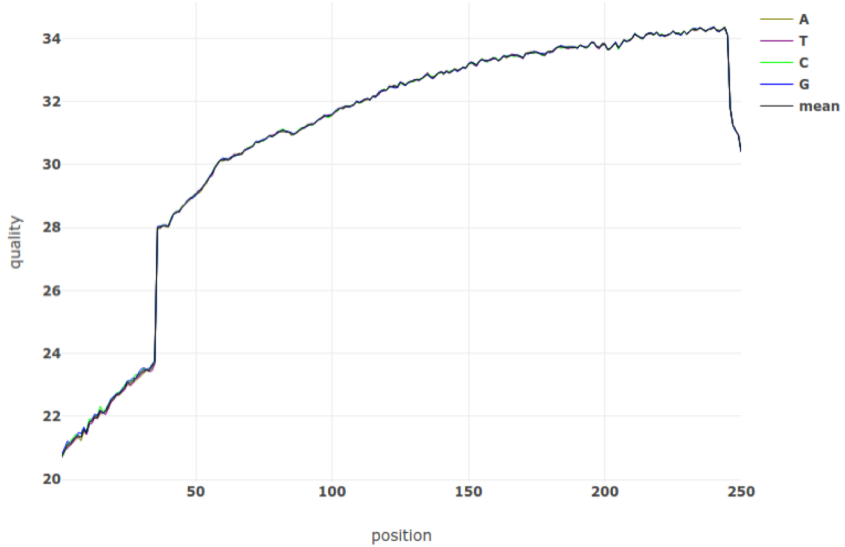


Fig.1: Before Filtering: read 2 quality

Fig. 3: Summary Quality Table Before and After Filtering

Before filtering

total reads:	1.165192 M
total bases:	291.298000 M
Q20 bases:	263.803236 M (90.561293%)
Q30 bases:	243.422655 M (83.564822%)
GC content:	50.555203%

After filtering

total reads:	1.165192 M
total bases:	285.290653 M
Q20 bases:	260.933421 M (91.462310%)
Q30 bases:	241.223586 M (84.553624%)
GC content:	50.558570%

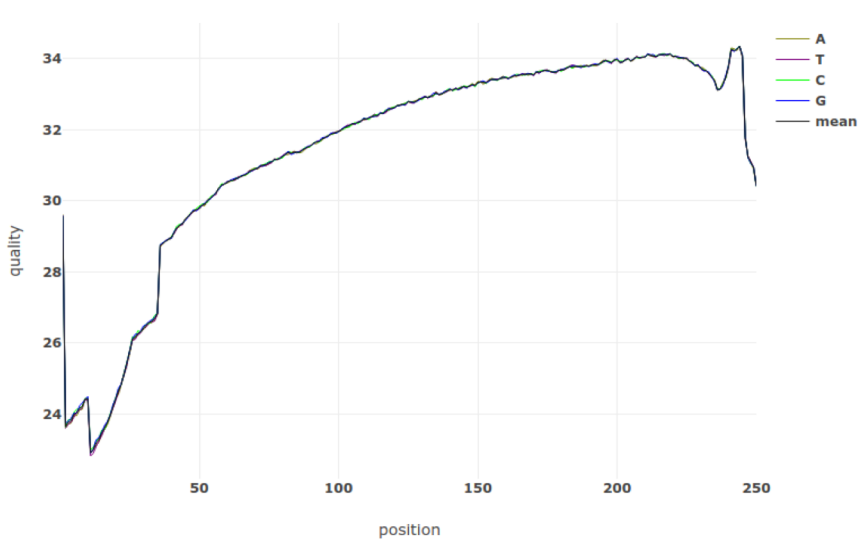
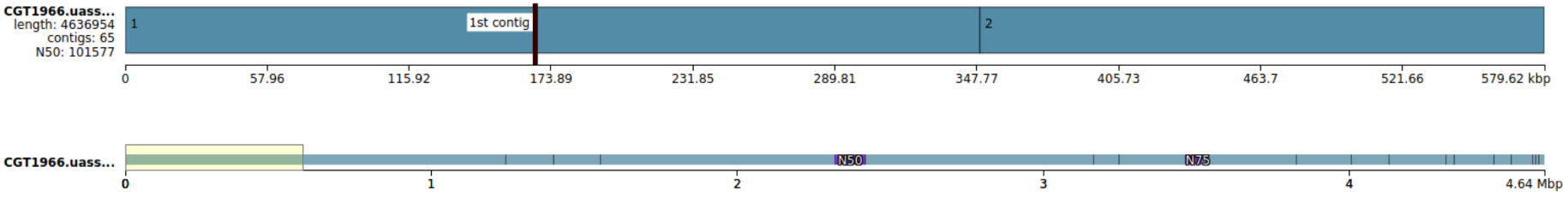
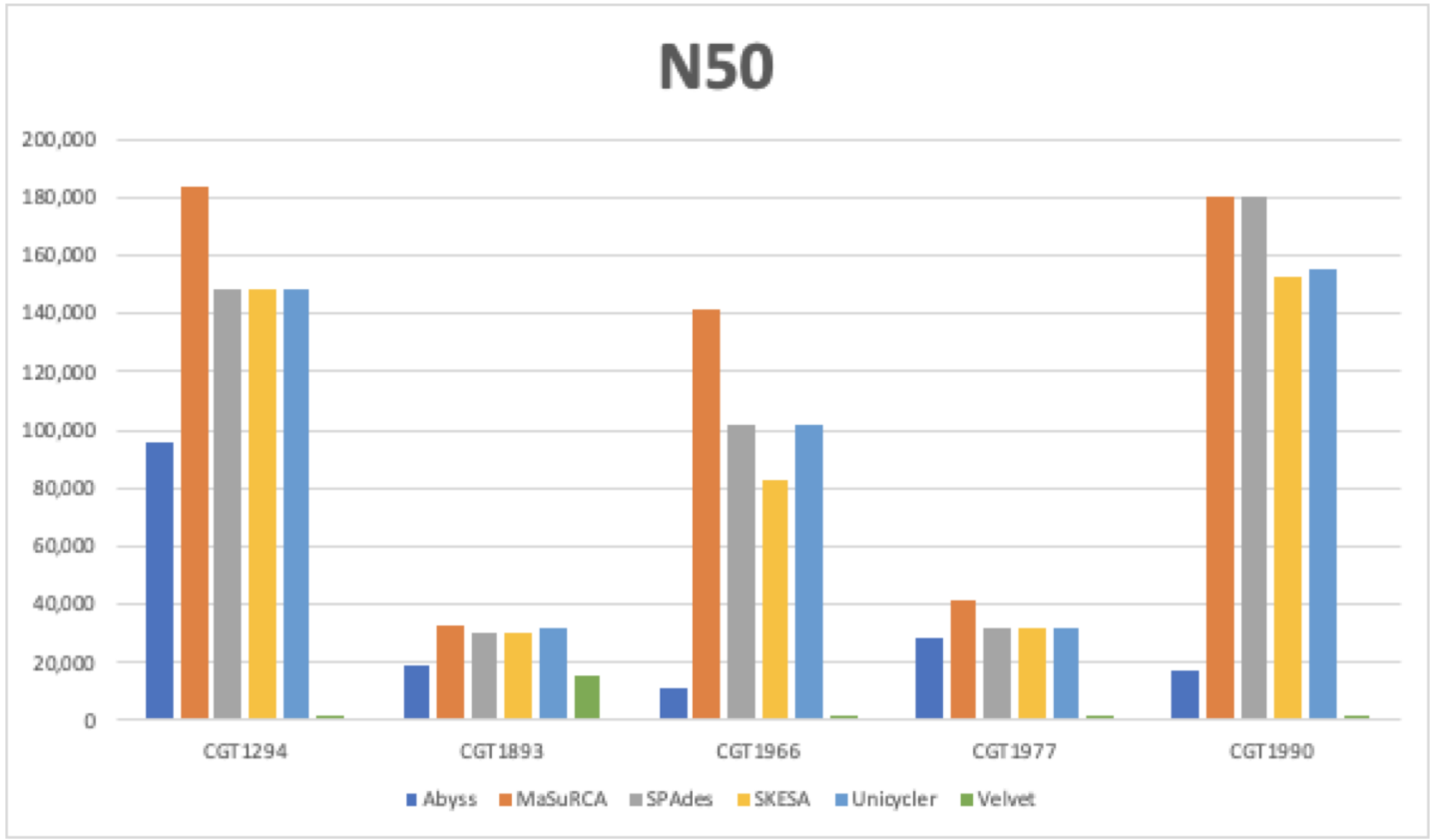


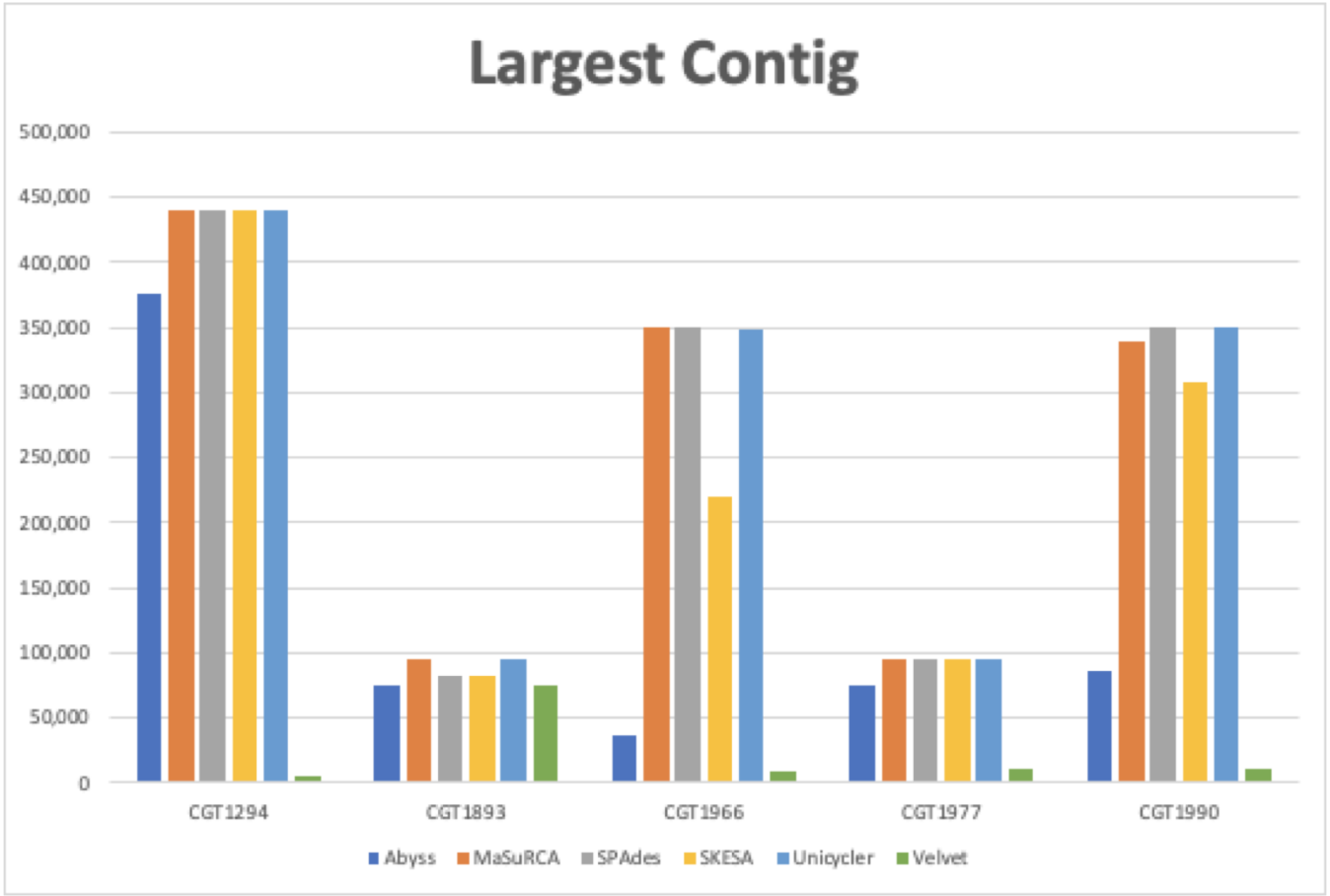
Fig. 2: After Filtering: read 2 quality

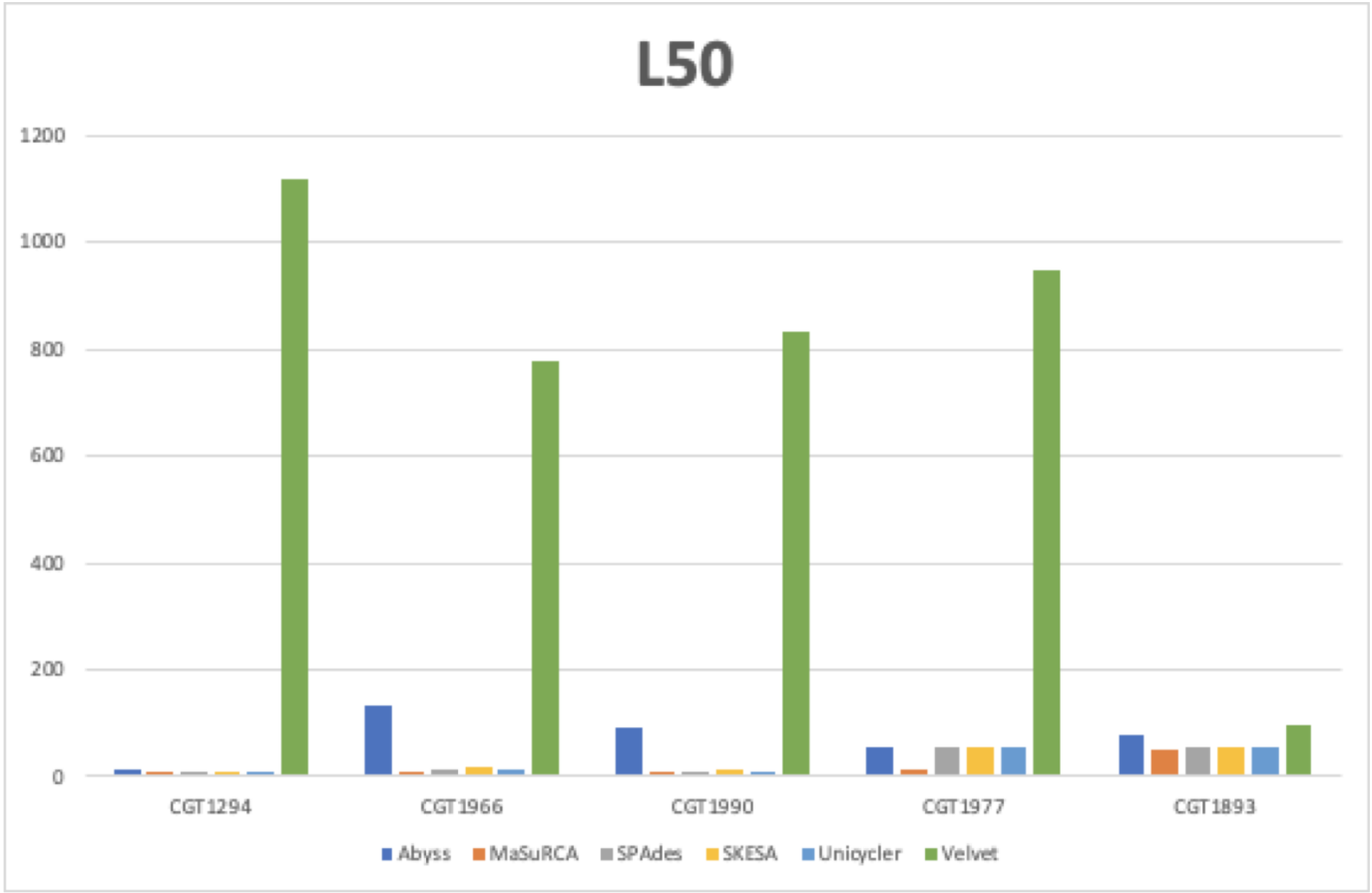
Genome Assembler Evaluation Criteria (QUAST)

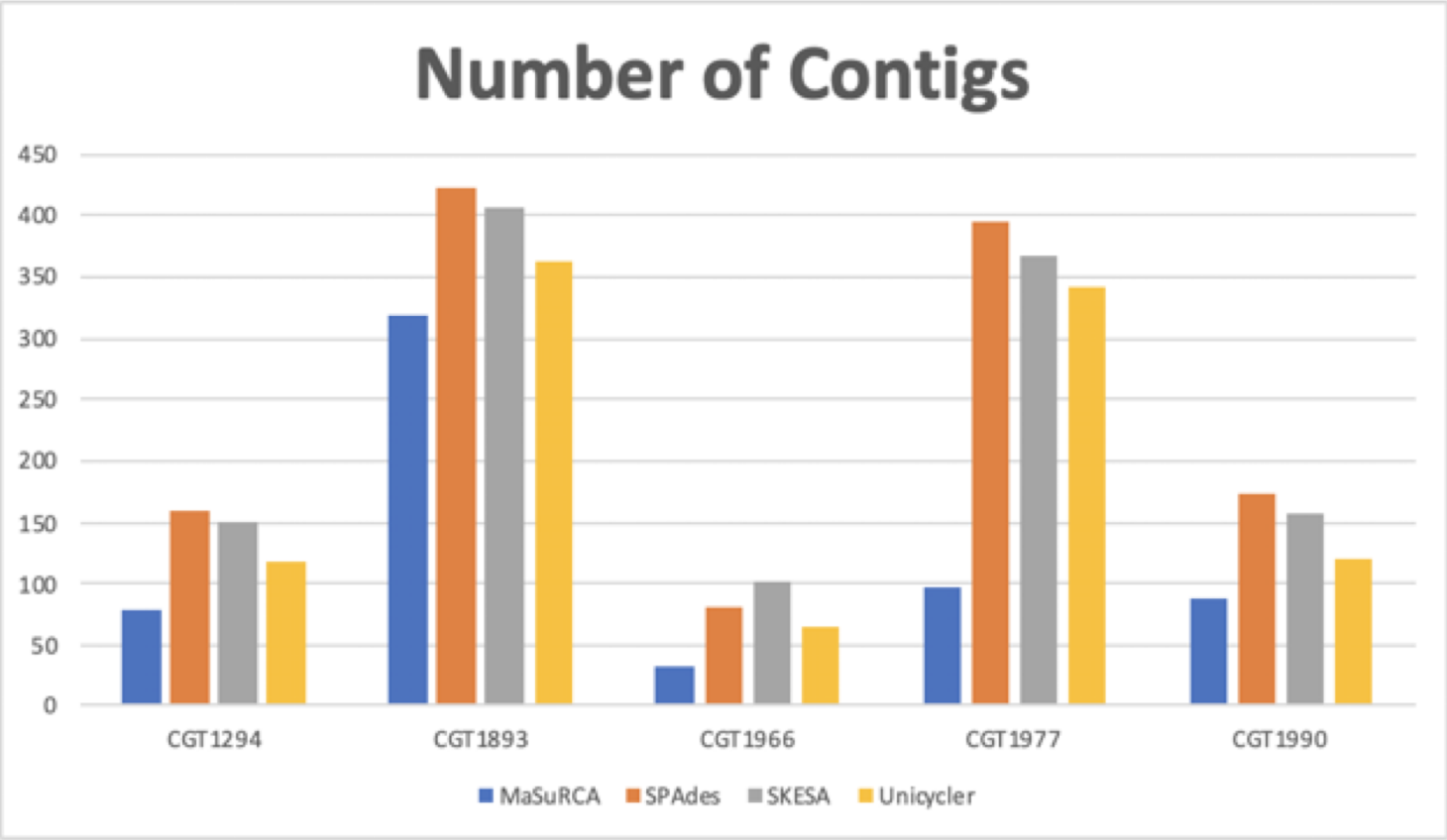
Metric	Description
N50	The minimum contig length crossing the 50% threshold of the total assembled size of the genome.
L50	An assembly is considered to have continuity if it's N90 > 5kb
Assembly Size	The total number of bases in the assembly
Contig statistics	Contigs may be joined into scaffolds or remain unscaffolded. This metric indicates how much of the assembly is represented by scaffolded contigs.

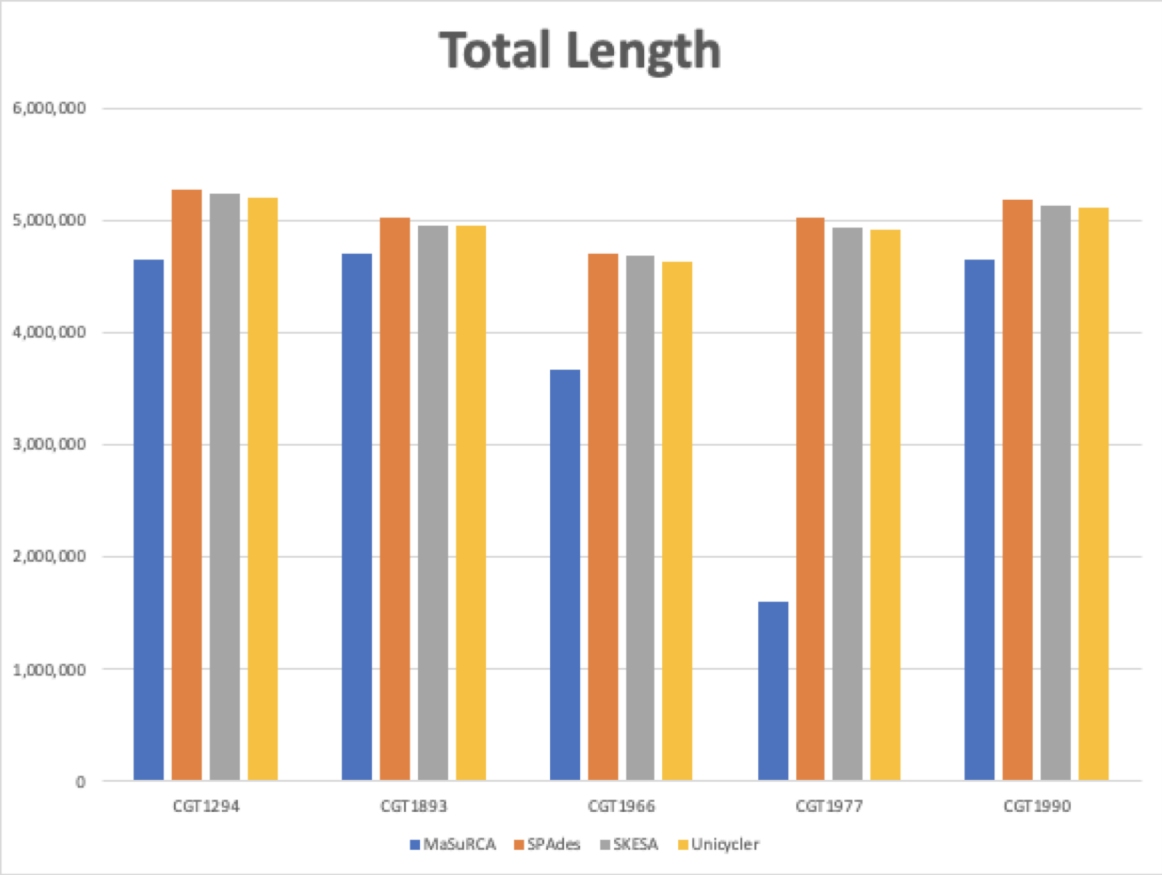



N50
200,000
180,000
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
CGT1294
CGT1893
CGT1966
CGT1977
CGT1990
Abyss
MaSuRCA
SPAdes
SKESA
Unicycler
Velvet


Largest Contig
500,000
450,000
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
CGT1294
CGT1893
CGT1966
CGT1977
CGT1990
Abyss
MaSuRCA
SPAdes
SKESA
Unicycler
Velvet


L50
1200
1000
800
600
400
200
0
CGT1294
CGT1966
CGT1990
CGT1977
CGT1893
Abyss
MaSuRCA
SPAdes
SKESA
Unicycler
Velvet


Number of Contigs
450
400
350
300
250
200
150
100
50
0
CGT1294
CGT1893
CGT1966
CGT1977
CGT1990
MaSuRCA
SPAdes
SKESA
Unicycler

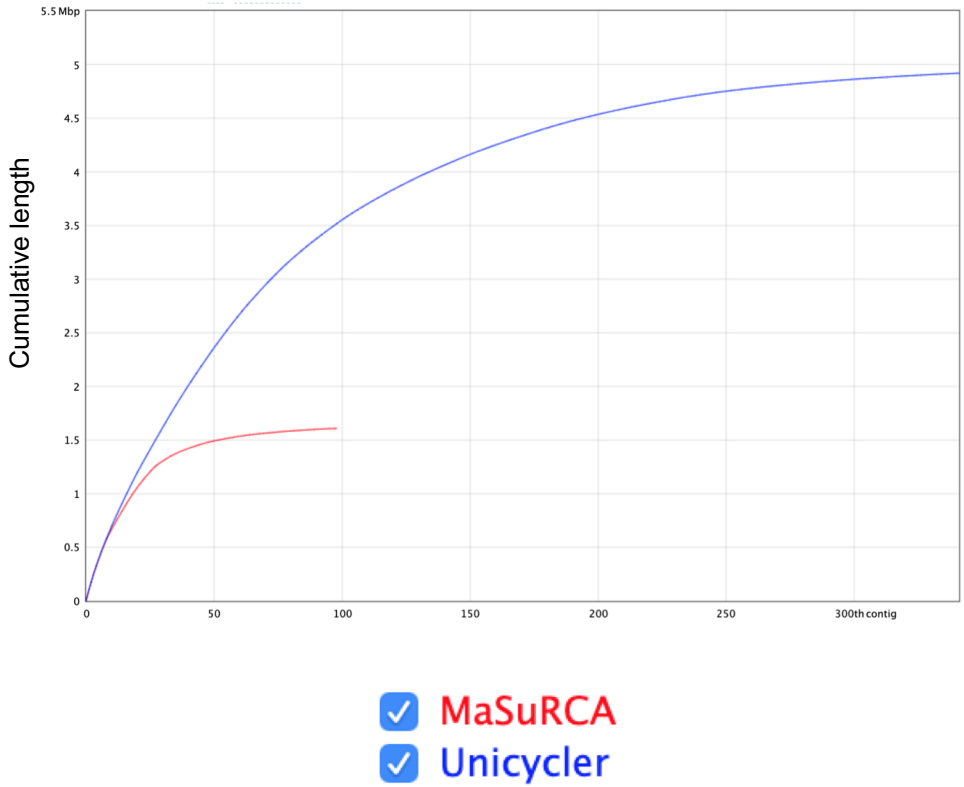

Total Length
6,000,000
5,000,000
4,000,000
3,000,000
2,000,000
1,000,000
0
CGT1294
CGT1893
CGT1966
CGT1977
CGT1990
MaSuRCA
SPAdes
SKESA
Unicycler

MaSuRCa and Unicycler comparison

Worst Median Best

Show heatmap

Statistics without reference	MaSuRCA	Unicycler
# contigs	98	341
# contigs (>= 0 bp)	98	341
# contigs (>= 1000 bp)	94	341
# contigs (>= 5000 bp)	50	212
# contigs (>= 10000 bp)	37	145
# contigs (>= 25000 bp)	26	72
# contigs (>= 50000 bp)	8	15
Largest contig	95 100	94 989
Total length	1 608 937	4 918 958
Total length (>= 0 bp)	1 608 937	4 918 958
Total length (>= 1000 bp)	1 605 848	4 918 958
Total length (>= 5000 bp)	1 492 690	4 598 663
Total length (>= 10000 bp)	1 395 498	4 113 840
Total length (>= 25000 bp)	1 232 255	2 995 458
Total length (>= 50000 bp)	576 292	958 822
N50	41 488	31 980
N75	26 681	14 129
L50	14	54
L75	26	109
GC (%)	51.36	51.01
Mismatches		
# N's	0	0
# N's per 100 kbp	0	0



Final pipeline

- Given the dataset of 50 paired end isolates, our script performs:
  - Quality control and trimming using fastp
  - Genome assembly using Unicycler and MaSuRCa
  - Evaluation of metrics using QUAST
  - Final output on the basis of the ranking into the output folder.

Genome Assembly Pipeline

This pipeline is designed to automate the assembly of a genome with the option to perform quality control prior to assembly and allows the user to pick between the MaSuRCa or Unicycler assemblers or the auto option, which will decide choose the assembler for the user based on quality metrics given by Quast.

Team 1- Genome Assembly

The Genome Assembly group members for Team 1 are:

- Cecilia (Hyeonjeong) Cheon
- Devishi Kesar
- Laura Mora
- Lawrence McKinney
- Jessica Mulligan
- Heather Patrick

Software Requirements

1. fastp (if performing read quality assessment and trimming)
2. MaSuRCa (if choosing MaSuRCa or the auto option for performing genome assembly)
3. Unicycler (if choosing Unicycler or the auto option for performing genome assembly)
4. Quast (tool for quality control metrics)

Usage

Update the paths of the tools downloaded from Software Requirements in the config.txt file prior to running the genome assembly pipeline.

```
./run_genome_assembly_pipeline.sh [-t <int>] -p <dir_path> -o <dir_name> [-q] -g <m|u|a> [-v] [-h]
    -t    number of threads; default is 4
    -p    path to directory containing gzipped fastq forward and backward reads
    -o    path to output directory
    -q    perform quality control and trimming using fastp
    -g    assembler of choice; can pick between MaSuRCa (m), Unicycler (u), or auto (a) (for auto, pipelin
    -v    activate verbose mode
    -h    print usage information
```

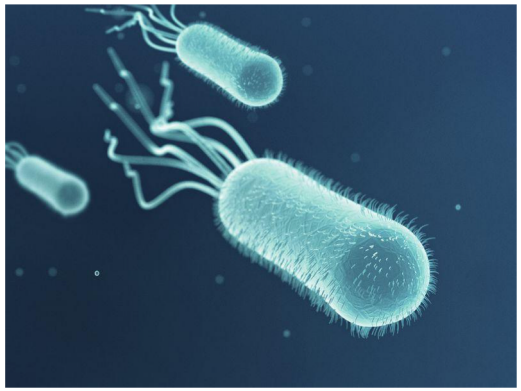
The `auto` option will pick MaSuRCa as the assembler of choice unless the length is more than 10% shorter than Unicycler.

References

- Shifu Chen, Yanqing Zhou, Yaru Chen, Jia Gu; fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics, Volume 34, Issue 17, 1 September 2018, Pages i884–i890, https://doi.org/10.1093/bioinformatics/bty560
- Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013 Nov 1;29(21):2669-77.
- Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13(6): e1005595. https://doi.org/10.1371/journal.pcbi.100559

Identification of pathogen

The identified pathogen for the 50 isolates is ***Escherichia coli***.



BLAST® » blastn suite » results for RID-4P9XZJSY014

Home Recent Results Saved Strategies Help

‹ Edit Search | Save Search | Search Summary

How to read this report? | BLAST Help Videos | Back to Traditional Results Page

Job Title	ctg7180000002909
RID	4P9XZJSY014 *Search expires on 02-19 06:54 am* Download All
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_33663
Description	None
Molecule type	dna
Query Length	95000
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to | E value to | Query Coverage to

Filter Reset

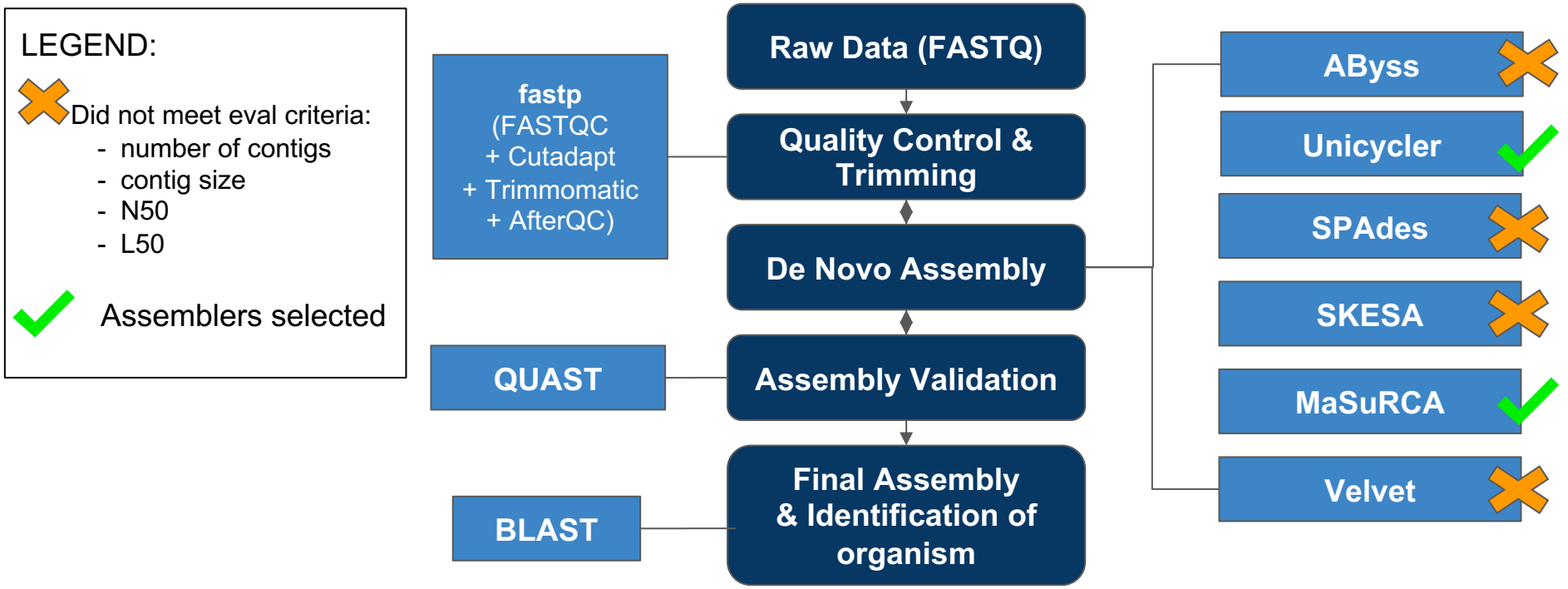
Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Manage Columns Show 100

select all *100 sequences selected* GenBank Graphics Distance tree of results

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Escherichia coli strain PSUO103, complete genome	1.752e+05	1.788e+05	100%	0.0	99.96%	CP014752.1
Escherichia coli strain 2013C-4404 chromosome, complete genome	1.725e+05	1.764e+05	100%	0.0	99.44%	CP027376.1
Escherichia coli strain HB-Coli0, complete genome	1.724e+05	1.762e+05	100%	0.0	99.42%	CP020933.1
Escherichia coli O104:H21 str. CFSAN002236 chromosome, complete genome	1.720e+05	1.757e+05	100%	0.0	99.34%	CP023541.1
Escherichia coli strain 94-3024, complete genome	1.720e+05	1.757e+05	100%	0.0	99.34%	CP009106.2
Escherichia coli isolate EC-TO75 genome assembly, chromosome: 1	1.719e+05	1.758e+05	100%	0.0	99.33%	LS998785.1
Escherichia coli isolate E. coli RL465 genome assembly, chromosome: RL465_chromosome	1.719e+05	1.758e+05	100%	0.0	99.33%	LT594504.1
Escherichia coli strain PT109 chromosome, complete genome	1.719e+05	1.758e+05	100%	0.0	99.33%	CP041031.1
Escherichia coli 042 chromosome, complete genome	1.719e+05	1.758e+05	100%	0.0	99.33%	CP042934.2
Escherichia coli strain AR24.2b chromosome, complete genome	1.719e+05	1.758e+05	100%	0.0	99.33%	CP035944.1

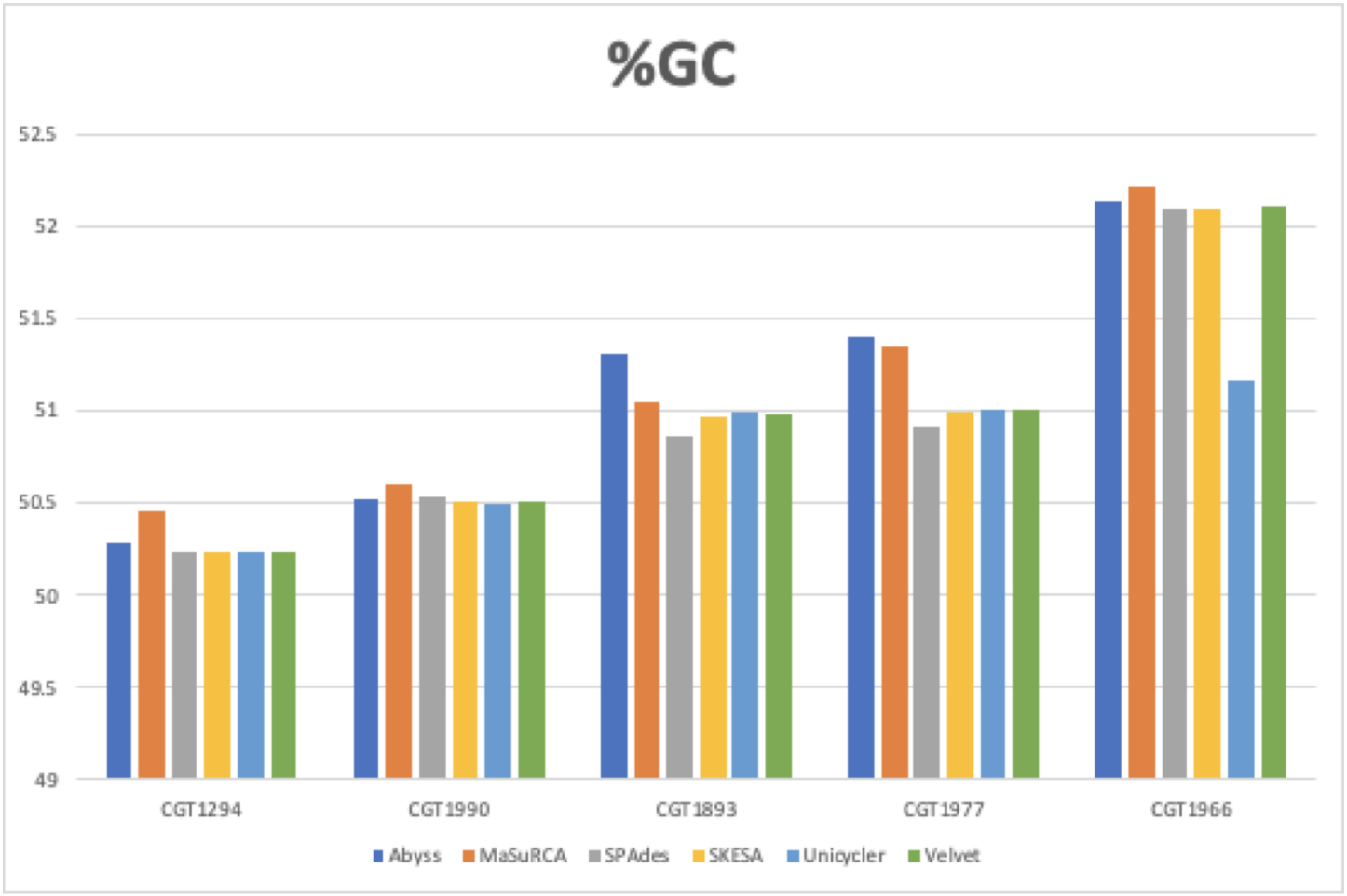
Final workflow

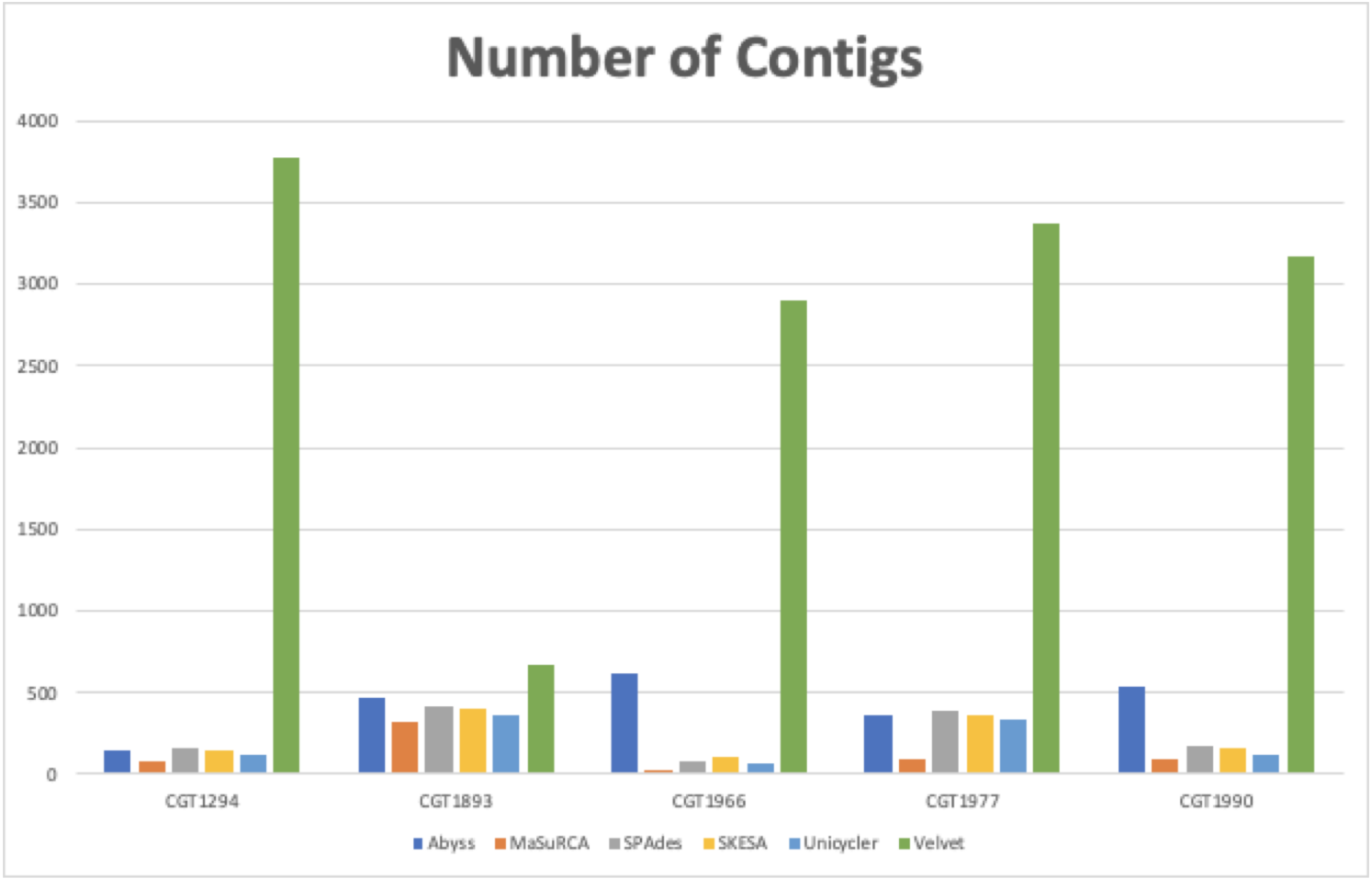


References

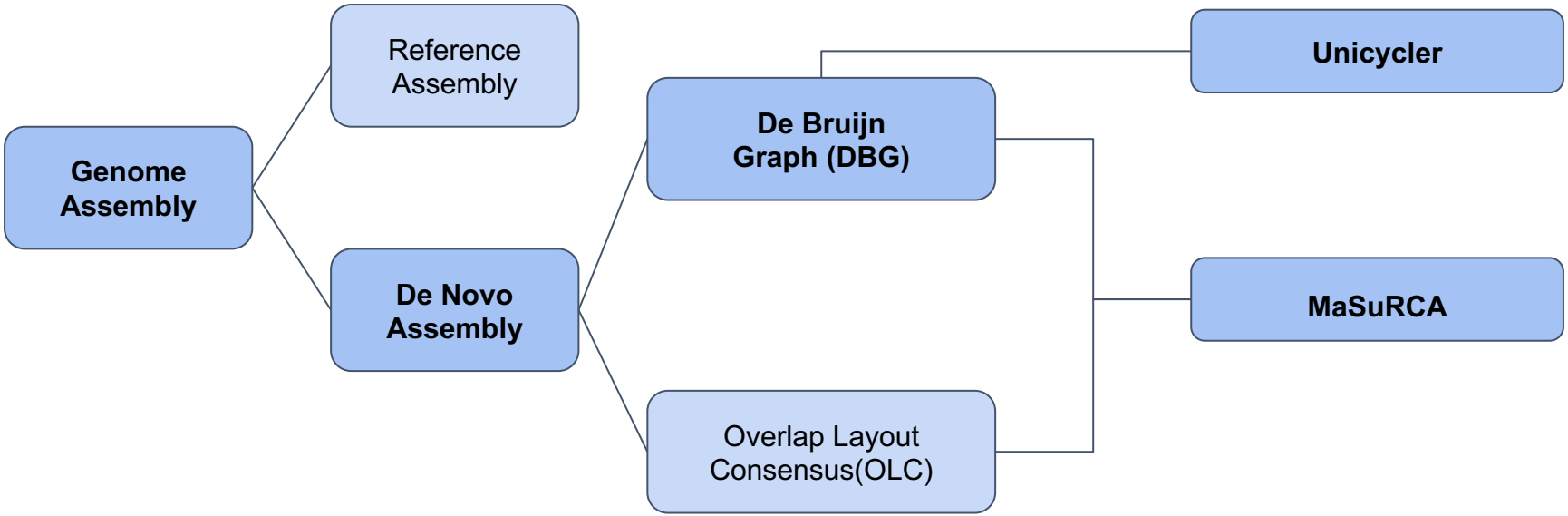
1. Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, Glenn Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics, Volume 29, Issue 8, 15 April 2013, Pages 1072–1075, https://doi.org/10.1093/bioinformatics/btt086
2. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455–477. doi:10.1089/cmb.2012.0021
3. Butler, Jonathan et al. "ALLPATHS: de novo assembly of whole-genome shotgun microreads." *Genome research* vol. 18,5 (2008): 810-20. doi:10.1101/gr.7337908
4. Earl, Dent et al. "Assemblathon 1: a competitive assessment of de novo short read assembly methods." *Genome research* vol. 21,12 (2011): 2224-41. doi:10.1101/gr.126599.111
5. Maccallum, Iain et al. "ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads." *Genome biology* vol. 10,10 (2009): R103. doi:10.1186/gb-2009-10-10-r103
6. Miller, Jason R et al. "Assembly algorithms for next-generation sequencing data." *Genomics* vol. 95,6 (2010): 315-27. doi:10.1016/j.ygeno.2010.03.001
7. Pritt, J., Chen, N. & Langmead, B. FORGe: prioritizing variants for graph genomes. Genome Biol 19, 220 (2018). https://doi.org/10.1186/s13059-018-1595-x
8. Quainoo, S., Coolen, J.P., Hijum, S.A., Huynen, M.A., Melchers, W.J., Schaik, W.V., & Wertheim, H.F. (2017). Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. *Clinical microbiology reviews, 30 4*, 1015-1063 .
9. Rahman, A., Pachter, L. CGAL: computing genome assembly likelihoods. *Genome Biol* 14, R8 (2013). https://doi.org/10.1186/gb-2013-14-1-r8
10. Salzberg, Steven L et al. "GAGE: A critical evaluation of genome assemblies and assembly algorithms." *Genome research* vol. 22,3 (2012): 557-67. doi:10.1101/gr.131383.111
11. Shifu Chen, Yanqing Zhou, Yaru Chen, Jia Gu; fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics, Volume 34, Issue 17, 1 September 2018, Pages i884–i890, https://doi.org/10.1093/bioinformatics/bty560
12. Sohn, Jang-il; Nam, Jin-Wu. "The present and future of de novo whole-genome assembly", Briefings in Bioinformatics, Vol 19.1 (2018). doi.org/10.1093/bib/bbw096
13. Souvorov A., Agarwala R., & Lipman D.J. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biology. 2018; 19(1). doi:10.1186/s13059-018-1540-z
14. Tanja Magoc, Stephan Pabinger, Stefan Canzar, Xinyue Liu, Qi Su, Daniela Puiu, Luke J. Tallon, Steven L. Salzberg, GAGE-B: an evaluation of genome assemblers for bacterial organisms, Bioinformatics, Volume 29, Issue 14, 15 July 2013, Pages 1718–1725, https://doi.org/10.1093/bioinformatics/btt273
15. Zerbino, D., & Birney, E. (n.d.). *Velvet: de novo assembly using very short reads*. Hinxton: European Bioinformatics Institute.

Supplementary slides


%GC
52.5
52
51.5
51
50.5
50
49.5
49
CGT1294
CGT1990
CGT1893
CGT1977
CGT1966
Abyss
MaSuRCA
SPAdes
SKESA
Unicycler
Velvet

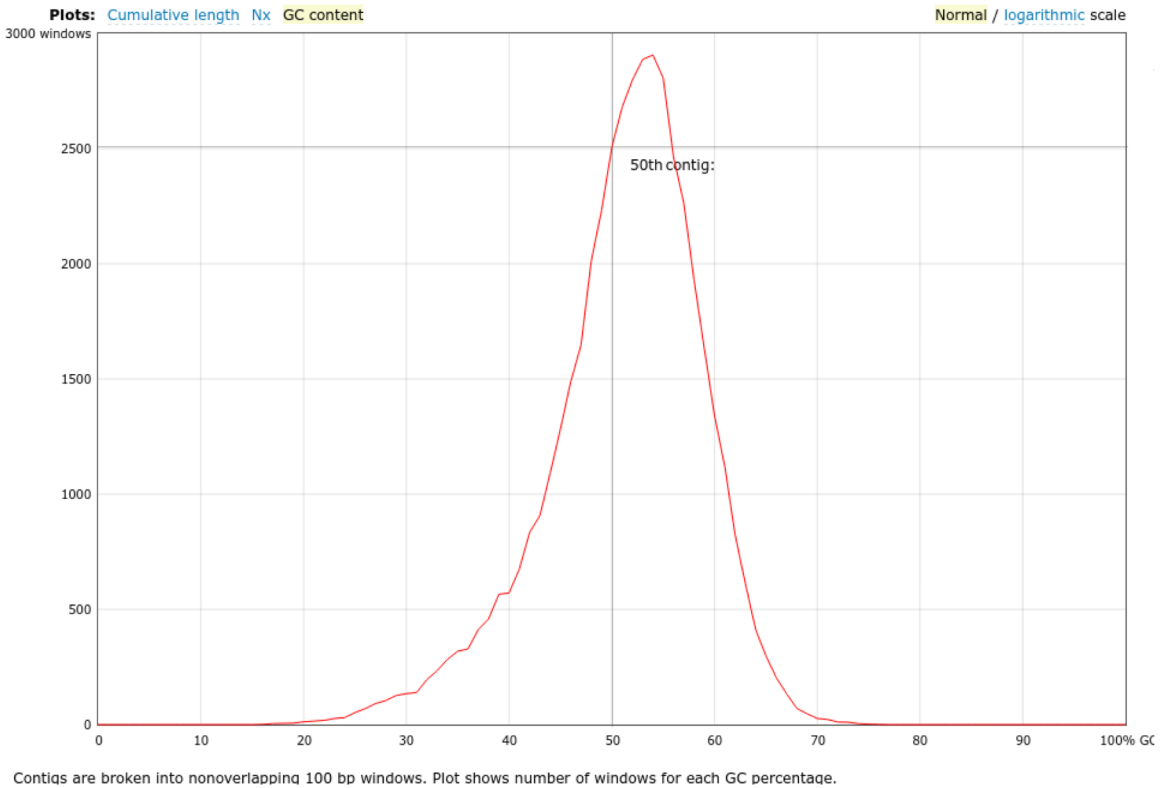

Number of Contigs
4000
3500
3000
2500
2000
1500
1000
500
0
CGT1294
CGT1893
CGT1966
CGT1977
CGT1990
Abyss
MaSuRCA
SPAdes
SKESA
Unicycler
Velvet

Final Pipeline

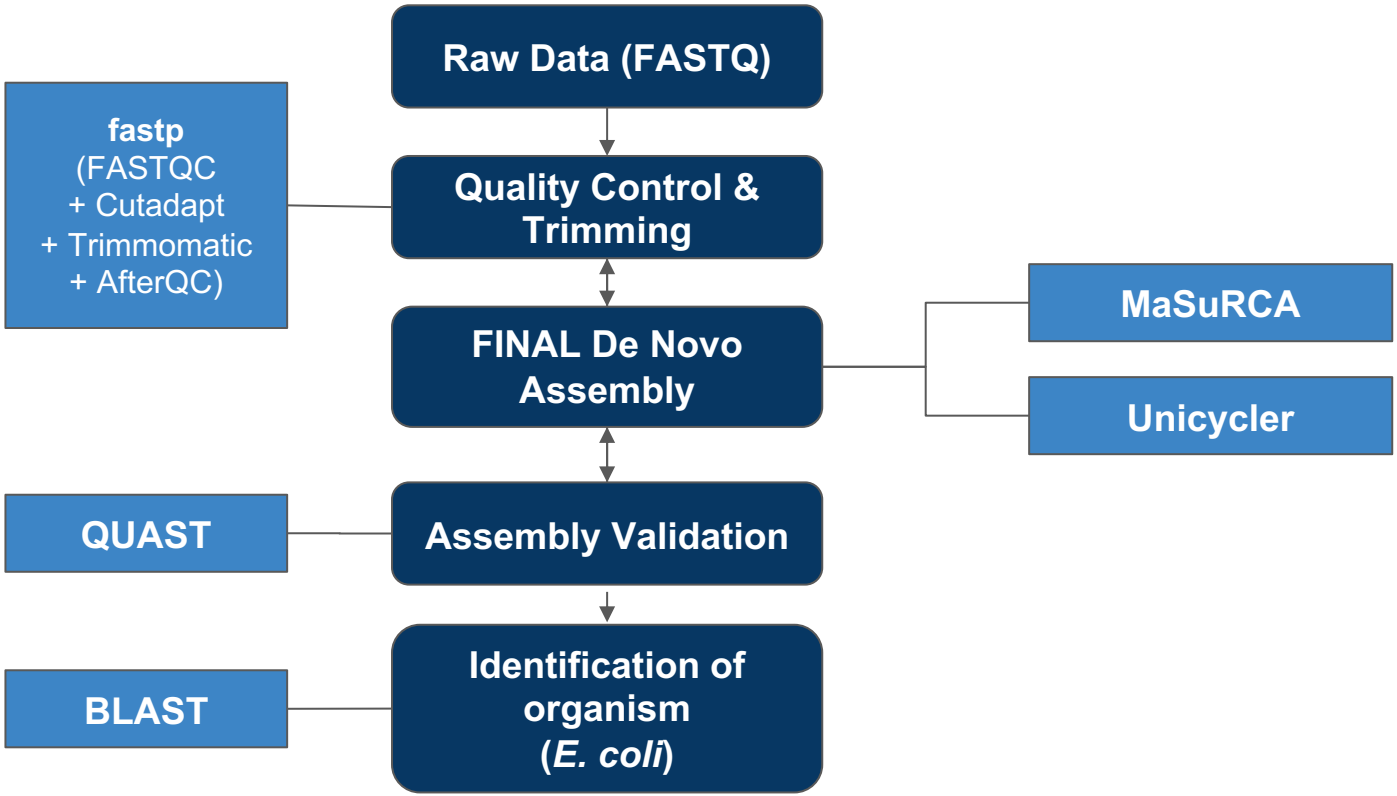


Quality Control Analysis: GC content

Statistics without reference	CGT1966.uassembly
# contigs	65
# contigs (>= 0 bp)	65
# contigs (>= 1000 bp)	65
# contigs (>= 5000 bp)	65
# contigs (>= 10000 bp)	61
# contigs (>= 25000 bp)	48
# contigs (>= 50000 bp)	30
Largest contig	349 043
Total length	4 636 954
Total length (>= 0 bp)	4 636 954
Total length (>= 1000 bp)	4 636 954
Total length (>= 5000 bp)	4 636 954
Total length (>= 10000 bp)	4 598 755
Total length (>= 25000 bp)	4 367 800
Total length (>= 50000 bp)	3 681 233
N50	101 577
N75	58 754
L50	14
L75	27
GC (%)	51.17
Mismatches	
# N's	0
# N's per 100 kbp	0



Final Pipeline



Assemblers Eliminated

Assemblers	**Elimination Criteria**
ALLPATHS-LG	Inappropriate input data
Velvet	Did not meet our evaluation criteria
AbySS	Did not meet our evaluation criteria
Skesa	Underperformed compared to Unicycler and MaSurCa