

## Feedback — Module 3 Exam

[Help Center](#)

Thank you. Your submission for this exam was received.

You submitted this exam on **Thu 24 Sep 2015 10:04 PM PDT**. You got a score of **20.00** out of **20.00**.

For this project, it is recommended that you use the VMBox virtual environment 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'.

[Click here to download the Project 3 Data Files](#)

### 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.

## Question 1

How many sequences were in the genome?

**You entered:**

Your Answer		Score	Explanation
7	✓	1.00	
Total		1.00 / 1.00	

## Question 2

What was the name of the third sequence in the genome file? Give the name only, without the ">" sign.

**You entered:**

Your Answer		Score	Explanation
Chr3	✓	1.00	
Total		1.00 / 1.00	

## Question 3

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

**You entered:**

Your Answer		Score	Explanation
mitochondria	✓	1.00	
Total		1.00 / 1.00	

## Question 4

How many index files did the operation create?

**You entered:**

Your Answer		Score	Explanation
6	✓	1.00	
Total		1.00 / 1.00	

## Question 5

What is the 3-character extension for the index files created?

**You entered:**

Your Answer		Score	Explanation
bt2	✓	1.00	
Total		1.00 / 1.00	

## Question 6

How many reads were in the original fastq file?

You entered:

147354

Your Answer		Score	Explanation
147354	✓	1.00	
Total		1.00 / 1.00	

## Question 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).

You entered:

137719 141044

Your Answer		Score	Explanation
137719	✓	0.50	
141044	✓	0.50	

Total

1.00 / 1.00

## Question 8

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

**You entered:**

Your Answer		Score	Explanation
137719	✓	0.50	
141044	✓	0.50	
Total		1.00 / 1.00	

## Question 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.

**You entered:**

Your Answer		Score	Explanation
43939	✓	0.50	
56105	✓	0.50	
Total		1.00 / 1.00	

## Question 10

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

**You entered:**

2782 2614

Your Answer		Score	Explanation
2782	✓	0.50	
2614	✓	0.50	
Total		1.00 / 1.00	

## Question 11

How many entries were reported for Chr3?

**You entered:**

360295

Your Answer		Score	Explanation
360295	✓	1.00	
Total		1.00 / 1.00	

## Question 12

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

**You entered:**

Your Answer		Score	Explanation
1150985	✓	1.00	
Total		1.00 / 1.00	

## Question 13

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

You entered:

Your Answer		Score	Explanation
1816	✓	1.00	
Total		1.00 / 1.00	

## Question 14

How many entries represent indels?

You entered:

Your Answer		Score	Explanation
1972	✓	1.00	
Total		1.00 / 1.00	

## Question 15

How many entries are reported for position 175672 on Chr1?

You entered:

Your Answer	Score	Explanation
2	✓ 1.00	
Total	1.00 / 1.00	

## Question 16

How many variants are called on Chr3?

You entered:

Your Answer	Score	Explanation
398	✓ 1.00	
Total	1.00 / 1.00	

## Question 17

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

You entered:



Your Answer		Score	Explanation
392	✓	1.00	
Total		1.00 / 1.00	

## Question 18

How many entries are indels?

You entered:

Your Answer		Score	Explanation
320	✓	1.00	
Total		1.00 / 1.00	

## Question 19

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

You entered:


Your Answer		Score	Explanation
2	✓	1.00	
Total		1.00 / 1.00	

## Question 20

What type of variant (i.e., SNP or INDEL) is called at position 11937923 on Chr3?

**You entered:**

SNP

Your Answer	Score	Explanation
SNP	 1.00	
Total	1.00 / 1.00	