Project 3

For this project, it is recommended that you use the VMBox virtual environment (***Docker instead for MacOS with M1 or M2 chips) provided with the Course package and the tools therein. You may also use your own system and software, however make sure that appropriate versions are installed. The answers are compatible with the following versions of the software: samtools v.1.2, bowtie v.2.2.5 and bcftools v.1.2.
As part of the effort to catalog genetic variation in the plant Arabidopsis thaliana, you re-sequenced the genome of one strain ('wu_0_A'; genome file: 'wu_0.v7.fas'), to determine genetic variants in this organism. The sequencing reads produced are in the file 'wu_0_A_wgs.fastq'. Using the tools bowtie2, samtools and bcftools, develop a pipeline for variant calling in this genome. NOTE: Genome and re-sequencing data have been obtained and modified from those generated by the 1001 Genomes Project, accession 'Wu_0_A'.

Apply to questions 1 - 5:
Generate a bowtie2 index of the wu_0_A genome using bowtie2-build, with the prefix 'wu_0'.

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# bowtie2-build wu_0.v7.fas wu_0
```

1. How many sequences were in the genome?

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# grep -c "^>" wu_0.v7.fas
```

   a. 7
2. What was the name of the third sequence in the gnome file? Give the name only, without the ">" sign.

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# grep "^>" wu_0.v7.fas | head -3 | tail -1
```

   a. >Chr3
3. What was the name of the last sequence in the genome file? Give the name only, without ">" sign.

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# grep "^>" wu_0.v7.fas | tail -1
```

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

Apply to questions 6 - 14:
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.

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# bowtie2 -p 4 -x wu_0/wu_0 wu_0_A_wgs.fastq -S wu_0.bt2.
sam
```

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

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# wc -l wu_0_A_wgs.fastq
```

   a. 589416 wu_0_A_wgs.fastq
   b. 589416/4 = 147354

7. How many matches (alignments) were reported for the original (full-match) setting? Exclude lines in the file containing unmapped reads.

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# 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
```

   a.
   b. 137719
8. How many matches (alignments) were reported with the local-match setting? Exclude lines in the file containing unmapped reads.

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# bowtie2 -p 4 --local -x wu_0/wu_0 wu_0_A_wgs.fastq -S w
u_0.loc.bt2.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
```

   a.
   b. 141044
9. How many reads were mapped in the scenario in Question 7?
   a. 137719
10. How many reads were mapped in the scenario in Question 8?
   a. 141044
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.
   a. 43939
12. How many reads had multiple matches in the scenario in Question 8? You can find this in the bowtie2 summary; note that by default bowtie2 only reports the best match for each read.
   a. 56105
13. How many alignments contained insertions and/or deletions, in the scenario in Question 7?

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# cat wu_0.bt2.sam | cut -f6 | grep "I" | grep "D" | wc -
l
42
(base) [root@522cc4c0d20e gencommand_proj3_data]# cat wu_0.bt2.sam | cut -f6 | grep "I" | wc -l
1429
(base) [root@522cc4c0d20e gencommand_proj3_data]# cat wu_0.bt2.sam | cut -f6 | grep "D" | wc -l
1395
```

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

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# cat wu_0.loc.bt2.sam | cut -f6 | grep "I" | grep "D" |
wc -l
85
(base) [root@522cc4c0d20e gencommand_proj3_data]# cat wu_0.loc.bt2.sam | cut -f6 | grep "I" | wc -l
1223
(base) [root@522cc4c0d20e gencommand_proj3_data]# cat wu_0.loc.bt2.sam | cut -f6 | grep "D" | wc -l
1475
```

   a.
   b. 2613

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.
Apply to questions 15 - 19:

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# samtools mpileup -uv -f wu_0.v7.fas wu_0.sorted.bt2.bam
.bam > wu_0.vcf
[mpileup] 1 samples in 1 input files
<mpileup> Set max per-file depth to 8000
```

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.

15. How many entries were reported for Chr3?

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# cat wu_0.vcf | cut -f1 | grep "Chr3" | wc -l
360296
```

    a.
    b. 360295

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

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# cat wu_0.vcf | grep -v "^#" | cut -f4 | grep -P "^A$" |
wc -l
1150985
```

    a.

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

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# cat wu_0.vcf | grep -v "^#" | grep -c "DP=20;"
1816
```

    a.

18. How many entries represent indels?

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# cat wu_0.vcf | grep -v "^#" | grep -c "INDEL"
1972
```

    a.

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

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# cat wu_0.vcf | grep -v "^#" | cut -f1,2 | grep "Chr1" |
 grep -w "175672" | wc -l
2
```

    a.

Apply to questions 20 - 24:
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.

20. How many variants are called on Chr3

```
(base) [root@522cc4c0d20e gencommand_proj3_data]# zcat wu_0.vcf.gz |cut -f1 |grep "Chr3" | wc -l
399
```

    a.
    b. 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@522cc4c0d20e gencommand_proj3_data]# zcat wu_0.vcf.gz | grep -v "^#"|cut -f4,5|grep -P "^A\t
T$" | wc -l
392
```

    a.

22. How many entries are indels?
    a.
```
(base) [root@522cc4c0d20e gencommand_proj3_data]# zcat wu_0.vcf.gz | grep -v "^#"|grep -c "INDEL"
320
```
23. How many entries have precisely 20 supporting reads (read depth)?
    a.
```
(base) [root@522cc4c0d20e gencommand_proj3_data]# zcat wu_0.vcf.gz | grep -v "^#"|grep -c "DP=20;"
2
```
24. What type of variant (i.e., SNP or INDEL) is called at position 11937923 on Chr3
    a.
```
2
(base) [root@522cc4c0d20e gencommand_proj3_data]# zcat wu_0.vcf.gz | grep -v "^#"|cut -f1-5| grep -w "119
37923"
Chr3    11937923        .       G       A
```