Deliverable 2D

Perform this first set of operations for all three datasets.

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

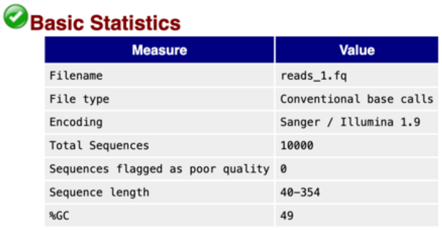
First let’s describe the data we have. For each set of fastq files, describe:

1. How many reads are in each file
    1. Lambda
        1. Reads 1.fq – 16874 paired
        2. Reads 2.fq – 16794 paired
        3. Longreads.fq 10458 single
    2. Ppar
        1. SRR6805880.tiny.fastq – 1000
        2. SRR6805881.tiny.fastq – 1248
        3. SRR6805882.tiny.fastq – 1104
        4. SRR6805883.tiny.fastq – 1134
        5. SRR6805884.tiny.fastq – 1173
        6. SRR6805885.tiny.fastq – 1258
    3. Day
        1. 10_S1_L001_R1_001.fastq – 28893152
        2. 10_S1_L001_R2_001.fastq – 28893152
        3. 11_S1_L001_R1_001.fastq – 29291552
        4. 11_S1_L001_R2_001.fastq – 29291552
        5. 12_S1_L001_R1_001.fastq – 29043844
        6. 12_S1_L001_R2_001.fastq – 29043844
        7. 13_S1_L001_R1_001.fastq – 29023016
        8. 13_S1_L001_R2_001.fastq – 29023016
        9. 14_S1_L001_R1_001.fastq – 24730770
        10. 14_S1_L001_R2_001.fastq – 24730770
        11. 15_S1_L001_R1_001.fastq – 28387419
        12. 15_S1_L001_R2_001.fastq – 28387419
        13. 1_S1_L001_R1_001.fastq – 32833451
        14. 1_S1_L001_R2_001.fastq – 32833451
        15. 2_S1_L001_R1_001.fastq – 33738336
        16. 2_S1_L001_R2_001.fastq – 33738336
        17. 3_S1_L001_R1_001.fastq – 35731214
        18. 3_S1_L001_R2_001.fastq – 35731214
        19. 4_S1_L001_R1_001.fastq – 36678316
        20. 4_S1_L001_R2_001.fastq – 36678316
        21. 5_S1_L001_R1_001.fastq – 36972680
        22. 5_S1_L001_R2_001.fastq – 36972680
        23. 6_S1_L001_R1_001.fastq – 31401357

        24. 6_S1_L001_R2_001.fastq – 31401357
        25. 7_S1_L001_R1_001.fastq – 35536673
        26. 7_S1_L001_R2_001.fastq – 35536673
        27. 8_S1_L001_R1_001.fastq – 24498096
        28. 8_S1_L001_R2_001.fastq – 24498096
        29. 9_S1_L001_R1_001.fastq – 29794050
        30. 9_S1_L001_R2_001.fastq – 29794050
2. The length of the reads and if they are single or paired-end
    1. Lambda
        1. Paired; 40-354
    2. Ppar
        1. Single; 80
    3. Day
        1. Paired; very varied
3. The overall quality of the reads and anything to be concerned about
    1. Lambda and ppar both were not the best quality (using head command) and for both there were a lot of random characters and not just letters. For the Day data, there were a lot of letters (way more than random characters) which means it was of higher quality than the other two.
4. Whether they appear to have adapter sequences that need to be trimmed
    1. Lambda and Day did not
    2. Ppar – TGCAG adapter sequence

Collect quality control data on the reads, in the form of an .html file produced by fastqc.

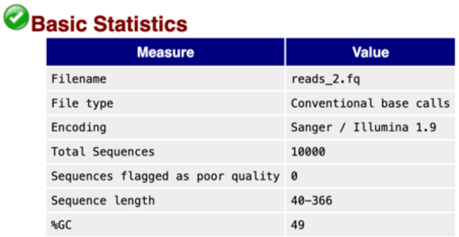
1. Lambda
    1. Reads 1.fq – 16874 paired



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 |

        1.
    2. Reads 2.fq – 16794 paired

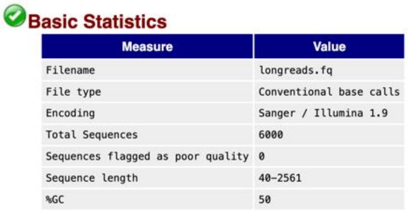


Basic Statistics

| 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 |

        1.

    3. Longreads.fq 10458 single

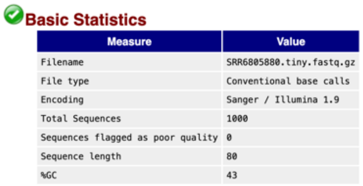


Basic Statistics

| 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 |

        1.

2. Ppar
    1. SRR6805880.tiny.fastq – 1000

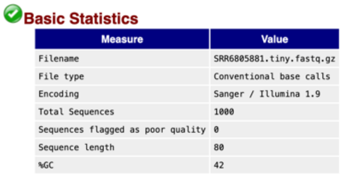


Basic Statistics

| 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 |

        1.

    2. SRR6805881.tiny.fastq – 1248

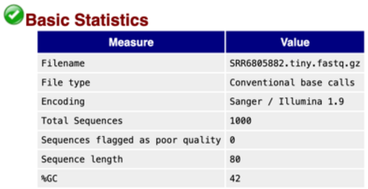


Basic Statistics

| 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 |

        1.

    3. SRR6805882.tiny.fastq – 1104

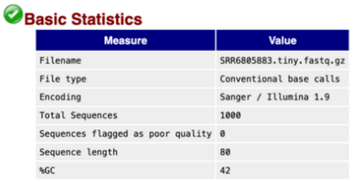


Basic Statistics

| 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 |

        1.

    4. SRR6805883.tiny.fastq – 1134

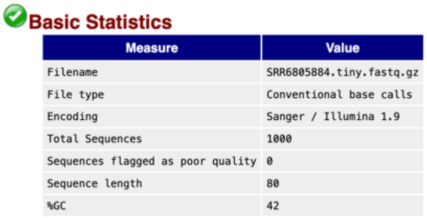


Basic Statistics

| 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 |

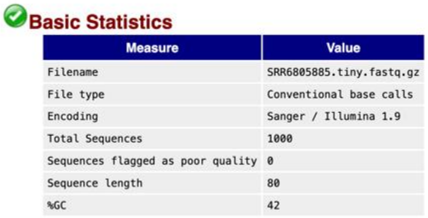
        1.

    5. SRR6805884.tiny.fastq – 1173

1. 

| 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 |

6. SRR6805885.tiny.fastq – 1258

1. 

| 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 |

**If the sequences of a project need trimming, perform this step as described in the Marine Genomics tutorial, using cutadapt.**

Ppar → wrote a shell script

```
For filename on *.tiny.fastq.gz

Do

Base=$(basenamme $filename .tiny.fastq.gz)

Echo ${base }

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

Done
```

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

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

```
(/courses/BIOL3411.202430/shared/cutadapt_env) [vyas.aas@c0184 week4]$ bowtie2-build ppar_tinygenome.fna.gz ppar_tinygenome
```

```
[vyas.aas@login-00 lambda]$ bowtie2-build lambda.fasta lambda
```

**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?)**

***Ppar →***

```
For filename on *.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 →***

```
[vyas.aas@c0184 lambda]$ bowtie2 -x
/home/vyas.aas/2B_deliverable/lambda/bowtie2/example/index/lambda_virus -1
/home/vyas.aas/2B_deliverable/lambda/bowtie2/example/reads/reads_1.fq -2
/home/vyas.aas/2B_deliverable/bowtie2/example/reads/reads_2.fq -S eg2.sam

10000 reads; of these:

  10000 (100.00%) were paired; of these:

    834 (8.34%) aligned concordantly 0 times

    9166 (91.66%) aligned concordantly exactly 1 time

    0 (0.00%) aligned concordantly >1 times

    ----

    834 pairs aligned concordantly 0 times; of these:

      42 (5.04%) aligned discordantly 1 time

    ----

    792 pairs aligned 0 times concordantly or discordantly; of these:

      1584 mates make up the pairs; of these:

        1005 (63.45%) aligned 0 times

        579 (36.55%) aligned exactly 1 time

        0 (0.00%) aligned >1 times

94.97% overall alignment rate
```

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

**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.**

***Ppar →***

```
For filename on *.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 →***

```
[vyas.aas@c0184 lambda]$ module load samtools/1.9

[vyas.aas@c0184 lambda]$ samtools view -bS eg2.sam > eg2.bam

[vyas.aas@c0184 lambda]$ samtools sort eg2.bam -o eg2.sorted.bam

[vyas.aas@c0184 lambda]$ ls

eg1.sam  eg2.bam  eg2.raw.bcf  eg2.sam  eg2.sorted.bam  eg3.sam  lambda_virus.fa
lamda_virus.1.bt2  lamda_virus.2.bt2  lamda_virus.3.bt2  lamda_virus.4.bt2  lamda_virus.rev.1.bt2
lamda_virus.rev.2.bt2
```

angsd

***ppar →***

```
For filename on *.sam

Do

Base=$(basename $filename .sam)
```

```
Echo ${base}

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

Done
```

Then we activate it using source activation from the shared folder.

```
[vyas.aas@c0184 week4]$ /courses/BIOL3411.202430/shared/angsd_env/angsd/angsd -bam
bam.filelist -GL 1 -out genotype_likelihoods -doMaf 2 -SNP_pval 1e-2 -doMajorMinor 1
```

And then →

```
(/courses/BIOL3411.202430/shared/cutadapt_env) [vyas.aas@c0184 week4] gunzip
genotype_likelihoods.mafs.gz

(/courses/BIOL3411.202430/shared/cutadapt_env) [vyas.aas@c0184 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        179190A      T      0.336645         3.103740e-03      3
```

***Lambda →***

Converted sam to bam again

```
(/courses/BIOL3411.202430/shared/bcftools_env) [vyas.aas@c0184 lambda]$ bcftools mpileup -f /courses/BIOL3411.202430/students/vyas.aas/2B_deliverable/lambda/bowtie2/example/reference/lambda_virus.fa eg2.sorted.bam | bcftools view -Ov - > eg2.raw.bcf
```

```
[mpileup] 1 samples in 1 input files
```

```
(/courses/BIOL3411.202430/shared/bcftools_env)[vyas.aas@c0184 lambda]$ bcftools view eg2.raw.bcf
```

HUGE FILE OUTPUT - did not include all data

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##bcftoolsVersion=1.6+htslib-1.6
##bcftoolsCommand=mpileup -f /courses/BIOL3411.202430/students/vyas.aas/2B_deliverable/lambda/bowtie2/example/reference/lambda_virus.fa eg2.sorted.bam
##reference=file:///courses/BIOL3411.202430/students/vyas.aas/2B_deliverable/lambda/bowtie2/example/reference/lambda_virus.fa
##contig=<ID=gi|9626243|ref|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=MQ0F,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_callret1_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">
##bcftools_viewVersion=1.6+htslib-1.6
##bcftools_viewCommand=view -Ov -; Date=Thu Mar 14 17:22:29 2024
##bcftools_viewCommand=view eg2.raw.bcf; Date=Thu Mar 14 17:22:41 2024
```

```
#CHROM          POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT          eg2.sorted.bam
gi|9626243|ref|NC_001416.1|     1       .       G       <*>     0       .
        DP=1;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQ0F=0          PL      0,0,0
gi|9626243|ref|NC_001416.1|     2       .       G       <*>     0       .
        DP=2;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQ0F=0          PL      0,0,0
gi|9626243|ref|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=0.0888889,0.91
1111;VDB=0.02;SGB=-0.379885;MQ0F=0              PL      36,1,0
gi|9626243|ref|NC_001416.1|     3       .       G       <*>     0       .
        DP=2;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQ0F=0          PL      0,0,0
gi|9626243|ref|NC_001416.1|     4       .       C       <*>     0       .
        DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQ0F=0          PL      0,0,0
gi|9626243|ref|NC_001416.1|     5       .       G       <*>     0       .
        DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQ0F=0          PL      0,0,0
gi|9626243|ref|NC_001416.1|     6       .       G       <*>     0       .
        DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQ0F=0          PL      0,0,0
gi|9626243|ref|NC_001416.1|     7       .       C       <*>     0       .
        DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQ0F=0          PL      0,0,0
gi|9626243|ref|NC_001416.1|     8       .       G       <*>     0       .
        DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQ0F=0          PL      0,0,0
gi|9626243|ref|NC_001416.1|     9       .       A       <*>     0       .
        DP=3;I16=0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0;QS=0,0;MQ0F=0          PL      0,0,0
```