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

2. **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

3. **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

4. **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

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	reads_1.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10000
Sequences flagged as poor quality	0
Sequence length	40-354
%GC	49

Per base sequence quality

Reads_2.fq

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	reads_2.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10000
Sequences flagged as poor quality	0
Sequence length	40-366
%GC	49

Per base sequence quality

Longreads.fq

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	longreads.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	6000
Sequences flagged as poor quality	0
Sequence length	40-2561
%GC	50

Per base sequence quality

Ppar:

SRR6805880.tiny.fastq.gz

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	SRR6805880.tiny.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1000
Sequences flagged as poor quality	0
Sequence length	80
%GC	43

Per base sequence quality

SRR6805881.tiny.fastq.gz

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	SRR6805881.tiny.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1000
Sequences flagged as poor quality	0
Sequence length	80
%GC	42

Per base sequence quality

SRR6805882.tiny.fastq.gz

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	SRR6805882.tiny.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1000
Sequences flagged as poor quality	0
Sequence length	80
%GC	42

Per base sequence quality

SRR6805883.tiny.fastq.gz

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	SRR6805883.tiny.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1000
Sequences flagged as poor quality	0
Sequence length	80
%GC	42

Per base sequence quality

SRR6805884.tiny.fastq.gz

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	SRR6805884.tiny.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1000
Sequences flagged as poor quality	0
Sequence length	80
%GC	42

Per base sequence quality

SRR6805885.tiny.fastq.gz

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	SRR6805885.tiny.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1000
Sequences flagged as poor quality	0
Sequence length	80
%GC	42

Per base sequence quality

Day-Sample of quality reports due to large amount

10_S1_L001_R1_001_fastqc.html

FastQC Report

Summary

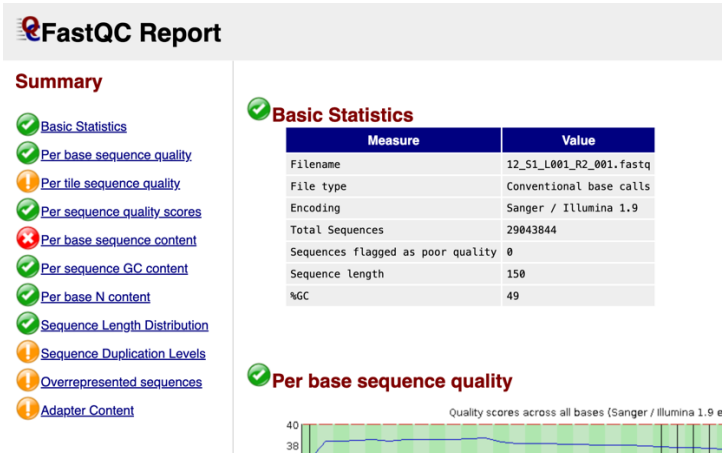
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	10_S1_L001_R1_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	28893152
Sequences flagged as poor quality	0
Sequence length	150
%GC	49

Per base sequence quality

12_S1_L001_R2_001_fastqc.html



FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	12_S1_L001_R2_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	29043844
Sequences flagged as poor quality	0
Sequence length	150
%GC	49

Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 e

40

38

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

```
  GNU nano 2.3.1                                                    File: trim.sh

for filename in *.tiny.fastq.gz
do

  base=$(basename $filename .tiny.fastq.gz)
  echo ${base}

  cutadapt -g TGCAG ${base}.tiny.fastq.gz -o ${base}.tiny_trimmed.fastq.gz

done
```

Day- no adapter to cut

Index the genome for each species using bowtie

Ppar

```
(/courses/BIOL3411.202430/shared/cutadapt_env) [ali.may@c0325 week4]$ module load bowtie/2.3.5.1
(/courses/BIOL3411.202430/shared/cutadapt_env) [ali.may@c0325 week4]$ bowtie2-build Ppar_tinygenome.fna.gz Ppar_tinygenome
```

Lambda

```
[ali.may@login-00 lambda]$ bowtie2-build lambda.fasta lambda
```

Map the reads using bowtie

Ppar

```
  GNU nano 2.3.1                                                    File: map.sh

for filename in *.tiny_trimmed.fastq.gz
do

  base=$(basename $filename .tiny_trimmed.fastq.gz)
  echo ${base}

  bowtie2 -x Ppar_tinygenome -U ${base}.tiny_trimmed.fastq.gz -S ${base}.sam

done
```

Lambda

```
[ali.may@login-00 lambda]$  bowtie2 -x lambda -U reads_1.fq -S egl.sam
10000 reads; of these:
  10000 (100.00%) were unpaired; of these:
    596 (5.96%) aligned 0 times
    9404 (94.04%) aligned exactly 1 time
    0 (0.00%) aligned >1 times
94.04% overall alignment rate
```

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

Ppar

```
  GNU nano 2.3.1                                                    File: bam.sh

for filename in *.sam
do

  base=$(basename $filename .sam)
  echo ${base}

  samtools view -bhS ${base}.sam | samtools sort -o ${base}.bam

done
```

Lambda

```
(/courses/BIOL3411.202430/shared/bcftools_env) [ali.may@c0279 lambda]$ module load samtools/1.18
(/courses/BIOL3411.202430/shared/bcftools_env) [ali.may@c0279 lambda]$ samtools view -bS eg2.sam > eg2.bam
```

Call Variants

Ppar-angsd

1. Convert sam to bam

```
  GNU nano 2.3.1                                                    File: bam.sh

for filename in *.sam
do

  base=$(basename $filename .sam)
  echo ${base}

  samtools view -bhS ${base}.sam | samtools sort -o ${base}.bam

done
```

2. Activate

```
source activate /courses/BIOL3411.202430/shared/angsd_env/
```

3. 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

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

```
(/courses/BIOL3411.202430/shared/cutadapt_env) [ali.may@c0325 week4]$ gunzip genotype_likelihoods.mafs.gz
(/courses/BIOL3411.202430/shared/cutadapt_env) [ali.may@c0325 week4]$ cat *.mafs
chromo   position        major   minor   unknownEM        pu-EM     nInd
KN882277.1       41498   G       T       0.332737         3.127339e-03    3
KN885472.1       10712   C       G       0.126253         1.118604e-03    6
KN885472.1       10741   T       A       0.205533         2.729806e-03    6
KN885472.1       10746   C       T       0.113382         1.394211e-03    6
KN894013.1       22082   T       C       0.098327         3.551274e-03    2
KN894013.1       22084   C       T       0.106562         3.241062e-03    2
KN883616.1       31041   C       A       0.422659         2.070393e-03    3
KN883616.1       31042   T       G       0.424129         1.269827e-03    3
KN883758.1       179190  A       T       0.336645         3.103740e-03    3
```

Lambda-bcftools

1. Convert sam to bam

```
(/courses/BIOL3411.202430/shared/bcftools_env) [ali.may@c0279 lambda]$ samtools sort eg2.bam -o eg2.sorted.bam
```

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

2.
```
(/courses/BIOL3411.202430/shared/bcftools_env) [ali.may@c0279 lambda]$ source activate /courses/BIOL3411.202430/shared/bcftools_env
```

3. Call variants

```
(/courses/BIOL3411.202430/shared/bcftools_env) [ali.may@c0279 lambda]$ bcftools mpileup -f /courses/BIOL3411.202430/students/ali.may/Assignments/U2B/lambda/lambda.fasta eg2.sorted.bam | bcftools view -Ov - > eg2.raw.bcf
[mpileup] 1 samples in 1 input files
```

4. View results

```
[mpileup] 1 samples in 1 input files
(/courses/BIOL3411.202430/shared/bcftools_env) [ali.may@c0279 lambda]$ bcftools view eg2.raw.bcf
```

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##bcftoolsVersion=1.6+htslib-1.6
##bcftoolsCommand=mpileup -f /courses/BIOL3411.202430/students/ali.may/Assignments/U2B/lambda/lambda.fasta eg2.sorted.bam
##reference=file:///courses/BIOL3411.202430/students/ali.may/Assignments/U2B/lambda/lambda.fasta
##contig=<ID=NC_001416.1,length=48502>
##ALT=<ID=*,Description="Represents allele(s) other than observed.">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of reads supporting an indel">
##INFO=<ID=IMF,Number=1,Type=Float,Description="Maximum fraction of reads supporting an indel">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias for filtering splice-site artefacts in RNA-seq data (bigger is better)",Version="3">
##INFO=<ID=RPB,Number=1,Type=Float,Description="Mann-Whitney U test of Read Position Bias (bigger is better)">
##INFO=<ID=MQB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality Bias (bigger is better)">
##INFO=<ID=BQB,Number=1,Type=Float,Description="Mann-Whitney U test of Base Quality Bias (bigger is better)">
##INFO=<ID=MQSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)">
##INFO=<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">
##INFO=<ID=MDBF,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">
##INFO=<ID=I16,Number=16,Type=Float,Description="Auxiliary tag used for calling, see description of bcf.calcrti_t in bam2bcf.h">
##INFO=<ID=QS,Number=R,Type=Float,Description="Auxiliary tag used for calling">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##bcftoolsViewVersion=1.6+htslib-1.6
##bcftoolsViewCommand=view -Ov -; Date=Wed Feb 21 17:32:29 2024
##bcftoolsViewCommand=view eg2.raw.bcf; Date=Wed Feb 21 17:32:41 2024
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  eg2.sorted.bam
NC_001416.1     1       .       G       <X>     0       .       DP=1;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQBF=0  PL      0,0,0
NC_001416.1     2       .       G       <X>     0       .       DP=2;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQBF=0  PL      0,0,0
NC_001416.1     2       .       GGCG    GGCGCGGGGGCG    0       .       INDEL;IDV=1;IMF=0.5;DP=2;I16=1,0,1,0,0,0,41,1681,24,576,42,1764,1,1,0,0;QS=0.8888889,0.911111;VDB=0.02;SGB=-0.379885;MQBF=0     PL      36,1,0
NC_001416.1     3       .       G       <X>     0       .       DP=2;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQBF=0  PL      0,0,0
NC_001416.1     4       .       C       <X>     0       .       DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQBF=0  PL      0,0,0
NC_001416.1     5       .       G       <X>     0       .       DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQBF=0  PL      0,0,0
NC_001416.1     6       .       C       <X>     0       .       DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQBF=0  PL      0,0,0
NC_001416.1     7       .       C       <X>     0       .       DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQBF=0  PL      0,0,0
NC_001416.1     8       .       G       <X>     0       .       DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQBF=0  PL      0,0,0
NC_001416.1     9       .       A       <X>     0       .       DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQBF=0  PL      0,0,0
NC_001416.1     10      .       C       <X>     0       .       DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQBF=0  PL      0,0,0
NC_001416.1     11      .       C       <X>     0       .       DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQBF=0  PL      0,0,0
NC_001416.1     12      .       T       <X>     0       .       DP=3;I16=1,0,0,0,31,961,0,0,42,1764,0,0,60,0,0;QS=1,0;MQBF=0  PL      0,3,31
NC_001416.1     13      .       C       <X>     0       .       DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQBF=0  PL      0,0,0
NC_001416.1     14      .       G       <X>     0       .       DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQBF=0  PL      0,0,0
NC_001416.1     15      .       C       <X>     0       .       DP=3;I16=1,0,0,0,36,1296,0,0,42,1764,0,0,11,121,0,0;QS=1,0;MQBF=0       PL      0,3,36
NC_001416.1     16      .       C       <X>     0       .       DP=4;I16=1,0,0,0,17,289,0,0,42,1764,0,0,12,144,0,0;QS=1,0;MQBF=0       PL      0,3,17
NC_001416.1     17      .       G       <X>     0       .       DP=4;I16=1,0,0,0,15,361,0,0,42,1764,0,0,13,169,0,0;QS=1,0;MQBF=0       PL      0,3,15
NC_001416.1     18      .       G       <X>     0       .       DP=4;I16=1,0,0,0,15,225,0,0,42,1764,0,0,14,196,0,0;QS=1,0;MQBF=0       PL      0,3,15
NC_001416.1     19      .       T       <X>     0       .       DP=3;I16=1,0,0,0,29,841,0,0,42,1764,0,0,0,0,0,0;QS=1,0;MQBF=0   PL      0,3,29
NC_001416.1     20      .       T       <X>     0       .       DP=5;I16=2,0,0,0,45,1125,0,0,84,3528,0,0,17,257,0,0;QS=1,0;MQBF=0       PL      0,6,42
```