GENOME ASSEMBLY

LAYLA DREYER

Eschericia coli

- I chose E. coli because working in the health care field I see a lot of bacterial infections due to E. coli

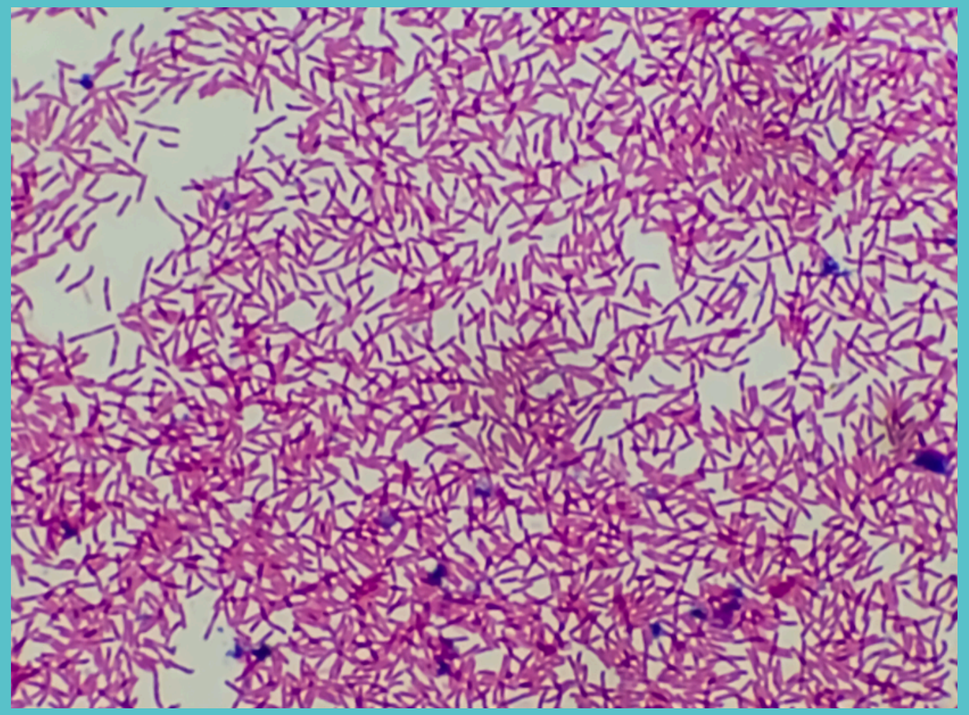

Table #1

contigs	59
scaffolds	78
N50	39089
Genome length	5.05 MBP

SPAdes output

- SPAdes is a genome assembler for genomes or regular and single cell project
- I obtained four files from SPAdes: assembly graph, assembly graph with scaffolds, contigs, scaffolds
- These are the files we will use to run a quality assessment

Reference genome

- I obtained a reference genome from NCBI
- During genome assembly a reference genome is a helpful tool to increase accuracy and be used almost as a “template” to compare the genome to and see where everything should fit

QUAST

We performed a quality assessment which gave us a report with a lot of information on the genome

Report

	SPAdes_on_data_8_and_data_7__Scaffolds	SPAdes_on_data_8_and_data_7__Scaffolds_broken
# contigs (>= 0 bp)	78	-
# contigs (>= 1000 bp)	46	52
Total length (>= 0 bp)	5053809	-
Total length (>= 1000 bp)	5040390	5038607
# contigs	59	67
Largest contig	570518	506009
Total length	5049193	5048393
Reference length	4641652	4641652
GC (%)	50.61	50.61
Reference GC (%)	50.79	50.79
N50	390869	312376
NG50	408096	312376
N90	94800	88270
NG90	150929	130956
auN	343882.1	276231.0
auNG	374075.2	300436.8
L50	6	7
LG50	5	7
L90	15	19
LG90	12	16
# misassemblies	120	113
# misassembled contigs	16	19
Misassembled contigs length	4640133	4483792
# local misassemblies	66	65
# scaffold gap ext. mis.	0	-
# scaffold gap loc. mis.	4	-
# unaligned mis. contigs	3	4
# unaligned contigs	16 + 33 part	17 + 36 part
Unaligned length	1013298	1012949
Genome fraction (%)	87.507	87.911
Duplication ratio	1.004	1.004
# N's per 100 kbp	15.84	0.00
# mismatches per 100 kbp	2239.37	2242.43
# indels per 100 kbp	44.54	45.29
# genomic features	7950 + 320 part	7953 + 341 part
Complete BUSCO (%)	98.65	98.65
Partial BUSCO (%)	0.00	0.00
# predicted rRNA genes	9 + 1 part	8 + 2 part
Largest alignment	188758	188758
Total aligned length	4034211	4033800
NA50	37783	37213
NGA50	44299	41861
NA90	-	-
NGA90	-	-
auNA	51347.4	50328.5
auNGA	55855.8	54738.7
LA50	35	36
LGA50	30	31
LA90	-	-
LGA90	-	-

QUAST

Some things to note

- BUSCO 98.65%
- GC % 50.61

Unaligned length	1013298	1012949
Genome fraction (%)	87.507	87.911
Duplication ratio	1.004	1.004
# N's per 100 kbp	15.84	0.00
# mismatches per 100 kbp	2239.37	2242.43
# indels per 100 kbp	44.54	45.29
# genomic features	7950 + 320 part	7953 + 341 part
Complete BUSCO (%)	98.65	98.65
Partial BUSCO (%)	0.00	0.00
# predicted rRNA genes	9 + 1 part	8 + 2 part
Largest alignment	188758	188758
Total aligned length	4034211	4033800

Report

	SPAdes_on_data_8_and_data_7__Scaffolds	SPAdes_on_data_8_and_data_7__Scaffolds_broken
# contigs (>= 0 bp)	78	-
# contigs (>= 1000 bp)	46	52
Total length (>= 0 bp)	5053809	-
Total length (>= 1000 bp)	5040390	5038607
# contigs	59	67
Largest contig	570518	506009
Total length	5049193	5048393
Reference length	4641652	4641652
GC (%)	50.61	50.61
Reference GC (%)	50.79	50.79
N50	390869	312376

Prokka results

- Prokka is used to annotate genes and identify protein coding regions
- In E. coli the genese found in the prokka results dnaA, recA, and 16S rRna

THANK YOU!