Brendan Kearney

Feb 3, 2019

BINF 6203 Genomics

Lab 2: Next Generation Sequencing Quality Control

INTRODUCTION

Sequence assembly in genomics is the process of putting fragmented DNA sequences back together. The process varies heavily in time and scope based on many factors, including the quality of the sequences and the complexity of the organism's genome. For the assembly of de novo sequences, the base building block are strings of nucleotides, k-mers, which can be assigned to nodes of *de Bruijn* graphs. One such assembler named SPAdes can use *de Bruijn* algorithms on small bacterial sequences to produce paired assembly graphs. In this lab, we used SPAdes on a chloroplast genome and a few *E. coli* sequence reads. The program QUAST was then used to analyze the results of the assembly.

METHODS

The following files were retrieved via web download:

- Ion Torrent Chloroplast genome sample (BC30)
- Paired-end *E.coli* Illumina sequences (ERR008613/ERR022075)
- PacBio CCS and CLR *E. coli* reads

The files were then transferred to HPC cluster "mamba".

Assembly

The following UNIX commands were submitted using qsub scripts. The default options listed in the first set of commands are only detailed once for the sake of simplicity.

- *BC30.sh*

```
#!/bin/bash

#NS  =======
#PBS -q mamba
#PBS -N cjg.sp
#PBS -l nodes=3:ppn=6
#PBS -l walltime=24:00:00
#PBS -l
prologue=/users/bkearney/torque/prologue.sh,epilogue=/users/bkearney/t
orque/epilogue.sh
# ===== END PBS OPTIONS =====
module load spades

spades.py -o BC30spades --iontorrent -k 21,33,55,77,99,121 -s
BC30.fastq
```

- *BC30_kmers.sh*

```
module load spades

spades.py -o BC30_kmers_spades --iontorrent -k 21,33,55,77 -s
BC30.fastq
```

- *BC30_careful.sh*

```
module load spades

spades.py -o BC30_careful_spades -iontorrent -k 21,33,55,77,99,121 --
careful -s BC30.fastq
```

- *BC30_noiontorrent.sh*

```
module load spades

spades.py -o BC30_noiontorrent_spades -k 21,33,55,77,99,121 -s
BC30.fastq
```

- *Ecoli.sh*

```
module load spades

spades.py --pe1-1 ERR008613sample_1.fastq --pe1-2
ERR008613sample_2.fastq --pe2-1 ERR022075sample_1.fastq --pe2-2
ERR022075sample_2.fastq -o ecoli_output
```

- *Ecoli_pacbio.sh*

```
module load spades

spades.py --pe1-1 ERR008613sample_1.fastq --pe1-2
ERR008613sample_2.fastq --pe2-1 ERR022075sample_1.fastq --pe2-2
ERR022075sample_2.fastq -o ecoli_pacbio_output --pacbio
PacBio_10kb_CLR.fastq
```

- *Ecoli_pacbio2.sh*

```
spades.py --pe1-1 ERR008613sample_1.fastq --pe1-2
ERR008613sample_2.fastq --pe2-1 ERR022075sample_1.fastq --pe2-2
```

```
ERR022075sample_2.fastq -o ecoli_pacbio2_output --pacbio
PacBio_10kb_CLR.fastq -s PacBio_2kb_CCS_500bp.fastq
```

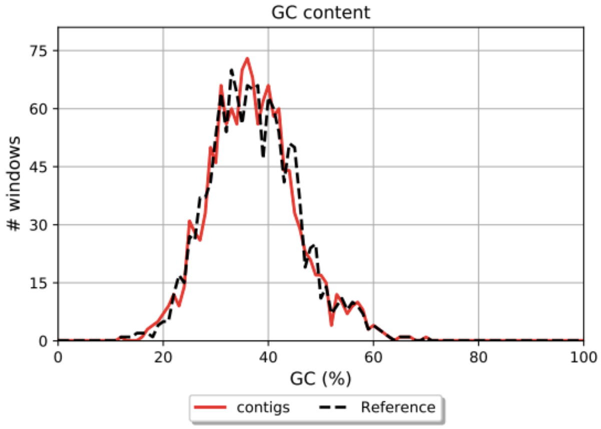
Analysis

The Quast tool was used to evaluate the assemblies.

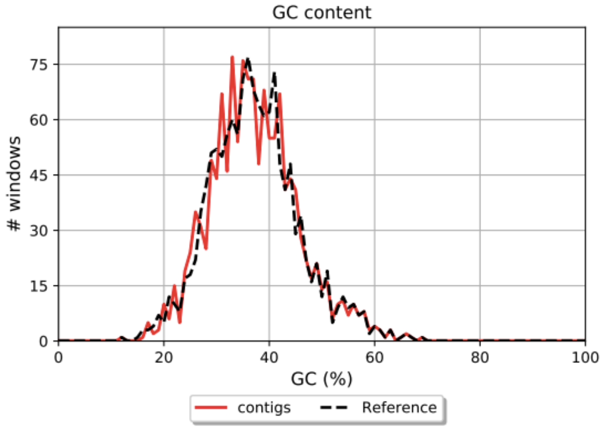
```
quast.py /Users/bkearney/BC30spades/contigs.fasta -r
/Users/bkearney/BC30spades/before_rr.fasta -g
/Users/bkearney/Downloads/assemblyfiles/references/NC_007898.gff --
min-contig 250 -o BC30_quast

quast.py /Users/bkearney/BC30_kmers_spades/contigs.fasta -r
/Users/bkearney/BC30_kmers_spades/before_rr.fasta -g
/Users/bkearney/Downloads/assemblyfiles/references/NC_007898.gff --
min-contig 250 -o BC30_kmers_quast

quast.py /Users/bkearney/BC30_careful_spades/contigs.fasta -r
/Users/bkearney/BC30_careful_spades/before_rr.fasta -g
/Users/bkearney/Downloads/assemblyfiles/references/NC_007898.gff --
min-contig 250 -o BC30_careful_quast

quast.py
/Users/bkearney/BC30_noiontorrent_spades/K21/final_contigs.fasta -r
/Users/bkearney/BC30_noiontorrent_spades/K21/before_rr.fasta -g
/Users/bkearney/Downloads/assemblyfiles/references/NC_007898.gff --
min-contig 250 -o BC30_noiontorrent_quast

quast.py /Users/bkearney/ecoli/ecoli_output/K55/final_contigs.fasta -r
/Users/bkearney/ecoli/ecoli_output/K55/before_rr.fasta -g
/Users/bkearney/Downloads/assemblyfiles/references/NC_000913.gff --
min-contig 250 -o ecoli_quast

quast.py
/Users/bkearney/ecoli/ecoli_pacbio_output/K55/final_contigs.fasta -r
/Users/bkearney/ecoli/ecoli_pacbio_output/K55/before_rr.fasta -g
/Users/bkearney/Downloads/assemblyfiles/references/NC_000913.gff --
min-contig 250 -o ecoli_quast_pacbio

quast.py /Users/bkearney/ecoli_pacbio2_output/K21/final_contigs.fasta
-r /Users/bkearney/ecoli_pacbio2_output/K21/before_rr.fasta -g
/Users/bkearney/Downloads/assemblyfiles/references/NC_000913.gff --
min-contig 250 -o ecoli_quast_pacbio2
```

RESULTS

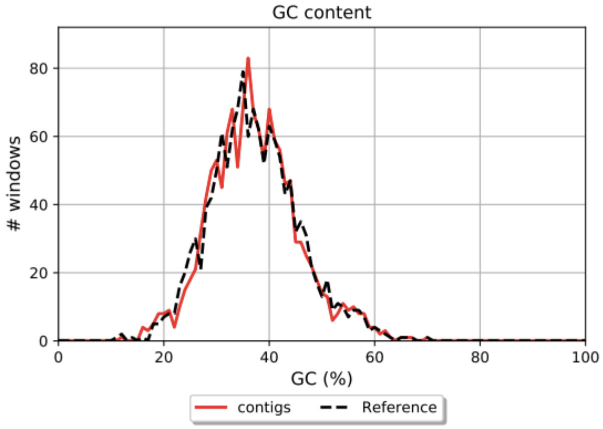
BC30



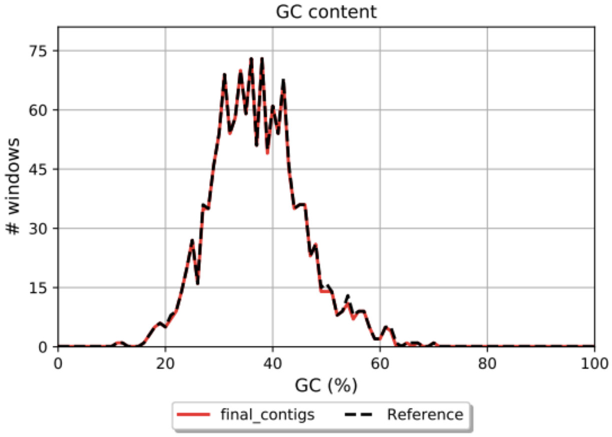
BC30_kmers (less kmer lengths)



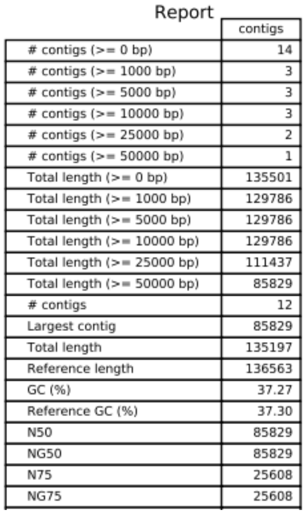
BC30_careful



BC30_noiontorrent



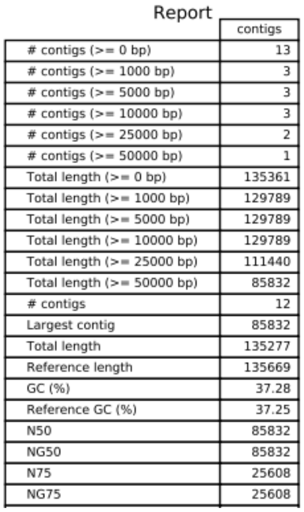
BC30



Report

	contigs
# contigs (>= 0 bp)	14
# contigs (>= 1000 bp)	3
# contigs (>= 5000 bp)	3
# contigs (>= 10000 bp)	3
# contigs (>= 25000 bp)	2
# contigs (>= 50000 bp)	1
Total length (>= 0 bp)	135501
Total length (>= 1000 bp)	129786
Total length (>= 5000 bp)	129786
Total length (>= 10000 bp)	129786
Total length (>= 25000 bp)	111437
Total length (>= 50000 bp)	85829
# contigs	12
Largest contig	85829
Total length	135197
Reference length	136563
GC (%)	37.27
Reference GC (%)	37.30
N50	85829
NG50	85829
N75	25608
NG75	25608

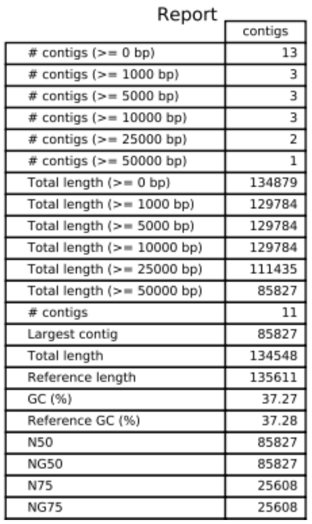
BC30_kmers



Report

	contigs
# contigs (>= 0 bp)	13
# contigs (>= 1000 bp)	3
# contigs (>= 5000 bp)	3
# contigs (>= 10000 bp)	3
# contigs (>= 25000 bp)	2
# contigs (>= 50000 bp)	1
Total length (>= 0 bp)	135361
Total length (>= 1000 bp)	129789
Total length (>= 5000 bp)	129789
Total length (>= 10000 bp)	129789
Total length (>= 25000 bp)	111440
Total length (>= 50000 bp)	85832
# contigs	12
Largest contig	85832
Total length	135277
Reference length	135669
GC (%)	37.28
Reference GC (%)	37.25
N50	85832
NG50	85832
N75	25608
NG75	25608

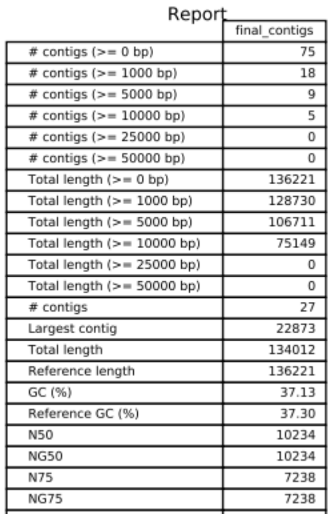
BC30_careful



Report

	contigs
# contigs (>= 0 bp)	13
# contigs (>= 1000 bp)	3
# contigs (>= 5000 bp)	3
# contigs (>= 10000 bp)	3
# contigs (>= 25000 bp)	2
# contigs (>= 50000 bp)	1
Total length (>= 0 bp)	134879
Total length (>= 1000 bp)	129784
Total length (>= 5000 bp)	129784
Total length (>= 10000 bp)	129784
Total length (>= 25000 bp)	111435
Total length (>= 50000 bp)	85827
# contigs	11
Largest contig	85827
Total length	134548
Reference length	135611
GC (%)	37.27
Reference GC (%)	37.28
N50	85827
NG50	85827
N75	25608
NG75	25608

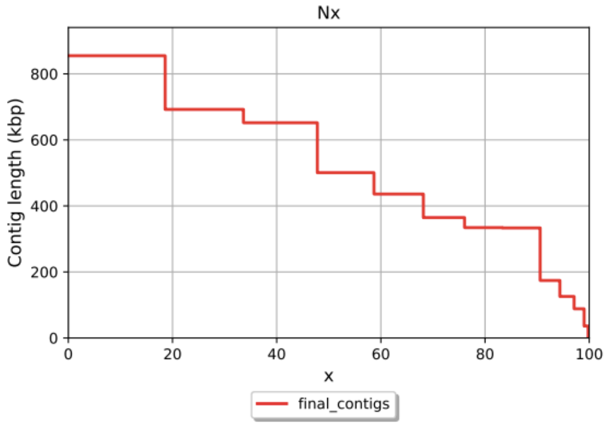
BC30_noiontorrent



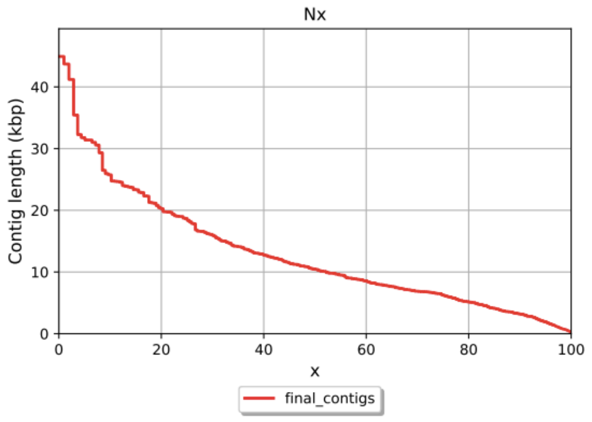
Report

	final_contigs
# contigs (>= 0 bp)	75
# contigs (>= 1000 bp)	18
# contigs (>= 5000 bp)	9
# contigs (>= 10000 bp)	5
# contigs (>= 25000 bp)	0
# contigs (>= 50000 bp)	0
Total length (>= 0 bp)	136221
Total length (>= 1000 bp)	128730
Total length (>= 5000 bp)	106711
Total length (>= 10000 bp)	75149
Total length (>= 25000 bp)	0
Total length (>= 50000 bp)	0
# contigs	27
Largest contig	22873
Total length	134012
Reference length	136221
GC (%)	37.13
Reference GC (%)	37.30
N50	10234
NG50	10234
N75	7238
NG75	7238

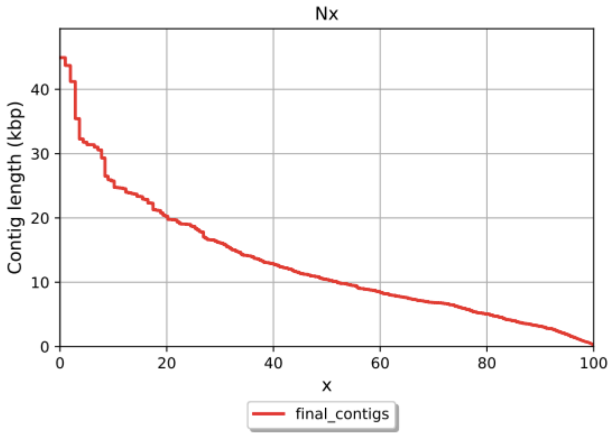
Ecoli_quast



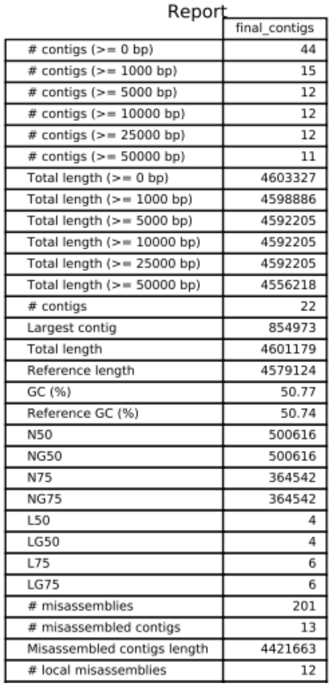
Ecoli_pacbio



Ecoli_pacbio2



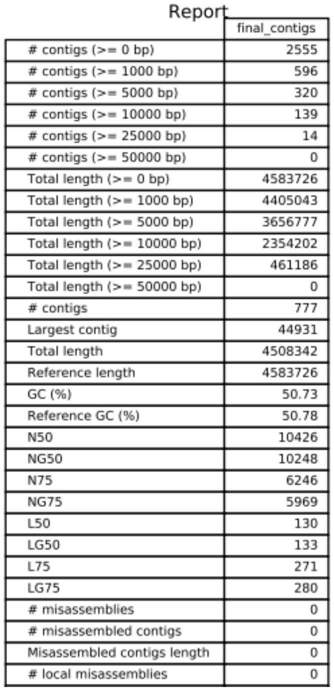
Ecoli_quast



Report

	final_contigs
# contigs (>= 0 bp)	44
# contigs (>= 1000 bp)	15
# contigs (>= 5000 bp)	12
# contigs (>= 10000 bp)	12
# contigs (>= 25000 bp)	12
# contigs (>= 50000 bp)	11
Total length (>= 0 bp)	4603327
Total length (>= 1000 bp)	4598886
Total length (>= 5000 bp)	4592205
Total length (>= 10000 bp)	4592205
Total length (>= 25000 bp)	4592205
Total length (>= 50000 bp)	4556218
# contigs	22
Largest contig	854973
Total length	4601179
Reference length	4579124
GC (%)	50.77
Reference GC (%)	50.74
N50	500616
NG50	500616
N75	364542
NG75	364542
L50	4
LG50	4
L75	6
LG75	6
# misassemblies	201
# misassembled contigs	13
Misassembled contigs length	4421663
# local misassemblies	12

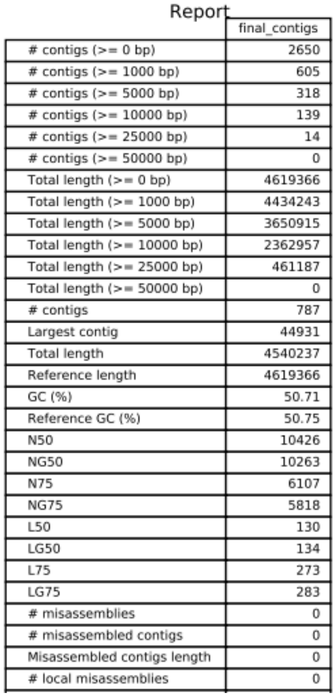
Ecoli_pacbio



Report

	final_contigs
# contigs (>= 0 bp)	2555
# contigs (>= 1000 bp)	596
# contigs (>= 5000 bp)	320
# contigs (>= 10000 bp)	139
# contigs (>= 25000 bp)	14
# contigs (>= 50000 bp)	0
Total length (>= 0 bp)	4583726
Total length (>= 1000 bp)	4405043
Total length (>= 5000 bp)	3656777
Total length (>= 10000 bp)	2354202
Total length (>= 25000 bp)	461186
Total length (>= 50000 bp)	0
# contigs	777
Largest contig	44931
Total length	4508342
Reference length	4583726
GC (%)	50.73
Reference GC (%)	50.78
N50	10426
NG50	10248
N75	6246
NG75	5969
L50	130
LG50	133
L75	271
LG75	280
# misassemblies	0
# misassembled contigs	0
Misassembled contigs length	0
# local misassemblies	0

Ecoli_pacbio2



Report

	final_contigs
# contigs (>= 0 bp)	2650
# contigs (>= 1000 bp)	605
# contigs (>= 5000 bp)	318
# contigs (>= 10000 bp)	139
# contigs (>= 25000 bp)	14
# contigs (>= 50000 bp)	0
Total length (>= 0 bp)	4619366
Total length (>= 1000 bp)	4434243
Total length (>= 5000 bp)	3650915
Total length (>= 10000 bp)	2362957
Total length (>= 25000 bp)	461187
Total length (>= 50000 bp)	0
# contigs	787
Largest contig	44931
Total length	4540237
Reference length	4619366
GC (%)	50.71
Reference GC (%)	50.75
N50	10426
NG50	10263
N75	6107
NG75	5818
L50	130
LG50	134
L75	273
LG75	283
# misassemblies	0
# misassembled contigs	0
Misassembled contigs length	0
# local misassemblies	0

DISCUSSION

For the tomato chloroplast data (**BC30**), the most important SPAdes parameter was the special indicator for Ion Torrent sequences (--iontorrent). Without it (**BC30_noiontorrent**), the results were drastically different, with number of contigs of shorter base pair lengths being substantially higher. In other assemblies with the command, most results resembled the expected qualities of the data, with rough coverage and sequencing inconsistencies. Two separate sets of k-mer sizes were tried based on the recommended Ion Torrent k values (**BC30 and BC30_kmers**), which yielded almost identical contig numbers and Nx statistics. Of the four assemblies, the only one which had any misassembled contigs was the first **BC30** assembly. In other assemblies, these misassemblies could've been removed by the --careful command or the adjusted k-mer lengths. The most substantial changes were in the GC content plots, which show the number of non-overlapping bp windows distributed over the percentage of guanine and cytosine content in the bases.

In the *E. coli* assemblies, the main feature that was compared was the effect of PacBio CSS and CLR reads. Using the commands for forward and reverse paired-end reads, the *E. coli* sequences were assembled first without the PacBio files (**Ecoli_quast**), and then with each of the PacBio reads. The PacBio CLR reads required the --pacbio command while the CSS reads only needed the -s option. K-mer sizes were not specified as SPAdes sets Illumina paired reads by default. The contig length filter of 250 was used for each of the QUAST runs, as that would still keep the longer reads for the *E. coli* fragments. From the QUAST results, there is an extreme increase in the number of contigs after adding the PacBio reads as options. Additionally, the Nx statistics, which shows what x% of the genome contains contigs of a certain length or smaller, reflect this difference in contig quantity. From the Nx graphs, there is a much smoother decline in the distribution of the contig lengths in the assemblies with PacBio reads, as well as much shorter contigs overall.