**Introduction:**

The scope of this lab consisted of assembling three different sets of genomic sequence reads (listing 1) using St. Petersburg genome assemble (SPAdes). The generated results from SPAdes was analyzed with the Quality Assessment Tool (QUAST). SPAdes is an assembly toolkit that contains various assembly pipelines, supporting paired-end reads, mate-pairs and unpaired reads [1]. QUAST is a tool for evaluating and visualizing assemblies by computing various metrics, such as N50, N75, number of misassembled contigs, etc. [2].

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Genome Description  (1) | Platform  (2) | Library Type  (3) | Reference  (QUAST)  (4) | Reference  (QUAST)  (5) | Reference  (SPAdes)  (6) |
| Chloroplast  (B30) | Ion Torrent | single-read | [BC30](https://www.dropbox.com/sh/mwmt6twf5dvcggf/AADdMfBmC7TEu1l1rOnYrfhga/BC30_BINF6350_Summer2014_13pm.fastq?dl=0) | [NC\_007898.fasta](https://www.dropbox.com/sh/mwmt6twf5dvcggf/AAAmDQqTqvCFVF250Hl6Tyr6a/NC_007898.fasta?dl=0)  [NC\_007898.gff](https://www.dropbox.com/sh/mwmt6twf5dvcggf/AAAMVt9vr2RfomADRmniUlCua/NC_007898.gff?dl=0) | none |
| E. Coli (200 bp) | Illumina | paired-end | [ERR008613\_1](https://www.dropbox.com/sh/mwmt6twf5dvcggf/AADnYKTwtbmuYfTv00MXa2jJa/ERR008613sample_1.fastq?dl=0)  [ERR008613\_2](https://www.dropbox.com/sh/mwmt6twf5dvcggf/AAC3sQnfv_vTHp6cOP8iFcb_a/ERR008613sample_2.fastq?dl=0) | [NC\_000913.fna](https://www.dropbox.com/sh/mwmt6twf5dvcggf/AABaf_sS-qMkOoq5Gm7v23sya/NC_000913.fna?dl=0)  [NC\_000913.gff](https://www.dropbox.com/sh/mwmt6twf5dvcggf/AAA3-ctsYc8FcZGKPmg6KMu6a/NC_000913.gff?dl=0) | [PacBio\_2kb\_500bp.fastq](https://www.dropbox.com/sh/mwmt6twf5dvcggf/AACPyRX674YhftOlgNO4JLt3a/PacBio_2kb_CCS_500bp.fastq?dl=0) |
| E. Coli (600 bp) | Illumina | paired-end | [ERR022075\_1](https://www.dropbox.com/sh/mwmt6twf5dvcggf/AADluGAzzPq-ChFKOmYZq6Raa/ERR022075sample_1.fastq?dl=0)  [ERR022075\_2](https://www.dropbox.com/sh/mwmt6twf5dvcggf/AACTo8C86Xd2h7B31GH_k3k4a/ERR022075sample_2.fastq?dl=0) | [NC\_000913.fna](https://www.dropbox.com/sh/mwmt6twf5dvcggf/AABaf_sS-qMkOoq5Gm7v23sya/NC_000913.fna?dl=0)  [NC\_000913.gff](https://www.dropbox.com/sh/mwmt6twf5dvcggf/AAA3-ctsYc8FcZGKPmg6KMu6a/NC_000913.gff?dl=0) | [PacBio\_2kb\_500bp.fastq](https://www.dropbox.com/sh/mwmt6twf5dvcggf/AACPyRX674YhftOlgNO4JLt3a/PacBio_2kb_CCS_500bp.fastq?dl=0) |

**Listing 1: Genomic Sequence Reads with Associated Genomic References**

**Methods/Materials:**

In order to assemble and analyze each dataset in listing 1, I completed the following steps to set up SPAdes and QUAST.

1. Installed/configured SPAdes on mamba.urc.uncc.edu server according to [lab 2-2 slides](https://uncc.instructure.com/courses/121225/files/7719084/download?download_frd=1).
2. Installed QUAST on my local machine according to installation instructions listed within QUAST 5.02 manual [2].
3. Downloaded chloroplast and E. coli genomic sequence read files (column 4, listing 1), reference E. coli files (column 5, listing 1), and PacBio file (column 6, listing 1) onto my local machine.
4. Transferred chloroplast and E. coli genomic sequence read files and PacBio file to mamba.urc.uncc.edu server via sftp for subsequent input for SPAdes.

Next, I utilized SPAdes to assemble each dataset (column 1, listing 1) twice under different configurations by submitting six jobs on the mamba.urc.uncc.edu server for processing. The script for each job is listed in the appendix at the end of this document (A1 & A2 [chloroplast], A3 & A4 [E. coli 200 bp], A5 & A6[E. coli 600 bp]).

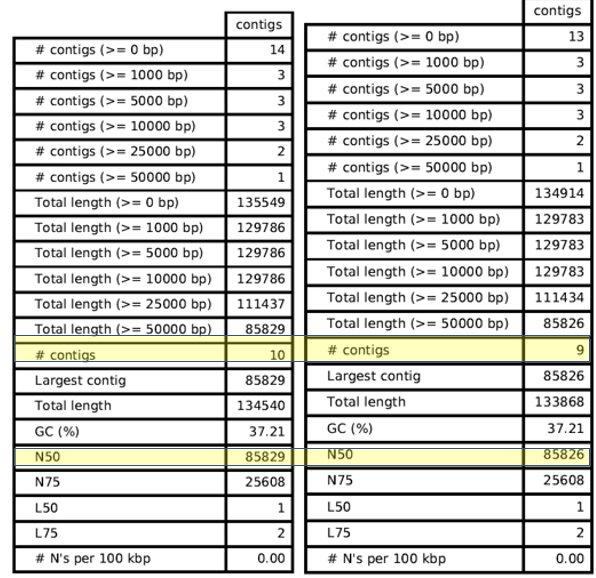
Finally, I zipped and transferred the generated files from SPAdes onto my local machine via sftp and then analyzed the results using QUAST. All six QUAST commands are listed within the appendix (A7 [chloroplast], A8 [E. coli 200 bp] & A9 [E. coli 600 bp]).

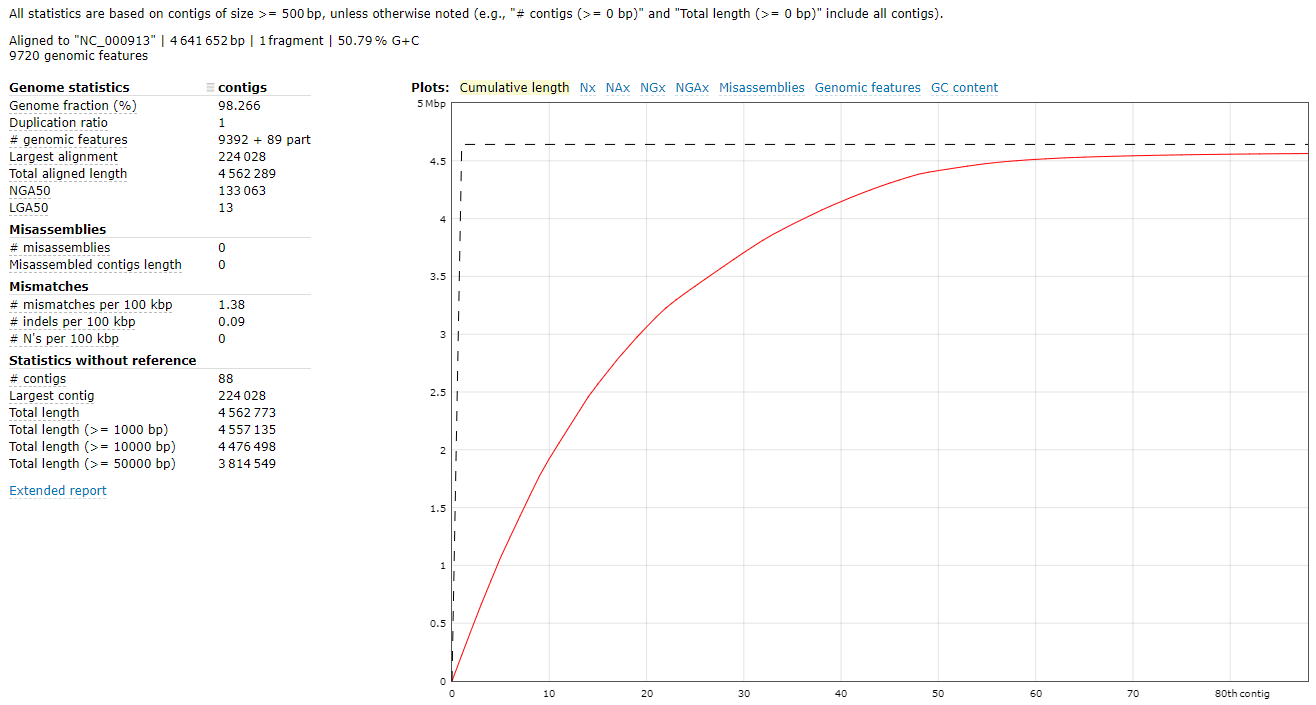
**Results:**

Assembly results for chloroplast, E. coli (200 bp) and E. coli (600 bp) are illustrated in figures 1 to 5. Listing 2 summarizes the key metrics from figures 1 to 5.

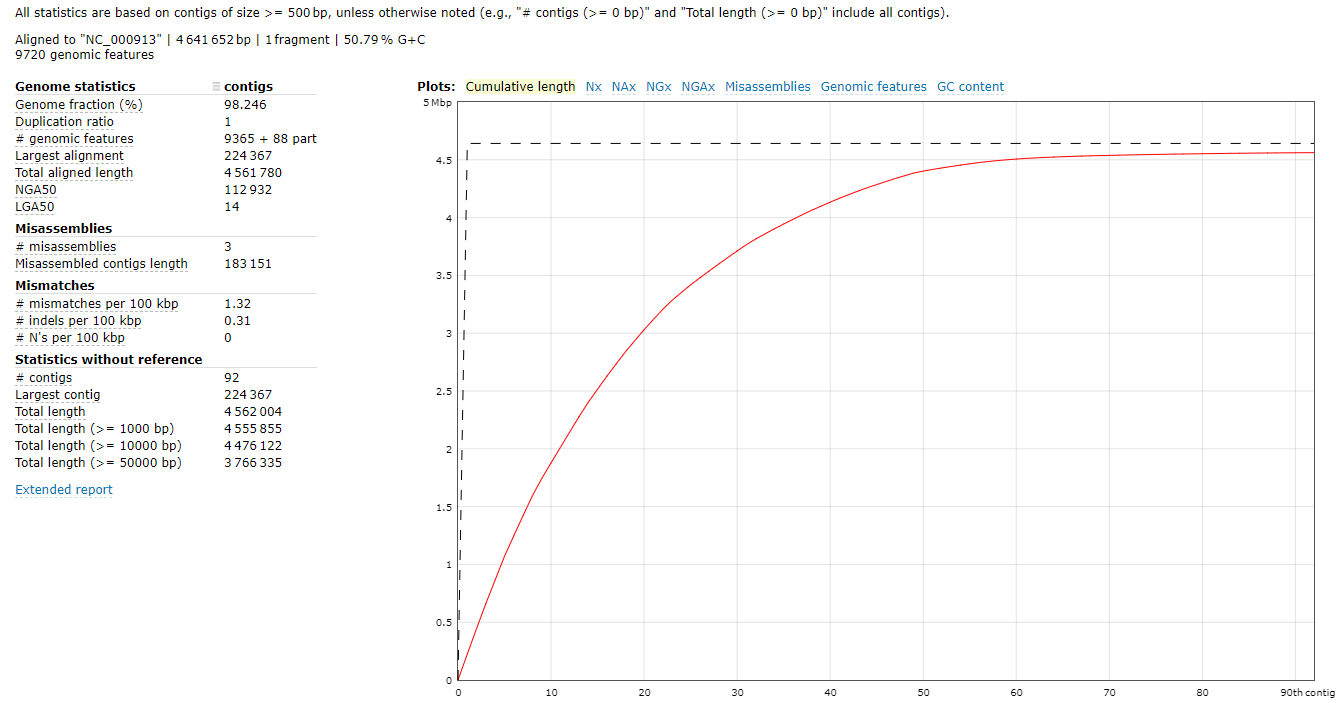
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Genomic Dataset** | **# contigs** | **N50/NG50** | **# misassemblies** | **# mismatches per 100 kbp** |
| Chloroplast  (without ***careful*** setting) | 10 | 85829 | N/A | N/A |
| Chloroplast (with ***careful*** setting) | 9 | 85826 | N/A | N/A |
| E. coli 200 bp (without PacBio ref) | 88 | 133063 | 0 | 1.38 |
| E. coli 200 bp (with PacBio ref) | 92 | 112932 | 3 | 1.32 |
| E. coli 600 bp (without PacBio ref) | 93 | 125811 | 0 | 1.62 |
| E. coli 600 bp (with PacBio ref) | 15 | 496531 | 3 | 3.89 |

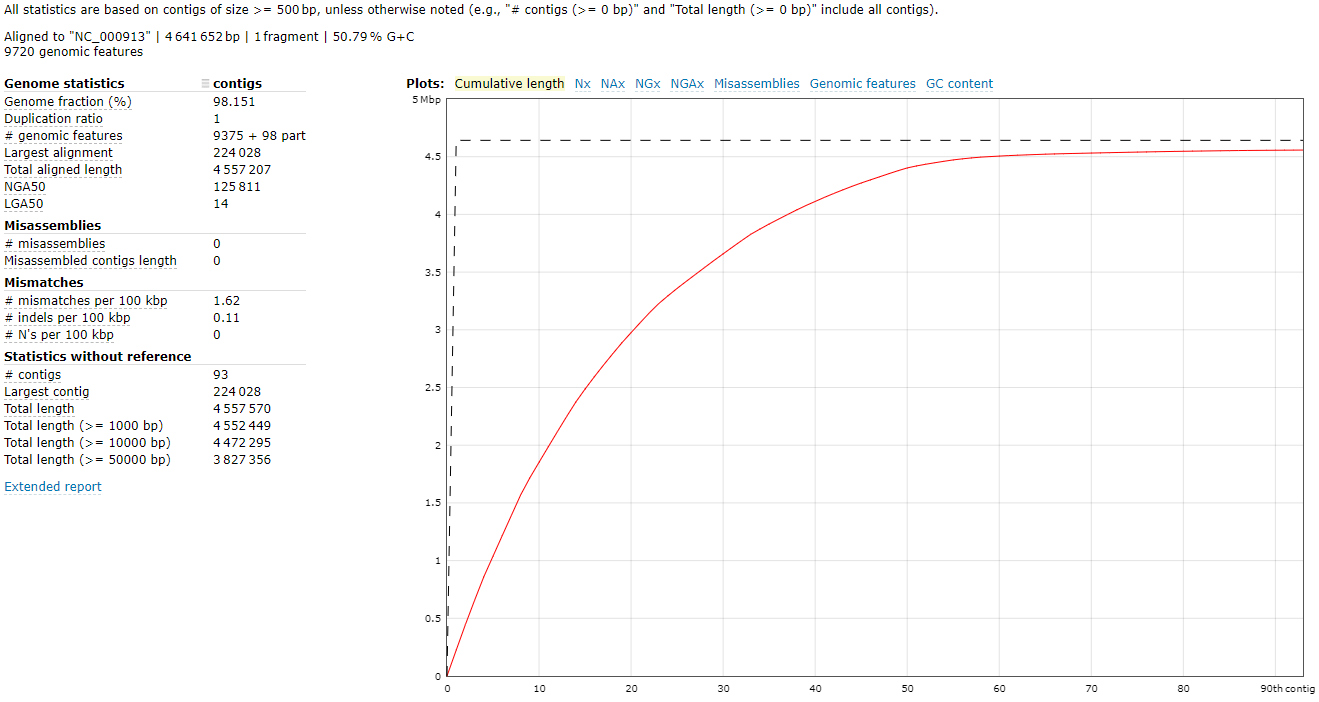
**Listing 2: Key Metrics from Figures 1 to 5.**

  
**Figure 1: Genomic Assembly Results for Chloroplast:**  The right table had the SPAdes setting of ***careful*** enabled, while the left table had it disabled.

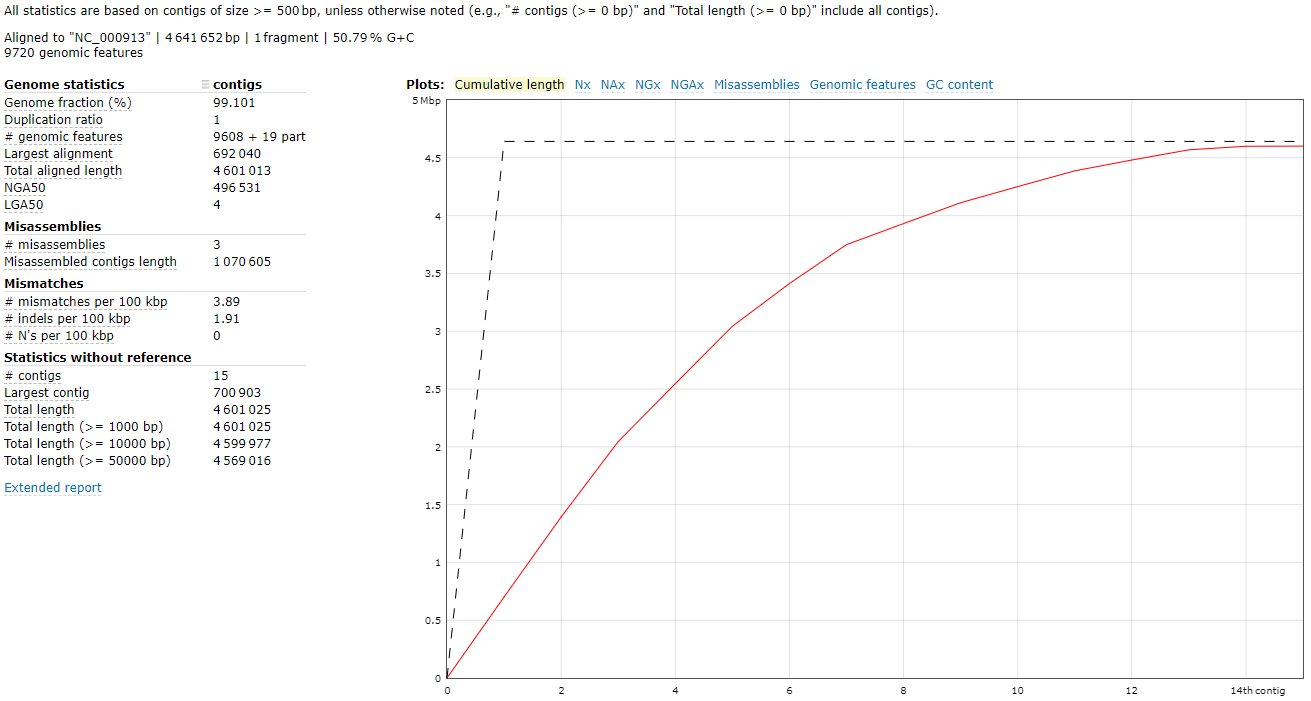


**Figure 2: Genomic Assembly Result for E. coli (200 bp):** PacBio sequence was not utilized as a supplementary SPAdes input file during assembly.

  
**Figure 3: Genomic Assembly Result for E. coli (200 bp):** PacBio sequence was utilized as a as a supplementary SPAdes input file during assembly.



**Figure 4: Genomic Assembly Result for E. coli (600 bp):** PacBio sequence was not utilized as a supplementary SPAdes input file during assembly.



**Figure 5: Genomic Assembly Result for E. coli (600 bp):** PacBio sequence was utilized as a supplementary SPAdes input file during assembly.

**Discussion:**

Utilizing SPAdes on the mamba.urc.uncc.edu server, I was able to successfully assemble all three genomic datasets under six use cases. By the way, I believe I saved a lot of time by running QUAST on my local machine instead of running it on the mamba.urc.uncc.edu server. For the chloroplast genome, I assembled it with the “careful” setting enabled and not enabled. For both versions of the E. coli dataset (200 bp & 600 bp), I assembled them with and without the PacBio dataset. Per the SPAdes manual *(“PacBio and Oxford Nanpore reads”* section), I utilized the PacBio CLF reads to help facilitate gap closure and repeat resolution [1]. Based upon this strategy, I hypothesized that the longer read configuration (E. coli 600 bp) and PacBio supplemental dataset would produce the least number of contigs against the E. coli reference (NC\_000913).

There was not significant difference between the two use cases of assembling the chloroplast genome (listing 2, rows 1 & 2). Also, there was not significant difference between the two E. coli 200 bp use cases (listing 2, rows 3 & 4). However, there was significant difference between the two E. coli 600 bp use cases (listing 2, rows 5 & 6). In fact, the number of contigs was reduced by nearly 84% (93 🡪 15). Moreover, N50 increased in size by about 295% (125,811 🡪 496,531). Therefore, it appears the use of a longer read library, coupled with PacBio reads, significantly contributed to improving both metrics of N50 and # of contigs.

**Work Cited/References:**

1. Center for Algorithmic Biotechnology---St. Peterburg State University, SPAdes 3.14.0 Manual, <http://cab.spbu.ru/files/release3.14.0/manual.html#sec1>
2. Center for Algorithmic Biotechnology---St. Peterburg State University, QUAST 5.0.2 Manual, <http://cab.spbu.ru/files/quast/latest-docs/manual.html#sec3.1>

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*APPENDIX \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*A1: BC30\_default.sh\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#!/bin/bash

#=====OPTIONS=====

#PBS -q mamba

#PBS -N nickname

#PBS -l nodes=2:ppn=16

#PBS -l walltime=2:00:00

#PBS -l prologue=/users/jyoun144/path/for/prologue.sh,epilogue=/users/jyoun144/path/for/epiloque.sh

#=====END PBS OPTIONS====

module load anaconda3

eval "$(conda shell.bash hook)"

conda activate spades

spades.py -s /users/jyoun144/BC30.fastq **--iontorrent -k 21,33,55,77,99,127** -o /users/jyoun144/lab02/results/BC30/result1/

conda deactivate

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*A2: BC30\_careful.sh\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#!/bin/bash

#=====OPTIONS=====

#PBS -q mamba

#PBS -N nickname

#PBS -l nodes=2:ppn=16

#PBS -l walltime=2:00:00

#PBS -l prologue=/users/jyoun144/path/for/prologue.sh,epilogue=/users/jyoun144/path/for/epiloque.sh

#=====END PBS OPTIONS====

module load anaconda3

eval "$(conda shell.bash hook)"

conda activate spades

spades.py -s /users/jyoun144/BC30.fastq **--iontorrent --careful -k 21,33,55,77,99,127** -o /users/jyoun144/lab02/results/BC30/result2/

conda deactivate

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*A3: ERR008613\_default.sh\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#!/bin/bash

#=====OPTIONS=====

#PBS -q mamba

#PBS -N nickname

#PBS -l nodes=2:ppn=16

#PBS -l walltime=2:00:00

#PBS -l prologue=/users/jyoun144/path/for/prologue.sh,epilogue=/users/jyoun144/path/for/epiloque.sh

#=====END PBS OPTIONS====

module load anaconda3

eval "$(conda shell.bash hook)"

conda activate spades

spades.py --pe1-1 /users/jyoun144/ERR008613sample\_1.fastq --pe1-2 /users/jyoun144/ERR008613sample\_2.fastq -k 21,33,55,77 -o /users/jyoun144/lab02/results/ERR008613/result1/

conda deactivate

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*A4: ERR008613\_PacBio\_2kb\_CCS\_500bp.sh\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#!/bin/bash

#=====OPTIONS=====

#PBS -q mamba

#PBS -N nickname

#PBS -l nodes=2:ppn=16

#PBS -l walltime=2:00:00

#PBS -l prologue=/users/jyoun144/path/for/prologue.sh,epilogue=/users/jyoun144/path/for/epiloque.sh

#=====END PBS OPTIONS====

module load anaconda3

eval "$(conda shell.bash hook)"

conda activate spades

spades.py --pe1-1 /users/jyoun144/ERR008613sample\_1.fastq --pe1-2 /users/jyoun144/ERR008613sample\_2.fastq -k 21,33,55,77 -o /users/jyoun144/lab02/results/ERR008613/result2/ -s /users/jyoun144/PacBio\_2kb\_CCS\_500bp.fastq

conda deactivate

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*A5: ERR022075\_default.sh\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#!/bin/bash

#=====OPTIONS=====

#PBS -q mamba

#PBS -N nickname

#PBS -l nodes=2:ppn=16

#PBS -l walltime=2:00:00

#PBS -l prologue=/users/jyoun144/path/for/prologue.sh,epilogue=/users/jyoun144/path/for/epiloque.sh

#=====END PBS OPTIONS====

module load anaconda3

eval "$(conda shell.bash hook)"

conda activate spades

spades.py --pe1-1 /users/jyoun144/ERR022075sample\_1.fastq --pe1-2 /users/jyoun144/ERR022075sample\_2.fastq -k 21,33,55,77 -o /users/jyoun144/lab02/results/ERR022075/result1/

conda deactivate

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*A6: ERR022075\_PacBio\_10kb\_CLR.sh\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#!/bin/bash

#=====OPTIONS=====

#PBS -q mamba

#PBS -N nickname

#PBS -l nodes=2:ppn=16

#PBS -l walltime=2:00:00

#PBS -l prologue=/users/jyoun144/path/for/prologue.sh,epilogue=/users/jyoun144/path/for/epiloque.sh

#=====END PBS OPTIONS====

module load anaconda3

eval "$(conda shell.bash hook)"

conda activate spades

spades.py --pe1-1 /users/jyoun144/ERR022075sample\_1.fastq --pe1-2 /users/jyoun144/ERR022075sample\_2.fastq -k 21,33,55,77 **--pacbio /users/jyoun144/PacBio\_10kb\_CLR.fastq** -o /users/jyoun144/lab02/results/ERR022075/result2/

conda deactivate

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*A7: BC30\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*result 1\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/home/youngjb/quast-5.0.2/quast.py /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/BC30/Spades/Default/contigs.fasta -t 4 -o /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/BC30/Quast/result1/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*result 2\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/home/youngjb/quast-5.0.2/quast.py /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/BC30/Spades/Careful/contigs.fasta -t 4 -o /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/BC30/Quast/result2/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*A8: ERR008613\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*result 1\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/home/youngjb/quast-5.0.2/quast.py /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/ERR008613/Spades/result1/contigs.fasta -t 4 -r /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/references/NC\_000913.fna -g /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/references/NC\_000913.gff -o /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/ERR008613/Quast/result1\_reference\_default\_contig/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*result 2\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/home/youngjb/quast-5.0.2/quast.py /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/ERR008613/Spades/result2/contigs.fasta -t 4 -r /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/references/NC\_000913.fna -g /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/references/NC\_000913.gff -o /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/ERR008613/Quast/result2\_reference\_default\_contig/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*A9: ERR022075\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*result 1\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/home/youngjb/quast-5.0.2/quast.py /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/ERR022075/Spades/result1/contigs.fasta -t 4 -r /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/references/NC\_000913.fna -g /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/references/NC\_000913.gff -o /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/ERR022075/Quast/result1\_reference\_default\_contig/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*result 2\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/home/youngjb/quast-5.0.2/quast.py /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/ERR022075/Spades/result2/contigs.fasta -t 4 -r /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/references/NC\_000913.fna -g /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/references/NC\_000913.gff -o /mnt/c/Users/young/Documents/UNCC\_Spring\_2020/Genomics/labs/lab02/ERR022075/Quast/result2\_reference\_default\_contig/