


Deliverable 2D

All Unit 2 work is due by Mar 10.


Each deliverable is complete when you have:

- answered each question
- saved a terminal session or screenshots demonstrating your performance of the commands (on Discovery or posted on GitHub)
- indicated in a “TOC” (table of contents) file where your work is found (GitHub repo or specific path/name_of_file on the cluster).

Beginning of exercise: You have three data sets consisting of NGS sequencing reads:

-  lambda phage from bowtie tutorial
- Ppar reads from Marine Genomics course
- Day lab reads

Perform this first set of operations for all three datasets.

 1. Prepare a separate directory for each project, with subdirectories as needed.

2. First let's describe the data we have. For each set of fastq files, describe:

1. How many reads are in each file.
 1. Lambda phage:
 1. longreads.fq: 6000
 2. reads_1.fq: 10000
 3. reads_2.fq: 10000
 2. Ppar (sea cucumber):
 1. 1000 reads in each of the six fastq files
 3. Day data:
 1. 10_S1_L001_R1_001.fastq.gz:28893152
 2. 10_S1_L001_R2_001.fastq.gz:28893152
 3. 11_S1_L001_R1_001.fastq.gz:29291552
 4. 11_S1_L001_R2_001.fastq.gz:29291552
 5. 12_S1_L001_R1_001.fastq.gz:29043844
 6. 12_S1_L001_R2_001.fastq.gz:29043844
 7. 13_S1_L001_R1_001.fastq.gz:29023016
 8. 13_S1_L001_R2_001.fastq.gz:29023016
 9. 14_S1_L001_R1_001.fastq.gz:24730770
 10. 14_S1_L001_R2_001.fastq.gz:24730770
 11. 15_S1_L001_R1_001.fastq.gz:28387419
 12. 15_S1_L001_R2_001.fastq.gz:28387419

13. 1_S1_L001_R1_001.fastq.gz.fastq:32833451
14. 1_S1_L001_R2_001.fastq.gz.fastq:32833451
15. 2_S1_L001_R1_001.fastq.gz.fastq:33738336
16. 2_S1_L001_R2_001.fastq.gz.fastq:33738336
17. 3_S1_L001_R1_001.fastq.gz.fastq:35731214
18. 3_S1_L001_R2_001.fastq.gz.fastq:35731214
19. 4_S1_L001_R1_001.fastq.gz.fastq:36678316
20. 4_S1_L001_R2_001.fastq.gz.fastq:36678316
21. 5_S1_L001_R1_001.fastq.gz.fastq:36972680
22. 5_S1_L001_R2_001.fastq.gz.fastq:36972680
23. 6_S1_L001_R1_001.fastq.gz.fastq:31401357
24. 6_S1_L001_R2_001.fastq.gz.fastq:31401357
25. 7_S1_L001_R1_001.fastq.gz.fastq:35536673
26. 7_S1_L001_R2_001.fastq.gz.fastq:35536673
27. 8_S1_L001_R1_001.fastq.gz.fastq:24498096
28. 8_S1_L001_R2_001.fastq.gz.fastq:24498096
29. 9_S1_L001_R1_001.fastq.gz.fastq:29794050
30. 9_S1_L001_R2_001.fastq.gz.fastq:29794050

2. The length of the reads and if they are single or paired-end

1. Lambda phage:

1. The length of the reads in each fastq file vary. A text file with the length of each read in each file can be found in my bowtie_lambda_phage_tutorial_20240221 directory in the Course directory.

2. Single-end / Paired-end??

2. Ppar (sea cucumber):

1. Each read is 80 bases in length.
2. Single-end.

3. Day data:

1. Each read is 150 bases in length.
2. Paired-end.

3. The overall quality of the reads and anything to be concerned about

1. Lambda phage:

1. The overall quality is low: the per base sequence quality is only about 16-18 and the per sequence quality score profile shows a small peak at quality score 4-5.
2. Other things to be concerned with: per base sequence content for longer reads, per sequence GC content distribution, per base N content, and varying sequence lengths shown in the sequence length distribution.

2. Ppar (sea cucumber):

1. Low quality reads: there are warning or failure symbols for the per base sequence quality and per sequence quality scores.

2. Other things to be concerned with: per sequence GC content, and several sequences are overrepresented.
3. Day data:
 1. Overall very high quality data.
 2. Other things to be concerned with: per tile sequence quality, sequence duplication levels, and adaptor content for the ends of the sequences. Per base GC content is not a concern; the summary indicates a failure however the pattern observed for bases 1-9 simply indicate that RNA was sequenced.
4. Whether they appear to have adapter sequences that need to be trimmed
 1. Lambda phage:
 1. No apparent adapter sequences
 2. Ppar (sea cucumber):
 1. Yes, there appears to be an adapter sequence that needs to be trimmed based on looking at the per base sequence content for bases 1-5 for several of the fastq files.
 3. Day data:
 1. No apparent adapter sequences
3. Collect quality control data on the reads, in the form of an .html file produced by fastqc.
4. If the sequences of a project need trimming, perform this step as described in the Marine Genomics tutorial, using cutadapt.

****For now, leave the Day data and perform the rest of the operations only on the lambda phage and Ppar (sea cucumber) data**.**

5. Index the genome for each species using bowtie.
6. Map the reads to the genome using bowtie. (How is the command used in the Marine Genomics tutorial different from that used in the bowtie tutorial?)
7. Convert the files containing mapped reads from sam to bam files using samtools.
8. There are two programs for determining variants (positions where the read sequences differ from the reference genome) that we were introduced to: bcftools and angsd. Use each of these to call variants for the lambda phage and sea cuke data, and compare the results.

Note that the bcftools protocol requires input files to be in a "sorted.bam" format, whereas angsd takes bam files as input.

Additional References:

- How to count length of reads in FASTQ file:
<https://www.biostars.org/p/459116/#:~:text=fq%2Ffastq.,%7Bprint%20length%7D'%20in%20put.>
-