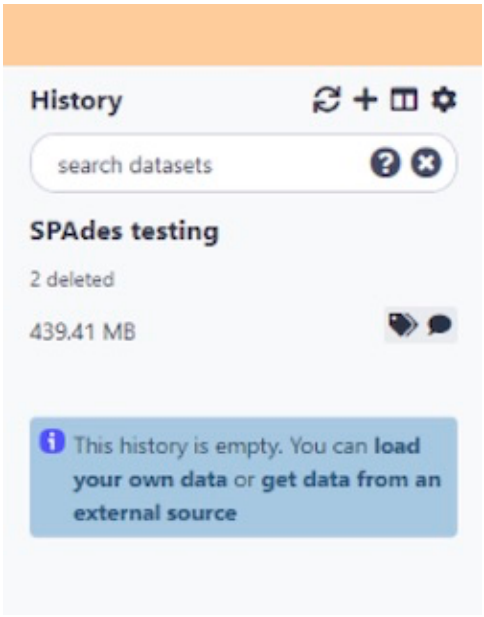
Software Challenge
Genome Assembly
02-604: Fundamentals of Bioinformatics

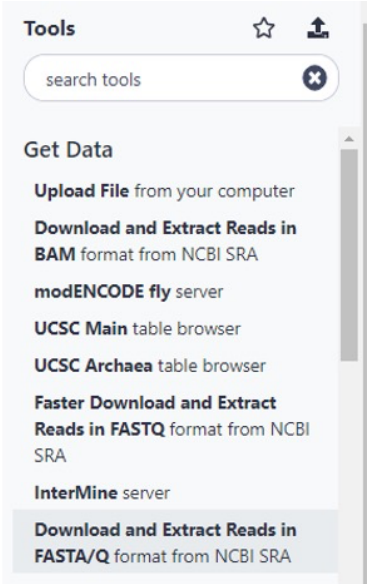
Every year in the United States, half a million patients contract a ***Staphylococcus* (Staph)** infection after surgery. Many of these patients are infected with drug-resistant strains such as **methicillin-resistant *Staphylococcus aureus* (MRSA)**, which can resist even last-resort antibiotics like Vancomycin and Daptomycin. As a result, MRSA causes over 20,000 deaths a year in the U.S. alone. Since there are over 40 different types of Staph bacteria that could be causing these infections, you want to determine which species is causing a Staph infection in a given patient by isolating this species in the patient and sequencing its genome. After you have sequenced its genome, scientists can start analyzing mutations that have led to antibiotics resistance.

Let's assume that we have isolated bacteria in the patient and generated reads for these bacteria. To assemble the genome from the reads, you will be using the **SPAdes** assembler (Bankevich *et al*, 2012) through the Galaxy service. Please follow these step-by-step instructions to register on Galaxy and run SPAdes:

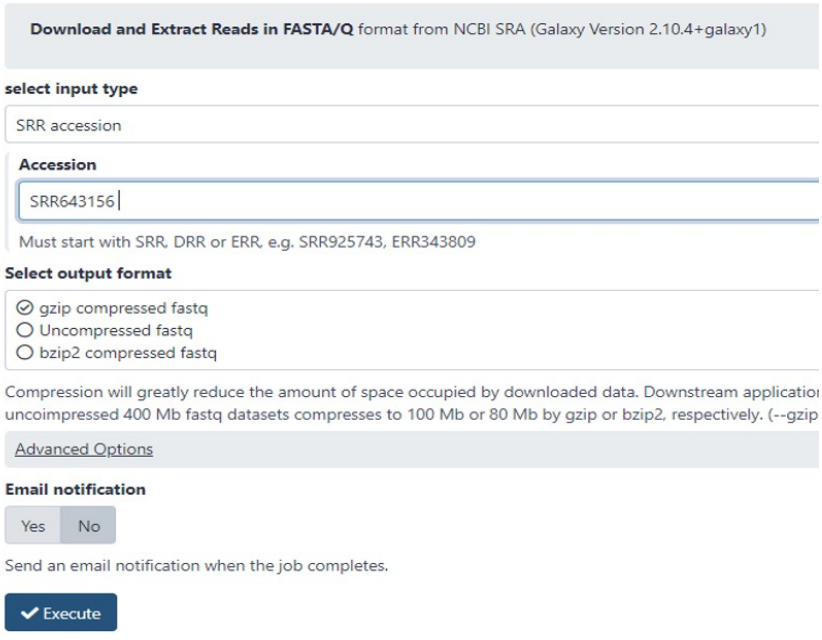
Register: Create an account on Galaxy here. You will need to fill out all fields. After logging into Galaxy, you will see the following dashboard on the right side of the page. Click on the plus '+' and then create a new history. Name it whatever you like. All of the following analysis will be performed under this history.



Next, we need to import our data. For this assignment, we will import raw data directly from the SRA database. Click on “Get Data” in the left side menu and search for the “**Download and Extract Reads in FASTA/Q** format from NCBI SRA”.



Launch the tool and use accession number **SRR643156** to import the data. Click "Execute" to import the data to your current history. It may take about 30 minutes for the app to begin execution and to load the data. When the tool is finished, you will be able to see the imported files and related information in your history on the right side of the page.



Next, we will use SPAdes to assemble the genome. Go to "tools" on the left side of the page and search for SPAdes , clicking on it to launch it. Use default settings except for the value of kmer size. For this homework assignment, you will need to run the program three separate time times (for $k = 25$, $k = 55$, and $k = 85$) to investigate how the choice of parameter k affects the resulting assembly.

For each value of k, enter the value of k and select 'Interleaved files' as the file format and select the file you imported in the previous step in the reads section. Hit execute to run SPAdes on the file.

A few important notes. First, you do not need to wait for the results to finish can queue all three runs at once – you will need to repeat all of the steps in this paragraph for each run. Second, Galaxy is an excellent public resource, but sometimes jobs may be slow to run or fail. If you have a failed run, or you would rather complete the analysis in this challenge without running SPAdes yourself, you can find files with the results of running SPAdes here:

https://drive.google.com/drive/folders/1L-cfjiyd9RJOwY1Slau-wQHGmO1g1xft?usp=sharing.

The contents of the above folder are as follows:

- SRR643156_(fastq-dump).fastqsanger.gz : Compressed interleaved fastqsanger file containing the reads
- Staph_genome.fasta : Reference staph genome
- k_25,55,85 : Folders containing the contigs.fasta and contigs.tabular files generated at the end of a spades run. The fasta file contains the contigs generated in the assembly process while the tabular file summarizes them. You can also find a log file for the Spades run within this folder.

If you are running SPAdes, then you will see the results appear in your history when the app has finished running. While the app runs, please continue reading.

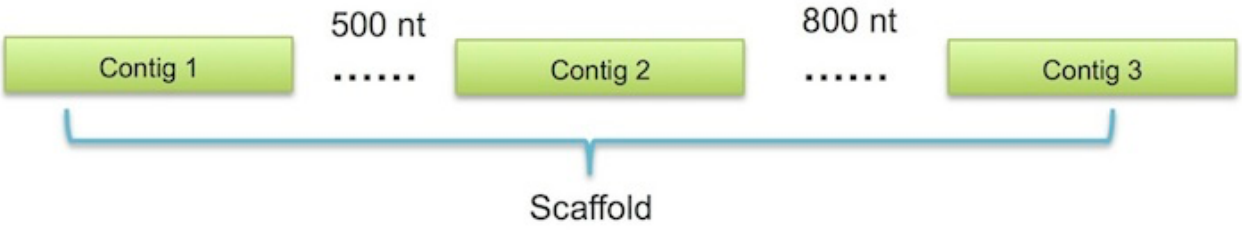
DEFINITIONS

There are many assembly tools, but none of them is perfect. Biologists therefore need to evaluate the quality of various assemblers by comparing their results. In our case, once we have run the SPAdes assembler on a set of reads, we need to test the quality of the resulting assembly.

Contig: A *contiguou*s segment of the genome that has been reconstructed by an assembly algorithm

Scaffold: An ordered sequence of contigs (possibly separated by gaps between them) that are reconstructed by an assembly algorithm. The order of contigs in a correctly assembled

scaffold corresponds to their order in the genome. Existing assemblers specify the approximate lengths of gaps between contigs in a scaffold.



N50 statistic: N50 is a statistic that is used to measure the quality of an assembly. N50 is defined as the maximal contig length for which all contigs greater than or equal to that length comprise at least half of the sum of the lengths of all the contigs. For example, consider the five toy contigs with the following lengths: [10, 20, 30, 60**,** 70]. Here, the total length of contigs is 190, and contigs of length 60 and 70 account for at least 50% of the total length of contigs (60 + 70 = 130), but the contig of length 70 does not account for 50% of the total length of contigs. Thus, N50 is equal to 60.

NG50 statistic: The NG50 length is a modified version of N50 that is defined when the length of the genome is known (or can be estimated). It is defined as the maximal contig length for which all contigs of at least that length comprise at least half of the length of the genome. NG50 allows for meaningful comparisons between different assemblies for the same genome. For example, consider the five toy contigs we considered previously: [10, 20, 30, 60, 70]. These contigs only add to 190 nucleotides, but say that we know that the genome from which they have been generated has length 300. In this example, the contigs of length 30, 60, and 70 account for at least 50% of the genome length (30 + 60 + 70 = 160); but the contigs of length 60 and 70 no longer account for at least 50% of the genome length (60 + 70 = 130). Thus, NG50 is equal to 30.

NGA50 statistic: If we already know a reference genome for a species, then we can test the accuracy of a newly assembled genome against this reference. The NGA50 statistic is a modified version of NG50 accounting for assembly errors (called **misassemblies**). To compute NGA50, errors in the contigs are accounted for by comparing contigs to a reference genome. All of the misassembled contigs are broken at **misassembly breakpoints**, resulting in a larger number of contigs with the same total length. For example, if there is a missasembly breakpoint at position 10 in a contig of length 30, this contig will be broken into contigs of length 10 and 20.

NGA50 is calculated as the NG50 statistic for the set of contigs resulting after breaking at misassembly breakpoints. For example, consider our example before, for which the genome length is 300. If the largest contig in [10, 20, 30, 60, 70] is broken into two contigs of length 20 and 50 (resulting in the set of contigs [10, 20, 20, 30, 50, 60]), then. contigs of length 20, 30, 50, and 60 account for at least 50% of the genome length (20 + 30 + 50 + 60 =

160). But contigs of length 30, 50, and 60 do not account for at least 50% of the genome length (30 + 50 + 60 = 140). Thus, NGA50 is equal to 20.

1. **Based on the above definition of N50, define N75 (3 points).**

N75 is the maximal contig length for which all contigs of at least that length comprise at least 75% of the sum of the lengths of all contigs.

2. **Compute N50 and N75 for the nine contigs with the following lengths: [20, 20, 30, 30, 60, 60, 80, 100, 200]. (4 points)**

N50 = 100, N75 = 60

3. **Say that we know that the genome length is 1000. What is NG50? (3 points)**

NG50 = 60.

4. **If the contig in our dataset of length 100 had a misassembly breakpoint in the middle of it, what would be the value of NGA50? (3 points)**

NGA50 = 50.

5. **Based on the definition of scaffolds, what information could we use to construct scaffolds from contigs? Justify your answer. (3 points)**

There are several reasonable answers. Three are given below.

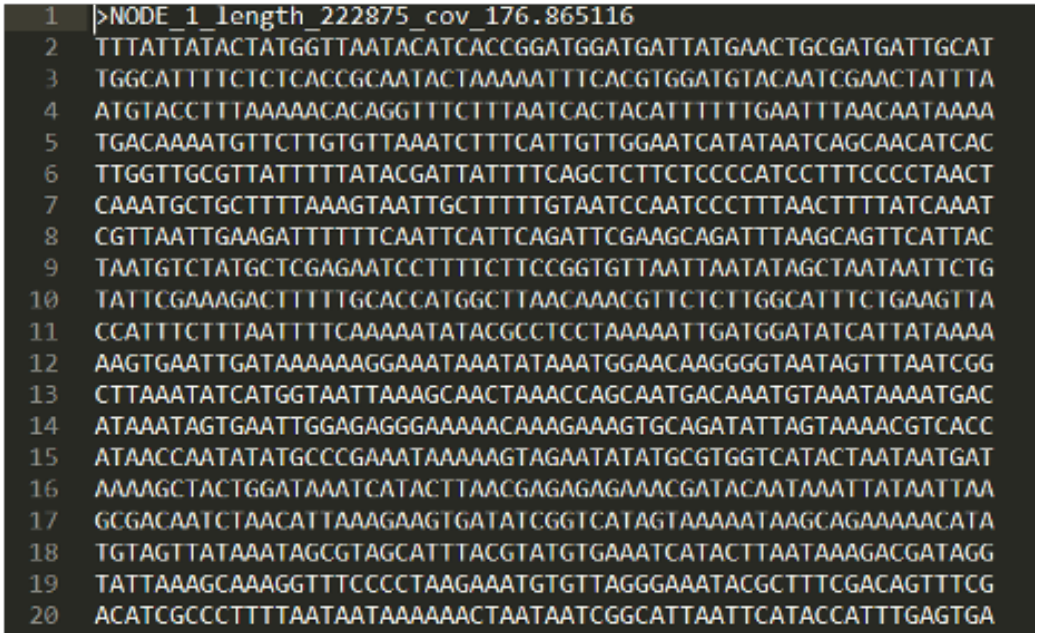
- Additional long reads could be generated in an attempt to find reads that bridge the gaps in contigs. In other words, if we find a long read that begins at the end of contig A, and ends at the beginning of contig B, then we can conclude that the read extends across the gap between the contigs.
- Contigs could be compared against a "reference genome", i.e., a complete genome sequenced from the same species (often at greater cost). The order of the contigs in the reference genome would indicate the order of the contigs in the desired scaffold.
- Information from read-pairs could be used. In particular, if the first read in a read-pair maps to contig A, and the second read in a read-pair maps to contig B, and we know the distance between the paired reads, then we can infer the distance between contigs A and B. By gathering this information for different pairs of contigs, we may be able to infer distances between contigs and therefore their ordering with respect to each other.

We will now answer questions concerning the assembly of the *Staphylococcus* reads. If you are running SPAdes yourself, then continue here as soon as your assembly of the Staph reads has completed. If you are following the completed runs, you can continue here at any time.

Consider the following three statistics:

- N50.
- The number of ***long* contigs**, i.e., contigs with length ≥ 1000 nucleotides. Biologists are mainly interested in long contigs and often discard short contigs, since short contigs often harbor only fragments of genes rather than complete genes.
- The total length of *long* contigs. This statistic can be combined with N50 and the number of long contigs; a good assembly is one that has relatively few long contigs, but the total length of long contigs is high, as is N50.

These three statistics can be found by analyzing the contigs.fasta file or the contigs.tabular file generated at the end of SPAdes execution. The tabular file is a summary sheet of all the contigs present in the assembly. Below is an example snapshot of what the fasta and tabular file look like, respectively, for $k = 25$.



	#name	length	coverage
1	NODE_1	222875	176.865116
2	NODE_2	145496	167.394848
3	NODE_3	114264	198.360367
4	NODE_4	88090	214.401419
5	NODE_5	87970	162.457036
6	NODE_6	79702	196.246156
7	NODE_7	78752	186.616795
8	NODE_8	78364	220.639171
9	NODE_9	78033	167.713837
10	NODE_10	77760	219.710208
11	NODE_11	76546	155.661936
12	NODE_12	75624	171.901004
13	NODE_13	64040	236.320503
14	NODE_14	63392	177.547036
15	NODE_15	61533	176.462428
16	NODE_16	59595	239.788249
17	NODE_17	48958	204.791920
18	NODE_18	44207	253.540175
19	NODE_19	42588	215.297253

You will use the Quality Assessment Tool for Genome Assembly **QUAST** (Gurevich *et al*, 2013) to evaluate the quality of your assembly using the Staph reference genome as the gold standard.

- Download the contigs.fasta file as part of the SPAdes output from each value of k.
- For each of the three files, go to QUAST (http://cab.cc.spbu.ru/quast/) and upload your contigs.fasta file with the "Add files" button.
- Leave the "Scaffolds" and "Find genes" boxes unchecked and keep the indicator on "Prokaryotic."
- Click on the "Another genome" link underneath "Genome." Fill in a name and upload the staph_genome.fasta file that we provided for the "Reference" file. (Note: we provide this file as a .txt, you will need to save it as .fasta). Leave the other two inputs ("Genes" and "Operons") blank and click "Evaluate."
- A link to the report should appear on the right side of the page in a few moments. Evaluate the report and answer the following questions.
- If you had difficulties running QUAST, please find the required reports for this part of the assignment in the Quast reports folder here : https://drive.google.com/drive/folders/1L-cfjiyd9RJOwY1Slau-wQHGmO1g1xft?usp=sharing
- Select the report (a html file) corresponding to the k value chosen. Download it and open it using any web browser.
- All the required statistics are present on the left panel of the report. Click on 'Extended report' to see all statistics required for the following questions.

6. **First, fill in the 9 missing values in the following 3 x 3 table (9 points; 1 point each):**

k	N50	#long contigs	total length of long contigs
25			
55			
85			

1 point for each of the following table values. Values may differ slightly due to updated software versions.

k	N50	#long contigs	total length of long contigs
25	59,595	110	2,802,857
55	159,616	38	2,821,839
85	188,896	37	2,825,752

7. **Which assembly performed the best in terms of each of these statistics? Justify your answer.Why do you think that the value you chose performed the best? (3 points)**

The total length of long contigs is about the same for all three values of *k*. Accordingly, we conclude that the assembly using $k = 85$ performed the best because it has a larger value of N50 and fewer contigs than $k = 25$, while having the same number of contains as $k = 55$.

$k = 85$ performs the best because if the reads are too short ($k = 25$ or 55), then the reads contain too little information, and repeats may make it difficult to identify where a read came from.

Please use the grading scheme below:

3 points available

1 point for identifying $k = 85$;

1 point for a reasonable justification of why $k = 85$ was the best choice according to statistics;

2 points for attempting a reasonable explanation of why $k = 85$ wound up being the best value of *k*.

Answer the following two questions using the QUAST reports.

8. **How many misassemblies were there? (2 points)**
9. **How significant is the effect of misassemblies on the resulting assembly? (2 points)**

There were 27 misassemblies. (number may vary slightly depending on the version used.) For k = 85 there were only about 37 long contigs (number may vary slightly depending on the version used), meaning that the misassemblies are likely causing most of the long contigs to have been broken into pieces.

3 points available

1 point: correctly identifying the number of misassemblies

2 points: identifying that the number of misassemblies is significant and giving a reasonable explanation

10. **What are NG50 and NGA50 for the QUAST run? (2 points)**
11. **How do they compare with the value of N50 that you previously calculated? Why? (2 points)**

(Note: values may vary slightly based on updates to the software.)

(1 point): NG50 = 202,267.

(1 point): NGA50 = 87,161.

(1 point): NG50 is larger than the N50 value previously obtained. (Note: it roughly corresponds to the value of N50 that was previously obtained, because the contigs generated cover the entire genome.)

(1 point): However, NGA50 is about half as large as N50 because of the effects of misassemblies.