Genome Analysis
Unit 2 Deliverable B
Aligning Another Set of Reads

For this example we will be working with the lambda phage data obtained from this tutorial https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#getting-started-with-bowtie-2-lambda-phage-example

1. To begin I indexed the files using the bowtie tool (bowtie/2.3.5.1)

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

This generates files ending in .1.bt2, .2.bt2, etc

```
[ali.may@login-00 lambda]$ ls
lambda.1.bt2  lambda.2.bt2  lambda.3.bt2  lambda.4.bt2  lambda.fasta  lambda.rev.1.bt2  lambda.rev.2.bt2  longreads.fq  reads_1.fq  reads_2.fq
```

2. Next I use the bowtie tool to align the reads, this will create an egl.sam file

```
[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
```

3. I then examined the lines of the sam file using the head command

```
[ali.may@login-00 lambda]$ head egl.sam
@HD     VN:1.0  SO:unsorted
@SQ     SN:NC_001416.1  LN:48502
@PG     ID:bowtie2      PN:bowtie2      VN:2.3.5.1      CL:"/shared/centos7/bowtie/2.3.5.1/bin/bowtie
r1      0       NC_001416.1     18401   42      122M    *       0       0       TGAATGCGAACTCCGGGACGC
+"@6<:27(F&5)9)"B:%B+A-%5A?2$HCB0B+0=D<7E/<.03#!.F77@6B==?C"7>;))%;,3-$.A06+<-1/@@?,26">=?*@'0;$:;??G
r2      0       NC_001416.1     8886    42      275M    *       0       0
```

4. Next I aligned the paired end reads using bowtie

```
(/courses/BIOL3411.202430/shared/cutadapt_env) [ali.may@c0279 lambda]$ bowtie2 -x lambda -1 reads_1.fq -2 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
```

5. Next I also aligned the long reads

```
(/courses/BIOL3411.202430/shared/bcftools_env) [ali.may@c0279 lambda]$ bowtie2 --local -x lambda -U longreads.fq -S eg3.sam
6000 reads; of these:
  6000 (100.00%) were unpaired; of these:
    158 (2.63%) aligned 0 times
    5636 (93.93%) aligned exactly 1 time
    206 (3.43%) aligned >1 times
97.37% overall alignment rate
```

6. Now I had sam files that I converted to bam files (compressed version) using samtools

```
(/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
```

7. Once converted to bam I again used samtools to convert the file into a sorted bam file

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

   This format makes the files nice for long term storage and easier for variant discovery

8. To generate variant calls I need to use bcftools. I had to access this tool from a shared environment and activate it. Also need to load Anaconda
   module load anaconda3/2022.05

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

9. 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
```

10. Finally to view the variants I used this command and bcftools

```
(/courses/BIOL3411.202430/shared/bcftools_env) [ali.may@c0279 lambda]$ bcftools view eg2.raw.bcf
```

This is what the command should generate.

```
##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=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=Wed Feb 21 17:32:29 2024
##bcftools_viewCommand=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       <*>     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
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
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=0.0888889,0.911111;VDB=0.02;SGB=-0.379885;MQ0F=0     PL      36,1,0
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
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
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
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
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
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
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
NC_001416.1     10      .       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
NC_001416.1     11      .       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
NC_001416.1     12      .       T       <*>     0       .       DP=3;I16=1,0,0,0,31,961,0,0,42,1764,0,0,8,64,0,0;QS=1,0;MQ0F=0  PL      0,3,31
NC_001416.1     13      .       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
NC_001416.1     14      .       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
NC_001416.1     15      .       C       <*>     0       .       DP=3;I16=1,0,0,0,36,1296,0,0,42,1764,0,0,11,121,0,0;QS=1,0;MQ0F=0       PL      0,3,36
NC_001416.1     16      .       G       <*>     0       .       DP=4;I16=1,0,0,0,17,289,0,0,42,1764,0,0,12,144,0,0;QS=1,0;MQ0F=0        PL      0,3,17
NC_001416.1     17      .       G       <*>     0       .       DP=4;I16=1,0,0,0,19,361,0,0,42,1764,0,0,13,169,0,0;QS=1,0;MQ0F=0        PL      0,3,19
NC_001416.1     18      .       G       <*>     0       .       DP=4;I16=1,0,0,0,15,225,0,0,42,1764,0,0,14,196,0,0;QS=1,0;MQ0F=0        PL      0,3,15
NC_001416.1     19      .       T       <*>     0       .       DP=5;I16=1,0,0,0,29,841,0,0,42,1764,0,0,0,0,0,0;QS=1,0;MQ0F=0   PL      0,3,29
NC_001416.1     20      .       T       <*>     0       .       DP=5;I16=2,0,0,0,45,1125,0,0,84,3528,0,0,17,257,0,0;QS=1,0;MQ0F=0        PL      0,6,42
```