Bacterial Genome Assembly

Chris Fields

PowerPoint by Chris Fields

Introduction

Exercise

1. Perform a bacterial genome assembly using PacBio HiFi data.

2. Evaluation and comparison of different datasets and parameters.

3. View the best assembly in Bandage.

Premise

1. We have sequenced the genomic DNA of a bacterial species that we are very interested in. Using other methods, we have determined that its genome size is approximately 1.7 Mbp.

2. We chose to use Pacific Biosciences HiFi technology for performing this analysis because our genome of interest is relatively small and PacBio HiFi gives us both long reads (700bp-20kbp) that are 99% accurate.

Dataset Characteristics

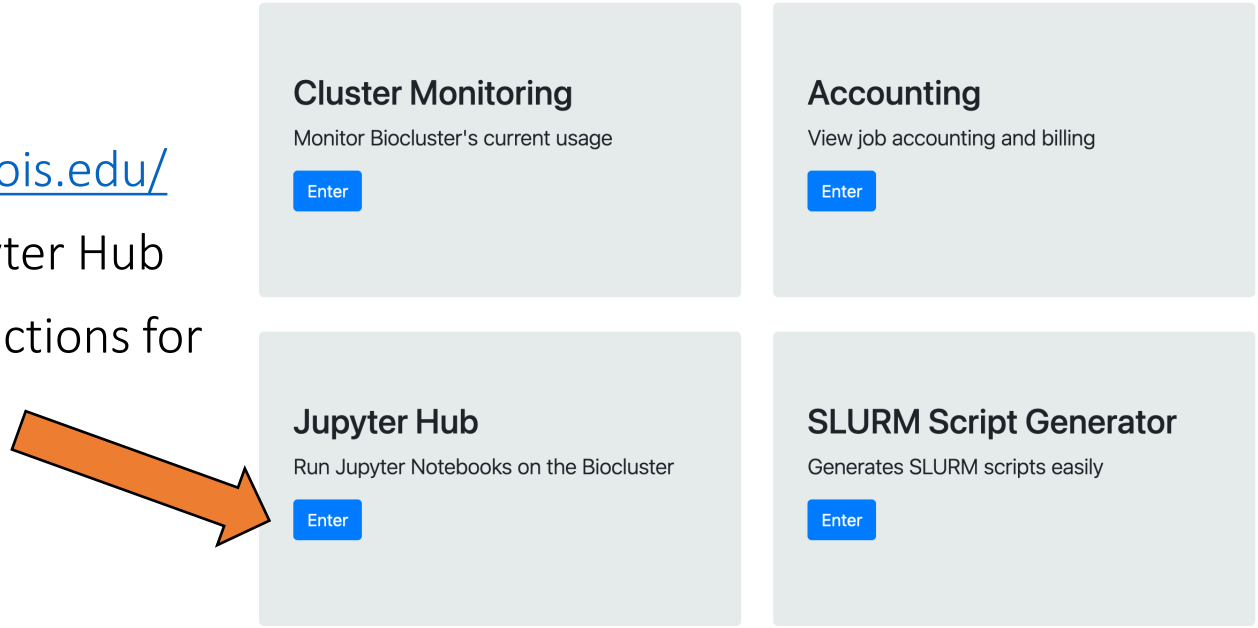
Dataset #	FQ Name	File Size
1	dataset1.fastq.gz	144 MB
2	dataset2.fastq.gz	36 MB
3	dataset3.fastq.gz	18 MB
4	dataset4.fastq.gz	7.1 MB

The assembler can read compressed files, so the FASTQ* files are compressed to save disk space. We will see what the data look like in a bit.

* FASTQ -> “FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letters and quality scores are encoded with a single ASCII character for brevity”. Excerpted from http://en.wikipedia.org/wiki/Fastq

Step 1: Start Jupyter Hub

1. Go to https://biocluster.igb.illinois.edu/
2. Click on **Enter** under Jupyter Hub
3. Follow general setup directions for all labs



Assembly

We will be using the hifiasm assembler, which is a very fast assembler developed for highly accurate long reads such as PacBio HiFi. The notebook, if copied over, will automatically run all the steps for genome assembly for you.

Step 1: Get sequence statistics

We know the size of the files, but we don't know **how many reads** there are, what the **maximum and minimum length** is, **total bases**, and so forth. This would be good to know, since we want to make sure we have adequate coverage for an assembly.

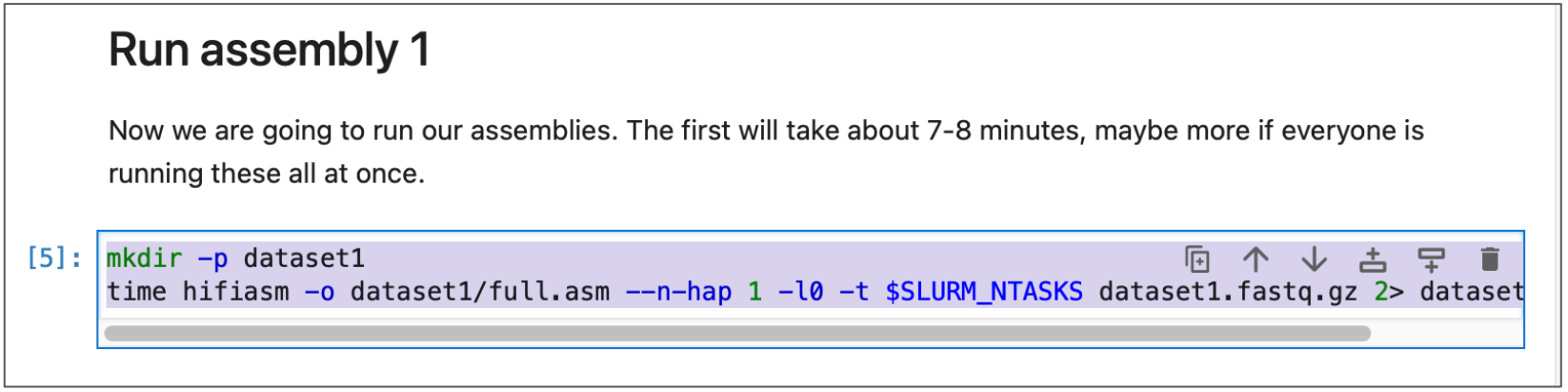
We can use a tool called 'seqkit' for this. The step in the notebook:

```
seqkit stats --quiet dataset*.fastq.gz
```

Will generate basic stats on the raw sequence reads. We can use this in a bit to summarize the data we have

Step 2: Assemblies

For the first assembly we use **dataset1.fastq.gz** (144MB). Here is the part in the notebook. Let's break this line down:



`time`	*Tool: time following command takes to complete*
`hifiasm`	*Tool: name of the assembler*
`-o dataset1/full.asm`	*Output prefix*
`--n-hap 1`	*Number of haplotypes (1)*
`-l0`	*Skips removing alternative haplotigs*
`-t $SLURM_NTASKS`	*Set the threads to the cores being used (2)*
`dataset1.fastq.gz`	*The reads we use for assembling the genome*
`2> dataset1/full.log`	*Redirect output to a log file named 'full.log'*

HiFiAsm Output: Legend

Once everything is finished you will have numerous files. Key ones are highlighted below:

File	Meaning
prefix.p_ctg.gfa	Primary assembly (a *haploid* representation). GFA format
prefix.p_ctg.fasta	Primary assembly (a *haploid* representation). FASTA format
prefix.a_ctg.gfa	Alternate assembly contig graph (alleles not in primary assembly). GFA format
prefix.r_utg.gfa	Raw unitig graph. GFA format. Keeps all haplotype information, including somatic mutations and recurrent sequencing errors.

Assembly Evaluation

What metrics do we use to evaluate the assembly?

Assembly Evaluation: Skeleton

	dataset1	dataset2	dataset3	dataset4
Genome Size (Mb)				
N50 (Mb)				
Number of contigs				
Longest contig (Mb)				
Shortest contig (bp)				
Mean contig size (Kb)				
GC content				

Definition of N50:

"Given a set of contigs, each with its own length, the *N50* length is defined as the shortest sequence length at 50% of the genome. It can be thought of as the point of half of the mass of the distribution. For example, 9 contigs with the lengths 2,3,4,5,6,7,8,9,and 10, their sum is 54, half of the sum is 27. 50% of this assembly would be 10 + 9 + 8 = 27 (half the length of the sequence). Thus the N50 = 8".

Excerpted from https://en.wikipedia.org/wiki/N50,_L50,_and_related_statistics#N50

Compare Assembly Statistics

	dataset1	dataset2	dataset3	dataset4
Genome Size (Mb)	1.663877	1.645745	1.437603	1.347800
N50 (Mb)	1.663877	1.645745	0.799713	0.105349
Number of contigs	1	1	4	18
Longest contig (bp)	1,663.877	1,645,745	799,713	325,120
Shortest contig (bp)	1,663,877	1,645,745	14,705	17,658
Mean contig size (bp)	1,663,877	1,645,745	359,401	74,878
GC content	39.17	39.19	39.03	39.17

Genome size is ~1.7 Mb; two of these assemblies are close. The genome coverage (the number of times each base is covered by a read) for dataset1 is about 44x, dataset2 is 22x, dataset3 is 11x and dataset4 is 4x. The lower the coverage the fewer bases recovered and more fragmented the genome.

Also note how many contigs each has; datasets 1 & 2 have fully assembled chromosomes!

Key takeaways

- Perform a basic bacterial genome assembly using hifiasm
- Understand basic file formats used or produced (FASTQ, FASTA, GFA)
  - FASTQ – sequence format for data generated from sequencing instruments. Includes quality scores per position indicating base call accuracy. Input for the assembler is normally FASTQ
  - FASTA – simple sequence format with name, optional description, and sequence. Output is normally FASTA and/or GFA
  - GFA – graph format used for capturing sequence graphs. Can be from assemblies, variation data, splice variants from RNA-Seq, or pangenome analyses. Can generate FASTA from this
- Understand the metrics used for evaluating reads (total bases, total reads, mean read length) and assemblies (total bases, N50)
- Understand the effect that decreased coverage has on genome assembly
- Basic visualization of the genome assembly graph using Bandage