Introduction:

- We've sequenced one strain wu_0_A to determine genetic variants
- Sequencing reads in are in the wu_0_A_wgs.fastq
- Develop a variant calling pipeline

Questions 1-5:

- build a genome index from wu_0_A

```
mv wu_0.v7.fas wu_0.v7.fasta
mkdir wu_0
bowtie2-build wu_0.v7.fasta wu_0/wu_0 # generated 6 files in wu_0\
```

1. How many sequences were in the genome? (Asking for the sequences in the fasta file)

```
more wu_0.v7.fasta | grep -c ">" # 7 chromosomes
```

2. Name of the third sequence.
3. Name of last sequence

```
more wu_0.v7.fasta | grep ">"
```

4. How many index files did the operation create? 6.
5. What is the 3-character extension for the index files created? .bt2

Questions 6-14: Use bowtie2 to align reads to the genome two ways

- report only full length matches of the reads
- allow local matches

```
bowtie2 -p 4 -x wu_0/wu_0 -U wu_0_A_wgs.fastq -S w0_0.sam # created
sam file, got a summary.

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

bowtie2 --local -p 4 -x wu_0/wu_0 -U wu_0_A_wgs.fastq -S
w0_0.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
```

6. How many reads in the original FASTQ files? 147354 reads.
7. How many matches (alignments) were reported for the original (full-match) setting? Exclude lines in the file containing unmapped reads. 137719
8. How many matches (alignments) were reported with the local-match setting? Exclude lines in the file containing unmapped reads. 141044
9. How many reads were mapped in the scenario in Question 7?
10. repeat 8
11. How many reads had multiple matches in the scenario in Question 7? You can find this in the bowtie2 summary; note that by default bowtie2 only reports the best match for each read. look at summary
12. How many reads had multiple matches in the scenario in Question 8? 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. look at summary.
13. How many alignments contained insertions and/or deletions, in the scenario in Question 7?

    ```
    samtools view w0_0.sam | cut -f6 | grep -E -c "I|D"
    ```

    2782
14. How many alignments contained insertions and/or deletions, in the scenario in Question 8?

    ```
    samtools view w0_0.local.sam | cut -f6 | grep -E -c "I|D"
    ```

    2614

For the following set of questions (15 - 24), use the set of full-length alignments calculated under scenario 1 only. Convert this SAM file to BAM, then sort the resulting BAM file.

```
samtools view -b w0_0.sam > w0_0.bam
```

Questions 15-19: compile the sites of variation using Samtools mpileup, use -uv and generate output in uncompressed vcf format.

```
(base) [root@12018efd4a72 project3]# samtools sort w0_0.bam
w0_0.sorted
```

```
(base) [root@12018efd4a72 project3]# samtools index w0_0.sorted.bam
```

Their approach was different

```
% samtools view –bT wu_0.v7.fas out.full.sam > out.full.bam
```

then sorting it:

```
% samtools sort out.full.bam out.full.sorted
```

```
samtools mpileup -v -u -f  wu_0.v7.fasta w0_0.sorted.bam > w0_0.vcf
```

15. How many entries were reported for Chr3?

```
(base) [root@12018efd4a72 project3]# more w0_0.vcf | grep -c "Chr3"
```

```
360296
```

- not sure exactly what went wrong here

```
(base) [root@12018efd4a72 project3]# more w0_0.vcf | grep -v "^#" |
cut -f1 | grep -c "^Chr3"
```

```
398
```

- pretty sure this was correct answer, may be because Chr is mentioned more than just in the column

16. How many entries have ‘A’ as the corresponding genome letter?

```
(base) [root@12018efd4a72 project3]# more w0_0.vcf | grep -v "##" |
cut -f4 | grep -P "^A$" | wc -l # ^A$ starts and ends with A
```

1150985

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

(base) [root@12018efd4a72 project3]# more w0_0.vcf | grep -c "DP=20"

1816

18. how many indels?

(base) [root@12018efd4a72 project3]# more w0_0.vcf | grep -v "##" | grep -c "INDEL"

1972

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

(base) [root@12018efd4a72 project3]# more w0_0.vcf | grep -v "##" | cut -f1-2 | grep "Chr1" | cut -f2 | grep -c "175672"

24

Wrong, % cat out.full.mpileup.vcf | grep –v “^#” | cut –f1,2 | grep Chr1 | grep 175672 was right.

Question 20-24: call variants with bcftools call

Rerurn samtools mpileup with bcf format -g

```
samtools mpileup -g -u -f  wu_0.v7.fasta w0_0.sorted.bam > w0_0.bcf
```

Run bcftools: show only variant sites, uncompressed vcf format

```
bcftools call -v -m -O v -o w0_0.vcf w0_0.bcf
```

20. How many variants are called on Chr3?

```
(base) [root@12018efd4a72 project3]# more w0_0.vcf | grep -v "##" |
grep -c "Chr3"
398
```

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

```
(base) [root@12018efd4a72 project3]# more w0_0.vcf | grep -v "##" |
cut -f4-5 | grep -P "^A\tT$" | wc -l
392
```

22. How many indels?

```
^[OB(base) [root@12018efd4a72 project3]# more w0_0.vcf | grep -v "##"
| grep -c "INDEL"
320
```

23. depth reads 20

```
(base) [root@12018efd4a72 project3]# more w0_0.vcf | grep -v "##" |
grep -c "DP=20"
2
```