Genome Analysis

Unit 2 Deliverable D

Working with all 3 Data Sets

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

1. **How many reads in each file**

**Use gep -c ‘@’**

Lambda

* Reads\_1.fq: 16874 (paired)
* Reads\_2.fq: 16794 (paired)
* Longreads.fq: 10458 (single)

Ppar

* SRR6805880.tiny.fastq: 1000
* SRR6805881.tiny.fastq: 1248
* SRR6805882.tiny.fastq: 1104
* SRR6805883.tiny.fastq: 1134
* SRR6805884.tiny.fastq: 1173
* SRR6805885.tiny.fastq: 1258

Day (paired)

* 10\_S1\_L001\_R1\_001.fastq:28893152
* 10\_S1\_L001\_R2\_001.fastq:28893152
* 11\_S1\_L001\_R1\_001.fastq:29291552
* 11\_S1\_L001\_R2\_001.fastq:29291552
* 12\_S1\_L001\_R1\_001.fastq:29043844
* 12\_S1\_L001\_R2\_001.fastq:29043844
* 13\_S1\_L001\_R1\_001.fastq:29023016
* 13\_S1\_L001\_R2\_001.fastq:29023016
* 14\_S1\_L001\_R1\_001.fastq:24730770
* 14\_S1\_L001\_R2\_001.fastq:24730770
* 15\_S1\_L001\_R1\_001.fastq:28387419
* 15\_S1\_L001\_R2\_001.fastq:28387419
* 1\_S1\_L001\_R1\_001.fastq:32833451
* 1\_S1\_L001\_R2\_001.fastq:32833451
* 2\_S1\_L001\_R1\_001.fastq:33738336
* 2\_S1\_L001\_R2\_001.fastq:33738336
* 3\_S1\_L001\_R1\_001.fastq:35731214
* 3\_S1\_L001\_R2\_001.fastq:35731214
* 4\_S1\_L001\_R1\_001.fastq:36678316
* 4\_S1\_L001\_R2\_001.fastq:36678316
* 5\_S1\_L001\_R1\_001.fastq:36972680
* 5\_S1\_L001\_R2\_001.fastq:36972680
* 6\_S1\_L001\_R1\_001.fastq:31401357
* 6\_S1\_L001\_R2\_001.fastq:31401357
* 7\_S1\_L001\_R1\_001.fastq:35536673
* 7\_S1\_L001\_R2\_001.fastq:35536673
* 8\_S1\_L001\_R1\_001.fastq:24498096
* 8\_S1\_L001\_R2\_001.fastq:24498096
* 9\_S1\_L001\_R1\_001.fastq:29794050
* 9\_S1\_L001\_R2\_001.fastq:29794050

1. **Length of reads (single or paired)**

Lambda reads paired

-length visible in Quality report, varied (40-354z0

Ppar reads are single

-length visible in quality reports (80)

Day data is paired

-length visible in quality reports

1. **Overall quality of reads?**

Lambda: using the head command to examine the reads I see many # and $ which may indicate the quality is not the best

Ppar: using the head command I again observed many characters rather than letters which may indicate that the data is not the best quality

Day: Using head I examined the reads and saw that there were more letter in the phred scores indicating the reads are of better quality

1. **Adapter sequences present**

Lambda: does not appear to have adapter sequences that need trimming

Ppar: I see the adapter sequence TGCAG

Day: I don’t see an adapter sequence

**Quality Control Data**

Lambda:

Reads\_1.fq

A screenshot of a computer

Description automatically generated

Reads\_2.fq

A screenshot of a report

Description automatically generated

Longreads.fq

A screenshot of a computer

Description automatically generated

Ppar:

SRR6805880.tiny.fastq.gz

A screenshot of a report

Description automatically generated

SRR6805881.tiny.fastq.gz

A screenshot of a report

Description automatically generated

SRR6805882.tiny.fastq.gz

A screenshot of a report

Description automatically generated

SRR6805883.tiny.fastq.gz

A screenshot of a report

Description automatically generated

SRR6805884.tiny.fastq.gz

A screenshot of a report

Description automatically generated

SRR6805885.tiny.fastq.gz

A screenshot of a report

Description automatically generated

Day-Sample of quality reports due to large amount

[**10\_S1\_L001\_R1\_001\_fastqc.html**](https://ood.discovery.neu.edu/pun/sys/dashboard/files/fs/courses/BIOL3411.202430/students/ali.may/Assignments/Day/10_S1_L001_R1_001_fastqc.html)

A screenshot of a report

Description automatically generated

[**12\_S1\_L001\_R2\_001\_fastqc.html**](https://ood.discovery.neu.edu/pun/sys/dashboard/files/fs/courses/BIOL3411.202430/students/ali.may/Assignments/Day/12_S1_L001_R2_001_fastqc.html)

A screenshot of a report

Description automatically generated

**Cut adapter sequences as necessary**

Lambda- no adapter sequences

Ppar: need to trim adapter

-use cutadapt, pre load anaconda3/2022.05, write shell script to trim all files

A computer screen shot of a black screen

Description automatically generated

Day- no adapter to cut

**Index the genome for each species using bowtie**

Ppar 

Lambda



**Map the reads using bowtie**

Ppar

A screen shot of a computer code

Description automatically generated

Lambda

A black screen with white text

Description automatically generated

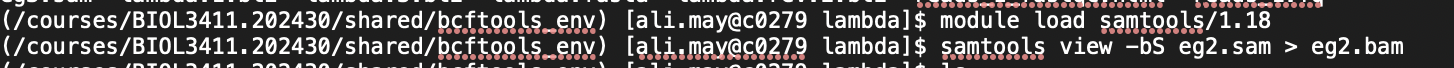
**Convert the files containing mapped reads from sam to bam files using samtools**

Ppar

A black screen with white text

Description automatically generated

Lambda



**Call Variants**

Ppar-angsd

1. Convert sam to bam

A black screen with white text

Description automatically generated

1. Activate

source activate /courses/BIOL3411.202430/shared/angsd\_env/

1. Call variants

/courses/BIOL3411.202430/shared/angsd\_env/angsd/angsd -bam bam.filelist -GL 1 -out genotype\_likelihoods -doMaf 2 -SNP\_pval 1e-2 -doMajorMinor 1

1. Examining the contents of the new mafs.gz file was done using gunzip and cat

A screen shot of a computer

Description automatically generated

Lambda-bcftools

1. Convert sam to bam 

Activate bcf (pre load : module load anaconda3/2022.05)

1. 
2. Call variants 
3. View results

