due in 2hr 6m

Submission Phase

1. Do assignment ☐ (/assembly-002/human_grading/view/courses/974903/assessments/3/submissions)

Evaluation Phase

2. Evaluate peers 🔒 (/assembly-002/human_grading/view/courses/974903/assessments/3/peerGradingSets)

Results Phase

3. See results 🔒 (/assembly-002/human_grading/view/courses/974903/assessments/3/results/mine)

☐ In accordance with the Honor Code, I certify that my answers here are my own work, and that I have appropriately acknowledged all external sources (if any) that were used in this work.

Save draft | Submit for grading

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.

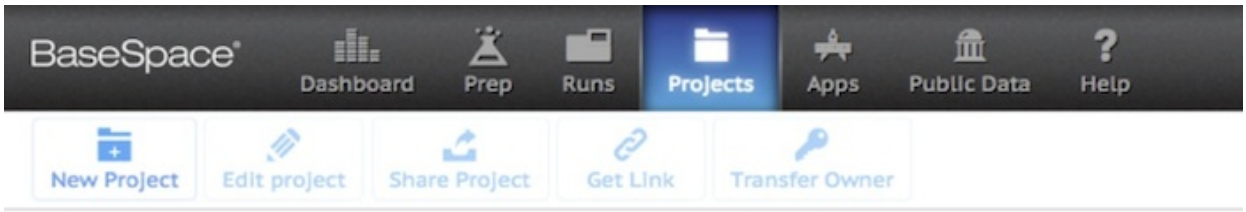
ASSEMBLY

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 Illumina's **BaseSpace** service. Please follow these step-by-step instructions to register on BaseSpace and run SPAdes:

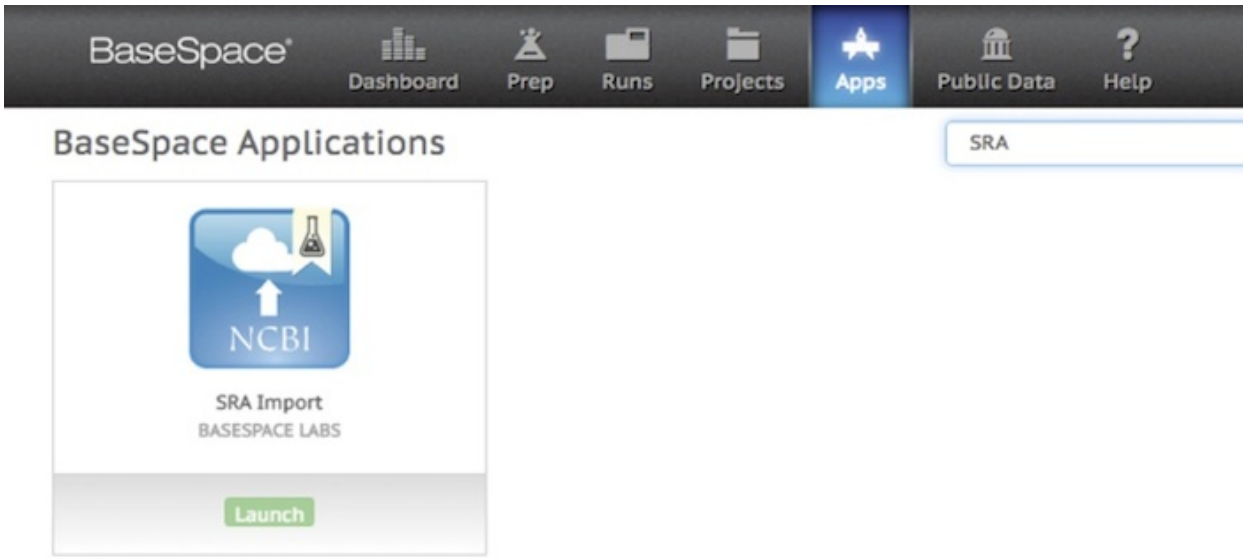
Register: Create an account on BaseSpace here (https://accounts.illumina.com/Account/Register?service=basespace&ReturnUrl=http%3a%2f%2fbasespace.illumina.com%2fdashboard). You will need to fill out all fields.

After logging into BaseSpace, you will see the following dashboard and menu at the top of the page. Click on "Projects" and then create a new project. Name it whatever you like. All of the following analysis will be

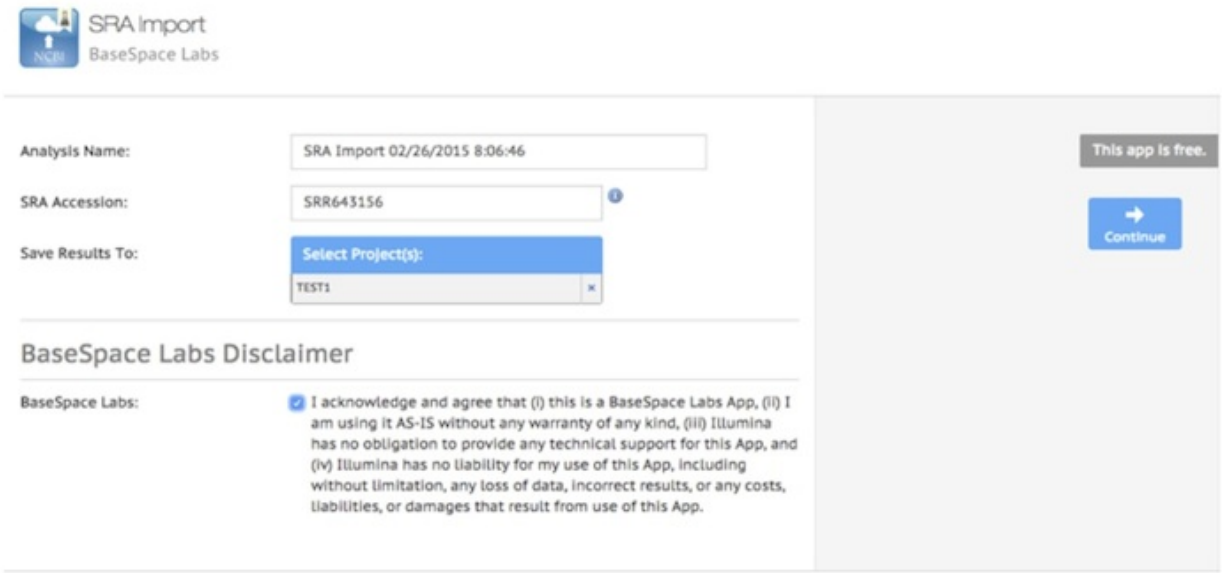
performed under this project.



Next, we need to import our data. For this assignment, we will import raw data directly from the SRA database. Click on “Apps” in the top menu and search for the SRA import app.

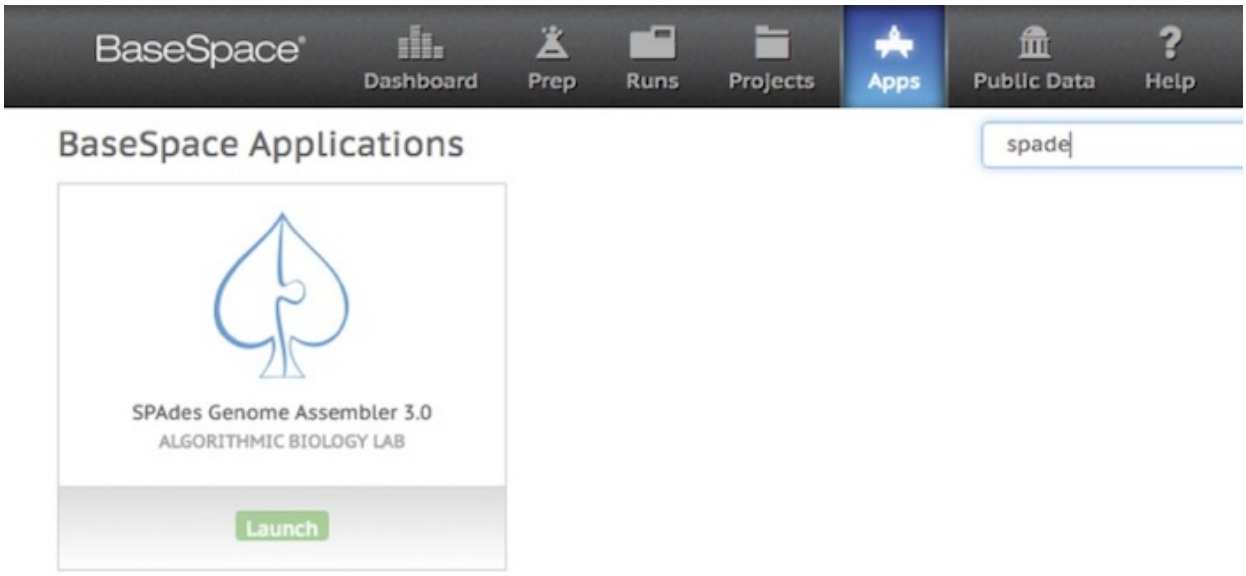


Launch the app and use accession number **SRR643156** to import the data. Click “Save to,” select your project, agree to the disclaimer, and click the continue button to import the data. It may take about 30 minutes for the app to begin execution and to load the data. When the app is finished, you will be able to see the imported files and related information in your project folder.



Next, we will use SPAdes to assemble the genome. Go to the App page again and search for SPAdes. Launch SPAdes (**please use version 3.0**) and click “Select Sample(s).” Click on the top sample, labeled **SRR643156**, and click ”Confirm.” Save results to the project that you made and specify the parameters below (use default settings except for the value of *k*-mer size). For this homework assignment, you will need to run the program three times (for $k = 25$, $k = 55$, and $k = 85$) to investigate how the choice of parameter *k* affects the assembly quality. Each run takes about 30 min, but you can queue all three runs

at once – you will need to repeat all of the steps in this paragraph for each run. You will see the results appear in your project folder when the app has finished running. While the app runs, please continue reading.

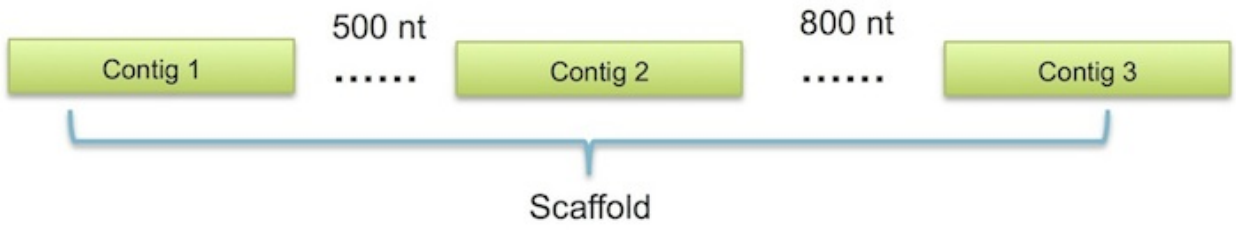


DEFINITIONS

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.

Based on the above definition of N50, define N75.

B *I* Link <code> Math Edit: Rich Preview

Words: 0 / 100

Compute N50 and N75 for the nine contigs with the following lengths: [20, 20, 30, 30, 60, 60, 80, 100, 200].

B *I* Link <code> Math Edit: Rich Preview

Words: 0 / 100

Say that we know that the genome length is 1000. What is NG50?

B *I* Link <code> Math Edit: Rich Preview

Words: 0 / 100

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?

B | I | Link | <code> | Math | Edit: Rich | Preview

Words: 0 / 100

Based on the definition of scaffolds, what information could we use to construct scaffolds from contigs? Justify your answer.

B | I | Link | <code> | Math | Edit: Rich | Preview

Words: 0 / 300

Continue here as soon as your assembly of the Staph reads has completed.

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.

Fill in the 9 missing values in the following 3 x 3 table:

k	**N50**	**#long contigs**	**total length of long contigs**
25			

55
85

B | I | Link | <code> | Math | Edit: Rich | Preview

Words: 0 / 200

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?

B | I | Link | <code> | Math | Edit: Rich | Preview

Words: 0 / 200

(Multiple choice) When you increase the length of *k*-mers, the de Bruijn graph _____. Justify your answer.

A) Becomes more tangled.
B) Contains more nodes.
C) Becomes less tangled.
D) Remains the same.

B | I | Link | <code> | Math | Edit: Rich | Preview

Words: 0 / 200

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 the best assembly you chose for question #8 above.
- Go to QUAST (http://quast.bioinf.spbau.ru/ (http://quast.bioinf.spbau.ru/)) 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 (http://bioinformaticsalgorithms.com/software_challenges/assembly/staph_genome.txt) 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.

1. How many misassemblies were there?

2. How significant is the effect of misassemblies on the resulting assembly?

B | I | Link | <code> | Math | Edit: Rich | Preview

1. What are NG50 and NGA50?

2. How do they compare with the value of N50 that you previously calculated? Why?

B | I | Link | <code> | Math | Edit: Rich | Preview

What is the known species of *Staphylococcus* that is most similar to the species that you assembled?

B | I | Link | <code> | Math | Edit: Rich | Preview

In accordance with the Honor Code, I certify that my answers here are my own work, and that I have appropriately acknowledged all external sources (if any) that were used in this work.

Save draft | Submit for grading