

Apply to questions 1 - 5:

Generate a bowtie2 index of the wu_0_A genome using bowtie2-build, with the prefix 'wu_0'.

Apply to questions 6 - 10:

Run bowtie2 to align the reads to the genome, under two scenarios: first, to report only full-length matches of the reads; and second, to allow partial (local) matches. All other parameters are as set by default.

For the following set of questions (11 - 20), use the set of full-length alignments calculated under scenario 1 only. Convert this SAM file to BAM, then sort the resulting BAM file.

Apply to questions 11 - 15:

Compile candidate sites of variation using SAMtools mpileup for further evaluation with BCFtools. Provide the reference fasta genome and use the option “-uv” to generate the output in uncompressed VCF format for easy examination.

Apply to questions 16 - 20:

Call variants using ‘BCFtools call’ with the multiallelic-caller model. For this, you will need to first re-run SAMtools mpileup with the BCF output format option (‘-g’) to generate the set of candidate sites to be evaluated by BCFtools. In the output to BCFtools, select to show only the variant sites, in uncompressed VCF format for easy examination.

1. How many sequences were in the genome?
2. What was the name of the third sequence in the genome file? Give the name only, without the ">" sign.
3. What was the name of the last sequence in the genome file? Give the name only, without the ">" sign.

```
[guest@centos6]$ cd project3
```

```
[guest@centos6 project3]$ gunzip gencommand_proj3_data.tar.gz
```

```
[guest@centos6 project3]$ tar xvc gencommand_proj3_data.tar
```

```
[guest@centos6 project3]$ ls -tl
```

```
[guest@centos6 project3]$ head wu_0.v7.fas
```

```
[guest@centos6 project3]$ cat wu_0.v7.fas | grep ">" | head
```

```
>Chr1
```

```
>Chr2
```

```
>Chr3
```

```
>Chr4
```

```
>Chr5
```

```
>chloroplast
```

```
>mitochondria
```

```
[guest@centos6 project3]$ cat wu_0.v7.fas | grep ">" | wc -l
```

```
7
```

4. How many index files did the operation create?
5. What is the 3-character extension for the index files created?

```
[guest@centos6 project3]$ mkdir wu_0
```

```
[guest@centos6 project3]$ bowtie2-build wu_0.v7.fas wu_0/wu_0
```

```
[guest@centos6 project3]$ ls wu_0
```

```
[guest@centos6 project3]$ ls wu_0 | wc -l
```

6. How many reads were in the original fastq file?

```
[guest@centos6 project3]$ head wu_0_A_wgs.fastq
```

```
[guest@centos6 project3]$ cat wu_0_A_wgs.fastq | wc -l
```

(note: the reads are counting lines/4)

7. How many matches (alignments) were reported for: i) the original (full-match) setting? and ii) with the local-match setting? Exclude lines in the file containing unmapped reads. Give these two numbers separated by a space (e.g., 23 53).

8. How many reads were mapped, in each scenario? Use the format above.

9. How many reads had multiple matches, in each scenario? Use the format above. You can find this in the bowtie2 summary; note that by default bowtie2 only reports the best match for each read.

```
[guest@centos6 project3]$ bowtie2 -p 4 -x wu_0/wu_0 wu_0_A_wgs.fastq  
-S wu_0.bt2.sam
```

147354 reads; of these:

147354 (100.00%) were unpaired; of these:

9635 (6.54%) aligned 0 times

93780 (63.64%) aligned exactly 1 time

43939 (29.82%) aligned >1 times

93.46% overall alignment rate

```
[guest@centos6 project3]$ bowtie2 -p 4 --local -x wu_0/wu_0  
wu_0_A_wgs.fastq -S wu_0.bt2.local.sam
```

147354 reads; of these:

147354 (100.00%) were unpaired; of these:

6310 (4.28%) aligned 0 times

84939 (57.64%) aligned exactly 1 time

56105 (38.07%) aligned >1 times

95.72% overall alignment rate

```
[guest@centos6 project3]$
```

10. How many alignments contained insertions and/or deletions, in each scenario? Use the format above.

```
[guest@centos6 project3]$ cat wu_0.bt2.sam | cut -f6 | grep "I" | grep "D" |  
wc -l
```

42

```
[guest@centos6 project3]$ cat wu_0.bt2.sam | cut -f6 | grep "I" | wc -l
```

1429

```
[guest@centos6 project3]$ cat wu_0.bt2.sam | cut -f6 | grep "D" | wc -l
```

1395

(note the answer is $1429 + 1395 - 42\ 1223 + 1476 - 85$)

11. How many entries were reported for Chr3?

```
[guest@centos6 project3]$ samtools view -bT wu_0.v7.fas wu_0.bt2.sam >
wu_0.bt2.bam
```

```
[guest@centos6 project3]$ samtools sort wu_0.bt2.bam
wu_0.bt2.sorted.bam
```

```
[guest@centos6 project3]$ samtools index wu_0.bt2.sorted.bam.bam
```

```
[guest@centos6 project3]$ samtools mpileup -f wu_0.v7.fas
wu_0.bt2.sorted.bam.bam > wu_0.mpileup
```

```
[guest@centos6 project3]$ samtools mpileup -uv -f wu_0.v7.fas
wu_0.bt2.sorted.bam.bam > wu_0.vcf
```

```
[guest@centos6 project3]$ more wu_0.vcf
```

```
[guest@centos6 project3]$ cat wu_0.vcf | cut -f1 | grep "Chr3" | head
##contig=<ID=Chr3,length=23042017>
```

Chr3

Chr3

Chr3

Chr3

Chr3

Chr3

Chr3

Chr3

Chr3

```
[guest@centos6 project3]$ cat wu_0.vcf | cut -f1 | grep "Chr3" | wc -l  
360296
```

(note: the answer is 360296 - 1 (1 count included in header))

12. How many entries have 'A' as the corresponding genome letter?

```
[guest@centos6 project3]$ cat wu_0.vcf | cut -f4 | more
```

```
[guest@centos6 project3]$ cat wu_0.vcf | cut -f4 | awk '$1 == "A"' | head
```

A

A

A

A

A

A

A

A

A

A

```
[guest@centos6 project3]$ cat wu_0.vcf | cut -f4 | awk '$1 == "A"' | wc -l  
1150985
```

13. How many entries have exactly 20 supporting reads (read depth)?

```
[guest@centos6 project3]$ cat wu_0.vcf | cut -f8 | grep "DP=20" | head
```

```
DP=20;I16=20,0,0,0,676,22856,0,0,18,18,0,0,231,3795,0,0;QS=1,0;MQ0F=0.1
```

```
DP=20;I16=20,0,0,0,665,22325,0,0,18,18,0,0,238,3656,0,0;QS=1,0;MQ0F=0.1
```

```
DP=20;I16=20,0,0,0,669,22475,0,0,18,18,0,0,244,3578,0,0;QS=1,0;MQ0F=0.1
```

```
DP=20;I16=20,0,0,0,678,22986,0,0,18,18,0,0,246,3598,0,0;QS=1,0;MQ0F=0.1
```

```
DP=20;I16=20,0,0,0,666,22252,0,0,18,18,0,0,248,3658,0,0;QS=1,0;MQ0F=0.1
```

```
DP=20;I16=20,0,0,0,662,21998,0,0,18,18,0,0,250,3758,0,0;QS=1,0;MQ0F=0.1
```

```
DP=20;I16=20,0,0,0,665,22195,0,0,18,18,0,0,252,3898,0,0;QS=1,0;MQ0F=0.1
```

```
DP=20;I16=17,0,3,0,574,19396,91,2821,15,15,3,3,232,3804,20,174;QS=0.85,0.15,0;VDB=0.17915;SGB=-0.511536;RPB=0.162817;MQB=0.95083;BQB=0.260295;MQ0F=0.1
```

```
DP=20;I16=20,0,0,0,654,21558,0,0,840,35280,0,0,245,4393,0,0;QS=1,0;MQ0F=0
```

```
DP=20;I16=20,0,0,0,669,22415,0,0,840,35280,0,0,234,3688,0,0;QS=1,0;M
```

Q0F=0

```
[guest@centos6 project3]$ cat wu_0.vcf | cut -f8 | grep "DP=20" | wc -l  
1816
```

14. How many entries represent indels?

```
[guest@centos6 project3]$ cat wu_0.vcf | cut -f8 | grep "INDEL" | head  
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the  
variant is an INDEL.">
```

```
INDEL;IDV=1;IMF=1;DP=1;I16=0,0,1,0,0,0,26,676,0,0,40,1600,0,0,24,576;  
QS=0,1;SGB=-0.379885;MQ0F=0
```

```
INDEL;IDV=1;IMF=1;DP=1;I16=0,0,1,0,0,0,164,26896,0,0,0,0,0,0,23,529;Q  
S=0,1;SGB=-0.379885;MQ0F=1
```

```
INDEL;IDV=1;IMF=1;DP=1;I16=0,0,1,0,0,0,35,1225,0,0,40,1600,0,0,22,484  
;QS=0,1;SGB=-0.379885;MQ0F=0
```

```
INDEL;IDV=1;IMF=1;DP=1;I16=0,0,1,0,0,0,154,23716,0,0,0,0,0,0,22,484;Q  
S=0,1;SGB=-0.379885;MQ0F=1
```

```
INDEL;IDV=1;IMF=1;DP=1;I16=0,0,1,0,0,0,34,1156,0,0,23,529,0,0,19,361;  
QS=0,1;SGB=-0.379885;MQ0F=0
```

```
INDEL;IDV=1;IMF=1;DP=1;I16=0,0,1,0,0,0,86,7396,0,0,0,0,0,0,20,400;QS=  
0,1;SGB=-0.379885;MQ0F=1
```

```
INDEL;IDV=1;IMF=1;DP=1;I16=0,0,1,0,0,0,31,961,0,0,23,529,0,0,18,324;Q  
S=0,1;SGB=-0.379885;MQ0F=0
```

```
INDEL;IDV=1;IMF=1;DP=1;I16=0,0,1,0,0,0,34,1156,0,0,40,1600,0,0,23,529
```



```
;QS=0,1;SGB=-0.379885;MQ0F=0
```

```
INDEL;IDV=1;IMF=1;DP=1;I16=0,0,1,0,0,0,29,841,0,0,23,529,0,0,10,100;Q
```

```
S=0,1;SGB=-0.379885;MQ0F=0
```

```
[guest@centos6 project3]$ cat wu_0.vcf | cut -f8 | grep "INDEL" | wc -l  
1973
```

(note: aws = 1973 -1 (1 count in header))

15. How many entries are reported for position 175672 on Chr1?

```
[guest@centos6 project3]$ cat wu_0.vcf | cut -f1,2 | grep "Chr1" | awk '$2  
== "175672"' | head
```

```
Chr1 175672
```

```
Chr1 175672
```

```
[guest@centos6 project3]$ cat wu_0.vcf | cut -f1,2 | grep "Chr1" | awk '$2  
== "175672"' | wc -l
```

```
2
```

16. How many variants are called on Chr3?

```
[guest@centos6 project3]$ samtools mpileup -g -f wu_0.v7.fas  
wu_0.bt2.sorted.bam.bam > wu_0.bcf
```

```
[guest@centos6 project3]$ bcftools call -v -m -O z -o wu_0.vcf.gz  
wu_0.bcf
```

```
[guest@centos6 project3]$ zcat wu_0.vcf.gz | cut -f1 | grep "Chr3" | head  
##contig=<ID=Chr3,length=23042017>
```

Chr3

Chr3

Chr3

Chr3

Chr3

Chr3

Chr3

Chr3

Chr3

```
[guest@centos6 project3]$ zcat wu_0.vcf.gz | cut -f1 | grep "Chr3" | wc -l
```

399

(note: aws = 399 -1 (1 count in header))

17. How many variants represent an A->T SNP? If useful, you can use 'grep -P' to allow tabular spaces in the search term.

```
[guest@centos6 project3]$ zcat wu_0.vcf.gz | cut -f4,5 | awk '$1 == "A" &&
```

```
$2 == "T"' | head
```

A T

A T

A T

A T

A T

A T

A T

A T

A T

A T

```
[guest@centos6 project3]$ zcat wu_0.vcf.gz | cut -f4,5 | awk '$1 == "A" &&
```

```
$2 == "T" | wc -l
```

392

18. How many entries are indels?

```
[guest@centos6 project3]$ zcat wu_0.vcf.gz | cut -f8 | grep "INDEL" | head
```

```
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the  
variant is an INDEL.">
```

```
INDEL;IDV=1;IMF=1;DP=1;SGB=-0.379885;MQ0F=0;ICB=1;HOB=0.5;AC=  
1;AN=2;DP4=0,0,1,0;MQ=40
```

```
INDEL;IDV=1;IMF=1;DP=1;SGB=-0.379885;MQ0F=0;ICB=1;HOB=0.5;AC=  
1;AN=2;DP4=0,0,1,0;MQ=40
```

```
INDEL;IDV=1;IMF=1;DP=1;SGB=-0.379885;MQ0F=0;ICB=1;HOB=0.5;AC=  
1;AN=2;DP4=0,0,1,0;MQ=40
```

```
INDEL;IDV=1;IMF=1;DP=1;SGB=-0.379885;MQ0F=0;ICB=1;HOB=0.5;AC=  
1;AN=2;DP4=0,0,1,0;MQ=40
```

```
INDEL;IDV=1;IMF=1;DP=1;SGB=-0.379885;MQ0F=0;ICB=1;HOB=0.5;AC=  
1;AN=2;DP4=0,0,1,0;MQ=40
```

```
INDEL;IDV=1;IMF=1;DP=1;SGB=-0.379885;MQ0F=0;ICB=1;HOB=0.5;AC=
```

```
1;AN=2;DP4=0,0,1,0;MQ=40
```

```
INDEL;IDV=1;IMF=1;DP=1;SGB=-0.379885;MQ0F=0;ICB=1;HOB=0.5;AC=
```

```
1;AN=2;DP4=0,0,1,0;MQ=40
```

```
INDEL;IDV=1;IMF=1;DP=1;SGB=-0.379885;MQ0F=0;ICB=1;HOB=0.5;AC=
```

```
1;AN=2;DP4=0,0,1,0;MQ=40
```

```
INDEL;IDV=1;IMF=1;DP=1;SGB=-0.379885;MQ0F=0;ICB=1;HOB=0.5;AC=
```

```
1;AN=2;DP4=0,0,1,0;MQ=40
```

```
[guest@centos6 project3]$ zcat wu_0.vcf.gz | cut -f8 | grep "INDEL" | wc -l
```

```
321
```

(note: aws = 321 -1 (remove 1 count in header))

19. How many entries have precisely 20 supporting reads (read depth)?

```
[guest@centos6 project3]$ zcat wu_0.vcf.gz | cut -f8 | grep "DP=20" | head
```

```
DP=20;VDB=0.0587288;SGB=-0.556411;RPB=0.639909;MQB=0.931063;
```

```
BQB=0.972484;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=16,0,4,0;MQ=
```

```
39
```

```
DP=20;VDB=0.255089;SGB=-0.676189;RPB=0.97436;MQB=0.499893;BQ
```

```
B=0.850154;MQ0F=0.05;ICB=1;HOB=0.5;AC=1;AN=2;DP4=9,0,11,0;MQ=
```

```
13
```

```
[guest@centos6 project3]$ zcat wu_0.vcf.gz | cut -f8 | grep "DP=20" | wc -l
```

```
2
```

20. What type of variant (i.e., SNP or INDEL) is called at position 11937923

on Chr3?

```
[guest@centos6 project3]$ zcat wu_0.vcf.gz | grep "Chr3" | awk '$2 ==  
"11937923"'
```

```
Chr3 11937923 . G A 13.73.
```

```
DP=20;VDB=0.0587288;SGB=-0.556411;RPB=0.639909;MQB=0.931063;
```

```
BQB=0.972484;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=16,0,4,0;MQ=
```

```
39 GT:PL 0/1:48,0,137
```