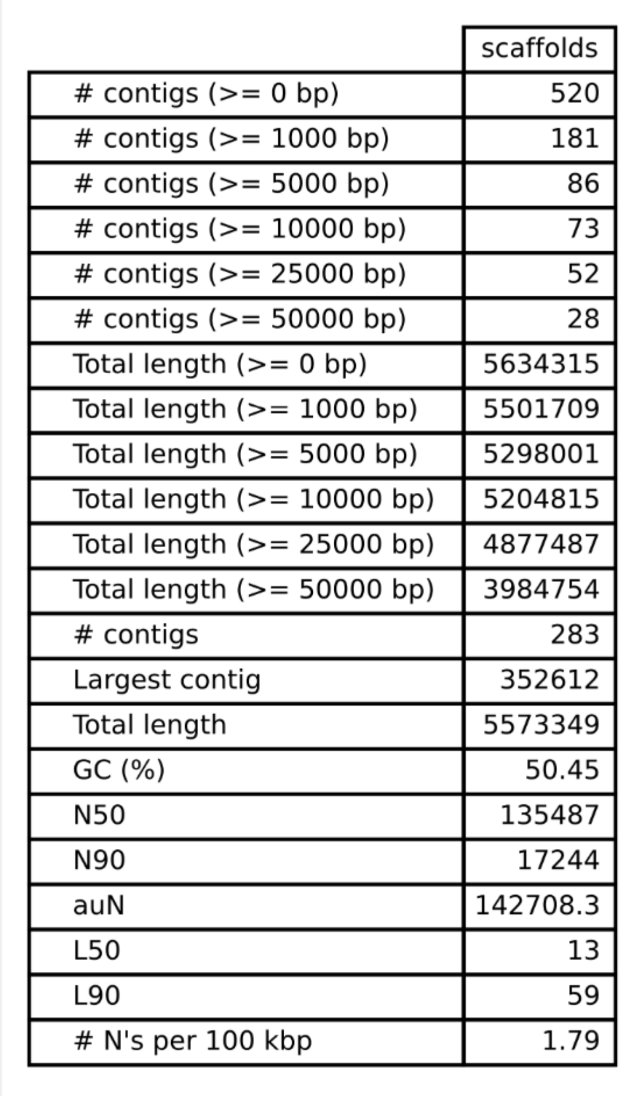
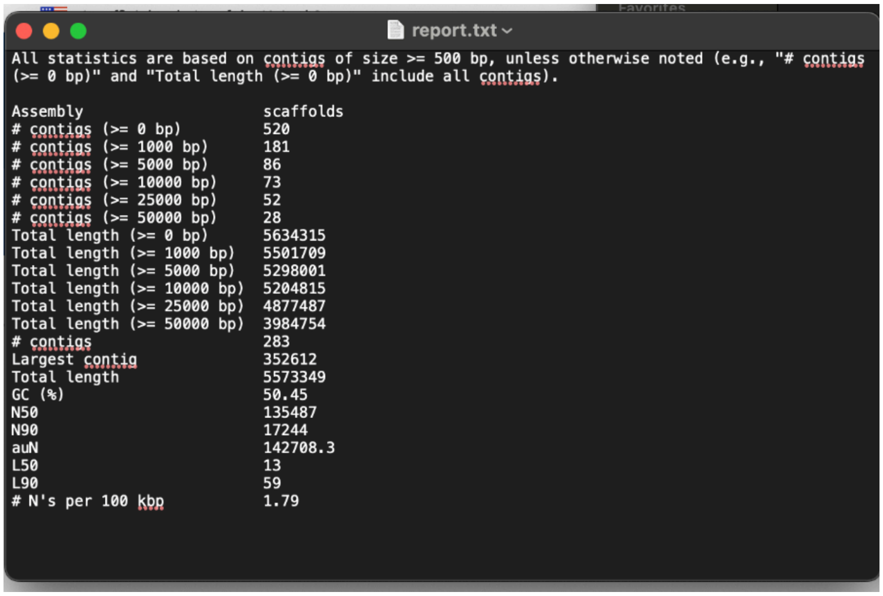


	scaffolds
# contigs (>= 0 bp)	520
# contigs (>= 1000 bp)	181
# contigs (>= 5000 bp)	86
# contigs (>= 10000 bp)	73
# contigs (>= 25000 bp)	52
# contigs (>= 50000 bp)	28
Total length (>= 0 bp)	5634315
Total length (>= 1000 bp)	5501709
Total length (>= 5000 bp)	5298001
Total length (>= 10000 bp)	5204815
Total length (>= 25000 bp)	4877487
Total length (>= 50000 bp)	3984754
# contigs	283
Largest contig	352612
Total length	5573349
GC (%)	50.45
N50	135487
N90	17244
auN	142708.3
L50	13
L90	59
# N's per 100 kbp	1.79

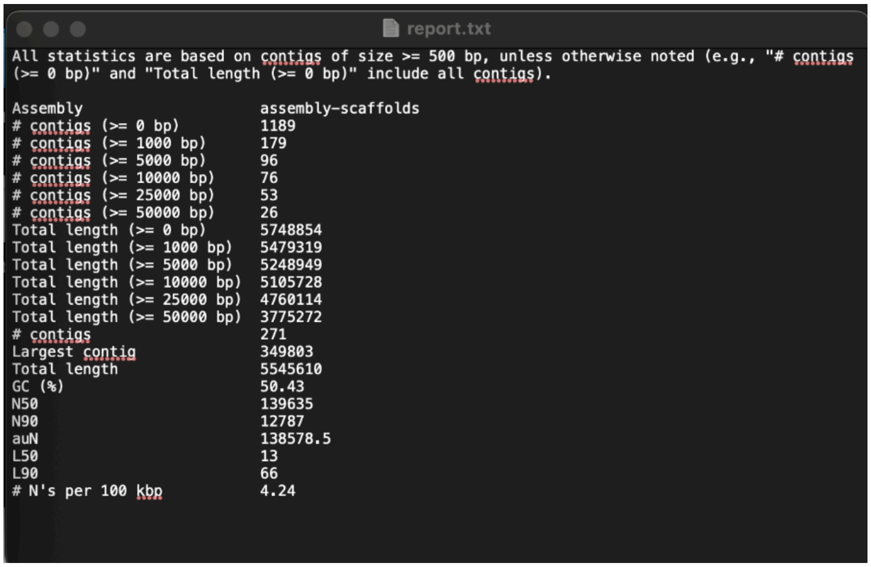
1.
2. Abyss-pe: the ABySS software that assembles reads into contigs and scaffolds. It assembles short-end reads of the input file into contigs.
   k: kmer size; a substring of length k in a DNA or RNA sequence.
   name=assembly: means that the genome is being assembled
   B: how much memory ABySS has access to; B is the bloom filter memory budget.
   in = insert the fastq files that are going to be assembled.
3. To modify the code I used to do a hybrid assembly with nanopore reads, I would use the two illumina files I already have, in addition to the long read sequence I would receive from PacBio. I would then type --nanopore <file_name> into the command line. This would file with Oxford nanopore reads. Hybrid assembly is when data from different sequencing technologies (short-read and long-read sequences) are combined to assemble a genome. Hybrid assembly can combine the short-read sequences from technology, such as Illumina, with long-read sequences from technology, such as Pac-Bio, to help resolve ambiguities and repetitive sequences in genomes.
4. Quast SPAdes:



```
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)         520
# contigs (>= 1000 bp)      181
# contigs (>= 5000 bp)      86
# contigs (>= 10000 bp)     73
# contigs (>= 25000 bp)     52
# contigs (>= 50000 bp)     28
Total length (>= 0 bp)      5634315
Total length (>= 1000 bp)   5501709
Total length (>= 5000 bp)   5298001
Total length (>= 10000 bp)  5204815
Total length (>= 25000 bp)  4877487
Total length (>= 50000 bp)  3984754
# contigs                   283
Largest contig              352612
Total length                5573349
GC (%)                      50.45
N50                         135487
N90                         17244
auN                         142708.3
L50                         13
L90                         59
# N's per 100 kbp           1.79
```

Quast ABySS:



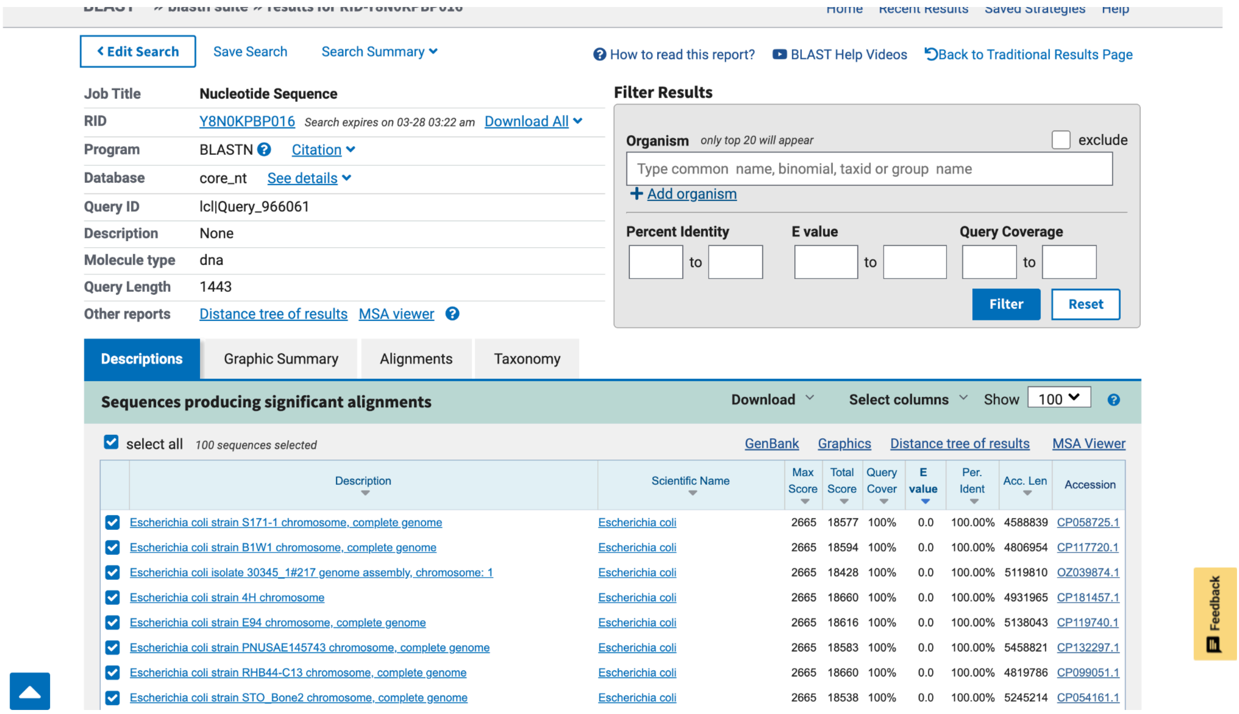
```
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs
(>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Assembly                    assembly-scaffolds
# contigs (>= 0 bp)         1189
# contigs (>= 1000 bp)      179
# contigs (>= 5000 bp)      96
# contigs (>= 10000 bp)     76
# contigs (>= 25000 bp)     53
# contigs (>= 50000 bp)     26
Total length (>= 0 bp)      5748854
Total length (>= 1000 bp)   5479319
Total length (>= 5000 bp)   5248949
Total length (>= 10000 bp)  5105728
Total length (>= 25000 bp)  4760114
Total length (>= 50000 bp)  3775272
# contigs                   271
Largest contig              349803
Total length                5545610
GC (%)                      50.43
N50                         139635
N90                         12787
auN                         138578.5
L50                         13
L90                         66
# N's per 100 kbp           4.24
```

5. Based on the statistics from my genome, I think that the SPAdes assembly is slightly better than the ABySS assembly. I believe this because the SPAdes assembly has a lower number of errors compared to the ABySS assembly. The SPAdes assembly also has a lower L90; lower L90 numbers indicate a more contiguous, complete strand. Additionally, the SPAdes assembly has a lower number of contigs.
6. Barrnap can be used to determine the species that I have because it can utilize the 16S rRNA sequence to help identify the species. Barrnap predicts the location of rRNA genes in genomes utilizing the 16S rRNA. Barrnap utilizes FASTA DNA sequences as an input, and generates a GFF3 as an output. The FASTA files have the 16S rRNA sequence that can be inserted into BLAST to determine the closest match. The 16S rRNA is a good tool to utilize for identifying species identity because almost all bacteria have a 16S rRNA gene, meaning that it can be universally used. Additionally, the 16S rRNA contains variable regions that allows it to be differentiated against other species.

But, it also contains conserved regions that can allow it to be matched to other species. However, the 16S rRNA gene is imperfect because the gene can be remarkably similar between closely related species; species in the same genus may not contain much variability in the 16S rRNA, therefore making it difficult to differentiate between the species. Additionally, databases may not contain much information about rare or newer species, therefore there would not be much data to compare the 16S rRNA gene to.

7.



Job Title	Nucleotide Sequence
RID	Y8N0KPBP016 *Search expires on 03-28 03:22 am* Download All
Program	BLASTN Citation
Database	core_nt See details
Query ID	lcl|Query_966061
Description	None
Molecule type	dna
Query Length	1443
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 □

Sequences producing significant alignments Show 100

☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Escherichia coli strain S171-1 chromosome, complete genome	Escherichia coli	2665	18577	100%	0.0	100.00%	4588839	CP058725.1
☑	Escherichia coli strain B1W1 chromosome, complete genome	Escherichia coli	2665	18594	100%	0.0	100.00%	4806954	CP117720.1
☑	Escherichia coli isolate 30345_1#217 genome assembly, chromosome: 1	Escherichia coli	2665	18428	100%	0.0	100.00%	5119810	OZ039874.1
☑	Escherichia coli strain 4H chromosome	Escherichia coli	2665	18660	100%	0.0	100.00%	4931965	CP181457.1
☑	Escherichia coli strain E94 chromosome, complete genome	Escherichia coli	2665	18616	100%	0.0	100.00%	5138043	CP119740.1
☑	Escherichia coli strain PNUSAE145743 chromosome, complete genome	Escherichia coli	2665	18583	100%	0.0	100.00%	5458821	CP132297.1
☑	Escherichia coli strain RHB44-C13 chromosome, complete genome	Escherichia coli	2665	18660	100%	0.0	100.00%	4819786	CP099051.1
☑	Escherichia coli strain STO_Bone2 chromosome, complete genome	Escherichia coli	2665	18538	100%	0.0	100.00%	5245214	CP054161.1

8. A genome annotation reads the designated genome and produces the genome’s protein sequence. Genome annotation identifies the functional elements of a genome, such as genes, regulatory regions, promoters, and enhancers. Annotations of genomes can include conducting quality and taxonomy assessments. Genome annotation is significant because it provides information about the biological features of a DNA sequence. Genome annotation can help individuals understand how a specific bacteria works at the molecular level. Additionally, genome annotations can aid scientists in the comparison of genomes of different organisms. Furthermore, genome annotation is important to conduct because it could be extremely helpful in clinical and medical settings.
9.

Gene	Location of Gene	What Program Tells Me About Gene	How the Outputs Differ

gyrA	MBHKAFJI_00089	The product is DNA gyrase subunit A
		EC_number="5.6.2.2”
			Dfast:
			inference = COORDINATES:ab initio prediction:MetaGeneAnnotator" "similar to AA sequence:RefSeq:WP_001281242.1”
			Prokka:
			inference = "ab initio prediction:Prodigal:002006" "similar to AA sequence:UniProtKB:P0AES5" /codon_start=1

rpsB	MBHKAFJI_01669	The product is 30S ribosomal protein S2
		No EC_number
			Dfast:
			inference = "COORDINATES:ab initio prediction:MetaGeneAnnotator" "similar to AA sequence:RefSeq:WP_000246884.1"
			Prokka:
			inference = "ab initio prediction:Prodigal:002006" "similar to AA sequence:UniProtKB:P0A7V0"

dnaA	MBHKAFJI_02006	The product is chromosomal replication initiator protein DnaA No EC_number	Dfast: Inference = "COORDINATES:ab initio prediction:MetaGeneAnnotator" "similar to AA sequence:RefSeq:WP_000059111.1" Prokka:

			Inference = "ab initio prediction:Prodigal:002006" "similar to AA sequence:UniProtKB:P03004"

10.

Scaffolds file	FASTA file	% Similarity	Alignment Length	Total Length
spadesout/scaffolds.fasta	neighbors/fergusonii.fasta	90.9639	1133	1741
spadesout/scaffolds.fasta	neighbors/albertii.fasta	90.0592	1254	1741

The first column, the scaffolds file, represents the source file name. The second column, the FASTA file, represents the target file name. The third column, the percent similarity, represents similarity between sequences. The sequences compared are the reference sequences (*E. fergusonii* and *E. albertii*) and the scaffold sequence (*E. coli*). The fourth column, the alignment length, represents the length of the aligned region. The fifth column, the total length, represents the total scaffold length. Based on the results that I received, I can conclude that both of the reference sequences are almost equally similar to the scaffold sequence. Because both sequences have about a 90% similarity, it can be predicted that the reference sequences are almost equally related to the scaffold sequence. I can also conclude that a majority of the scaffold is being aligned to the reference sequences in both cases. I can predict this because the high alignment length suggests that the assembly was successful in reconstructing meaningful sequences. Finally, these results support that the scaffolds sequence is in the genus *Escherichia* because of the percent similarity.