1. From the ABySS output, create a table for the unitigs, contigs, and scaffolds with the number of each, N50 for each, and predicted genome length.

Name	n	N50	Predicted Genome Length
Assembly-untigs.fa	247	112494	5388023
Assembly-contigs.fa	134	593361	5411307
Assembly-scaffolds.fa	125	593361	5411685

2. In your own words, please summarize the function of each of the commands (e.g., abyss-pe, k, B, etc) that you included in your code.

For the command abyss-pe name=assembly k=96 B=2G in='SRR32657023_1.fastq.gz SRR32657023_2.fastq.gz'.
Abyss-pe organizes the genome assembly process, K refers to the k-mer length (length of short DNA sequences), B refers to the size of the Bloom filter, and in refers to the input read files used for assembly.

3.Based on this manual, can you identify how you could modify the code you used to do a hybrid assembly with nanopore reads? Please explain what a hybrid assembly is and why someone might want to do that.

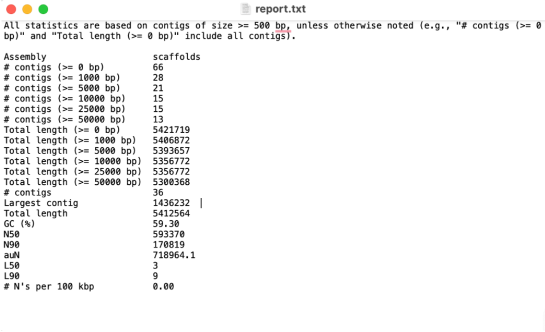
You could modify the code we used to do a hybrid assembly with nanopore reads similar to this:

```
spades.py \
-1 Lillyanna_1.fastq.gz \
-2 Lillyanna_2.fastq.gz \
--nanopore nanopore_reads.fastq \
-o spades_hybrid_out
```

Hybrid assembly combines short reads with long reads in order to use strengths from both technologies. The short reads correct errors in the long reads and the long reads help span complex regions that the short reads are unable to read.

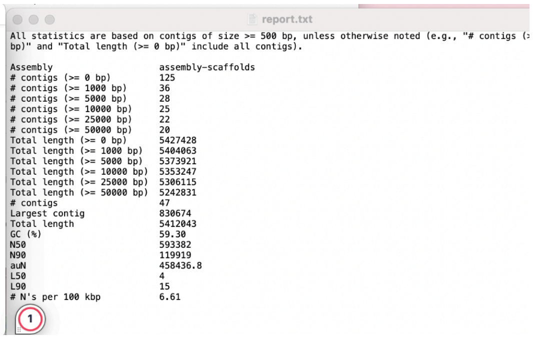
4. Include a screenshot of the QUAST assembly statistics for the ABySS and SPAdes assembly.

Spades on the left, ABySS on the right.



```
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0
bp)" and "Total length (>= 0 bp)" include all contigs).

Assembly                    scaffolds
# contigs (>= 0 bp)         66
# contigs (>= 1000 bp)      28
# contigs (>= 5000 bp)      21
# contigs (>= 10000 bp)     15
# contigs (>= 25000 bp)     15
# contigs (>= 50000 bp)     13
Total length (>= 0 bp)      5421719
Total length (>= 1000 bp)   5406872
Total length (>= 5000 bp)   5393657
Total length (>= 10000 bp)  5356772
Total length (>= 25000 bp)  5356772
Total length (>= 50000 bp)  5300368
# contigs                   36
Largest contig              1436232
Total length                5412564
GC (%)                      59.30
N50                         593370
N90                         170819
auN                         718964.1
L50                         3
L90                         9
# N's per 100 kbp           0.00
```



```
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>
bp)" and "Total length (>= 0 bp)" include all contigs).

Assembly                    assembly-scaffolds
# contigs (>= 0 bp)         125
# contigs (>= 1000 bp)      36
# contigs (>= 5000 bp)      28
# contigs (>= 10000 bp)     25
# contigs (>= 25000 bp)     22
# contigs (>= 50000 bp)     20
Total length (>= 0 bp)      5427428
Total length (>= 1000 bp)   5404063
Total length (>= 5000 bp)   5373921
Total length (>= 10000 bp)  5353247
Total length (>= 25000 bp)  5306115
Total length (>= 50000 bp)  5242831
# contigs                   47
Largest contig              830674
Total length                5412043
GC (%)                      59.30
N50                         593382
N90                         119919
auN                         458436.8
L50                         4
L90                         15
# N's per 100 kbp           6.61
```

5. Based on the statistics from your genome, which assembly do you think is best? Why? This is the assembly you can use going forward.

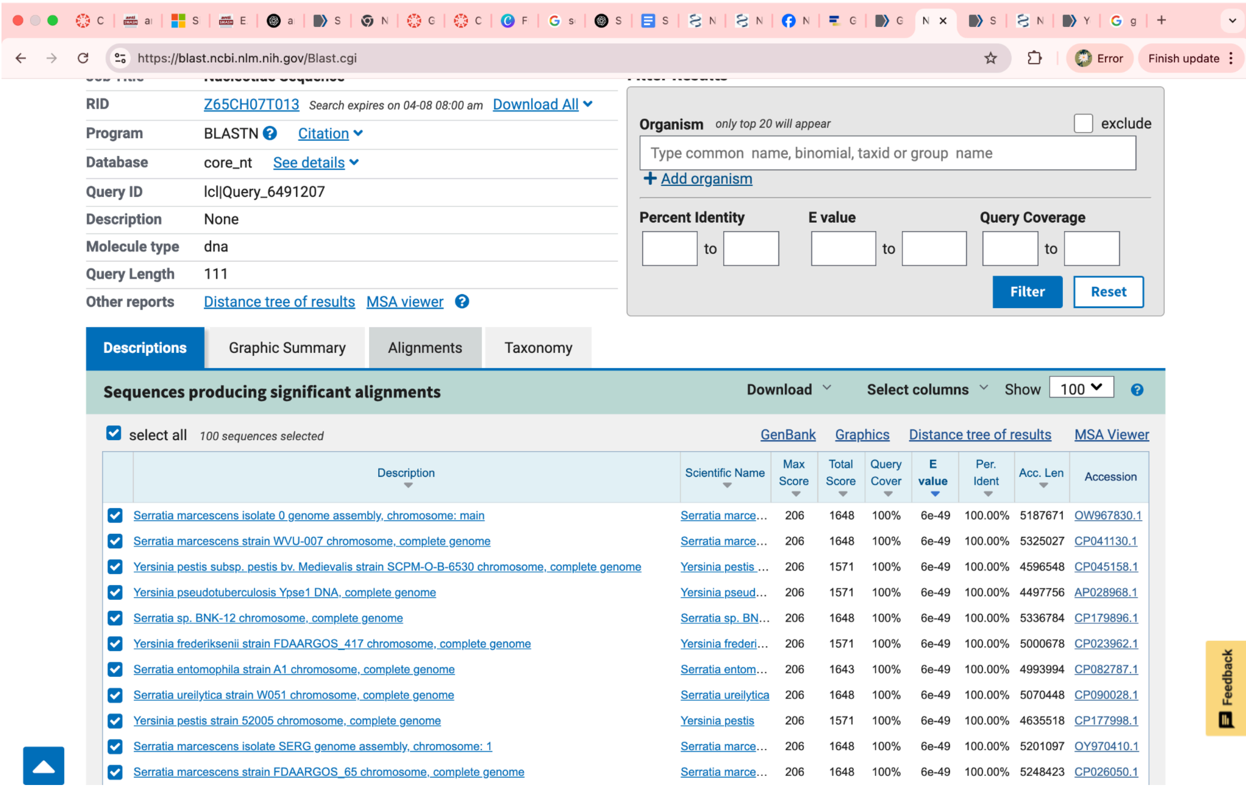
Spades is the better option because the n is lower and it has fewer errors. Spades has 0 errors while ABySS has 6.61. N90, L50, and L90 all have lower values as well.

6. How can we use barrnap to figure out what species we have? Why is using the 16S rRNA sequence a good, but imperfect, tool for identifying species identity?

We can use barrnap to figure out what species we have by being able to obtain the 16s rRNA sequence of the species. You can then copy this sequence and paste it into a blastn search on NCBI. This will be able to show you what species is most similar.

7. What species do you have? Include a screenshot of your top NCBI results.

Serratia marcescens isolate 0 genome assembly, chromosome: main



RID	Z65CH07T013 *Search expires on 04-08 08:00 am* Download All
Program	BLASTN Citation
Database	core_nt See details
Query ID	lcl|Query_6491207
Description	None
Molecule type	dna
Query Length	111
Other reports	Distance tree of results MSA viewer

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

Sequences producing significant alignments

☑ select all *100 sequences selected*

GenBank Graphics Distance tree of results MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Serratia marcescens isolate 0 genome assembly, chromosome: main	Serratia marce...	206	1648	100%	6e-49	100.00%	5187671	OW967830.1
☑	Serratia marcescens strain WVU-007 chromosome, complete genome	Serratia marce...	206	1648	100%	6e-49	100.00%	5325027	CP041130.1
☑	Yersinia pestis subsp. pestis bv. Medievalis strain SCPM-O-B-6530 chromosome, complete genome	Yersinia pestis ...	206	1571	100%	6e-49	100.00%	4596548	CP045158.1
☑	Yersinia pseudotuberculosis Ypse1 DNA, complete genome	Yersinia pseud...	206	1571	100%	6e-49	100.00%	4497756	AP028968.1
☑	Serratia sp. BNK-12 chromosome, complete genome	Serratia sp. BN...	206	1648	100%	6e-49	100.00%	5336784	CP179896.1
☑	Yersinia frederiksenii strain FDAARGOS_417 chromosome, complete genome	Yersinia frederi...	206	1571	100%	6e-49	100.00%	5000678	CP023962.1
☑	Serratia entomophila strain A1 chromosome, complete genome	Serratia entom...	206	1643	100%	6e-49	100.00%	4993994	CP082787.1
☑	Serratia ureilytica strain W051 chromosome, complete genome	Serratia ureilytica	206	1648	100%	6e-49	100.00%	5070448	CP090028.1
☑	Yersinia pestis strain 52005 chromosome, complete genome	Yersinia pestis	206	1571	100%	6e-49	100.00%	4635518	CP177998.1
☑	Serratia marcescens isolate SERG genome assembly, chromosome: 1	Serratia marce...	206	1648	100%	6e-49	100.00%	5201097	OY970410.1
☑	Serratia marcescens strain FDAARGOS_65 chromosome, complete genome	Serratia marce...	206	1648	100%	6e-49	100.00%	5248423	CP026050.1

8. What is genome annotation? Why is it important to do that?

Genome annotation identifies genes and other important aspects of the genome sequence. It also helps identify ORFs, promoters, operons, exons, introns, and other RNAs. For functional purposes, it is helpful in assigning functions which are often based on sequence homology.

9. Perform a genome annotation using two different programs. Find 3 of the 5 genes/features in your results file and create a table of those results: recA, gyrA, 16S rRNA, rpsB, dnaA. What is the location of the genes you chose? What does each program tell you about the gene? How are the outputs different between the two programs.

Prokka	**Location**	**Function**
recA	Reverse(minus) strand: 144,849-145,910	Homologous recombination, activates SOS response, genome stability.

gyrA	Forward(positive) strand: 374,133-376,769	Function: introducing negative supercoils into DNA
dnaA	Forward(positive) strand: 669-52,069	Initiation of **DNA replication**

dfast	**Location**	**Function**
recA	Reverse(minus) strand: 45,561-46,625	DNA repair and homologous recombination
gyrA	Forward(positive) strand: 66,669-66,725	encodes the **DNA topoisomerase (ATP-hydrolyzing) subunit A**
dnaA	Forward (positive) strand: 50,999-52,324	Initiation of DNA replication

10. Create a table for your ANI results. How do you interpret these results? What do each of the columns represent?

Query Genome	Reference Genome	ANI (%)	Fragments Used	Total Fragments
spadesout/scaffolds.fasts	neighbors/typhimurium.fasta	98.9482	1,483	1,610
spadesout/scaffolds.fasta	neighbors/bongori.fasta	90.0417	1,241	1,610

- The query genomes represent the assembled genomes
- The reference genomes represent the comparison between the strains
- The ANI % represents the average nucleotide identity percentage between the strains and shows the genetic similarity between the respective genomes
- The fragments used include the number of conserved DNA fragments aligned for ANI calculation
- The total fragments represent the total sampled fragments from the query genome
- Result interpretation: The genome was identified and supported with a 98.95% ANI match, which confirms species-level identity. In contrast the 90.04% ANI demonstrates the expected genus-level divergence between the species.