Module 2 Exam

20/20 points (100%)

Quiz, 20 questions

✓ **Congratulations! You passed!**

Next Item

✓ 1 / 1 points

1.

How many alignments does the set contain?

221372

Correct Response

First, a couple of introductory comments. A BAM file contains alignments for a set of input reads. Each read can have 0 (none), 1 or multiple alignments on the genome. These questions explore the relationships between reads and alignments.

The number of alignments is the number of entries, excluding the header, contained in the BAM file, or equivalently in its SAM conversion. To find the number of alignments, we can apply ('%' denotes the terminal prompt):

which will list the number of alignments on line 1. An alternate method would be to count the number of lines in the converted SAM file (header excluded):

Note that, if the file was created with a tool that includes unmapped reads into the BAM file, we would need to exclude the lines representing unmapped reads, i.e. with a '*' in column 3 (chrom):

✓ 1 / 1 points

2.

How many alignments show the read’s mate unmapped?

65521

Correct Response

An alignment with an unmapped mate is marked with a ‘*’ in column 7. Note that the question asks for alignments, not reads, so we simply count the number of lines in the SAM file with a ‘*’ in column 7:

✓ 1 / 1 points

3.

How many alignments contain a deletion (D)?

2451

Correct Response

Deletions are be marked with the letter ‘D’ in the CIGAR string for the alignment, shown in column 6:

✓ 1 / 1 points

4.

How many alignments show the read’s mate mapped to the same chromosome?

150913

Correct Response

Alignments with the read’s mate mapped to the same chromosome are marked with a ‘=’ in column 7:

1 / 1 points

5.

How many alignments are spliced?

0

Correct Response

A spliced alignment will be marked with an 'N' (intron gap) in the CIGAR field:

1 / 1 points

6.

How many alignments does the set contain?

7081

Correct Response

We first need to construct the reduced set, i.e. to extract from the original set only those alignments in the specified region. For this, we need to sort and index the file:

This will create the file 'athal_wu_0_A.sorted.bam'. We then index this file:

This will create the index file 'athal_wu_0_A.sorted.bam.bai' in the current directory. Lastly, we extract alignments in the specified range:

The option '-b' will generate output in BAM format. The resulting BAM file will be sorted, so it can be indexed directly if needed.Common pitfalls: make sure to specify the correct reference sequence ('Chr3', not 'chr3') and exclude ',' when representing the query coordinates. Also, make sure to use the sorted and index BAM file. To determine the number of alignments in the new (region) file, we can use the same commands as for Q1, e.g.:

Module 2 Exam

20/20 points (100%)

Quiz, 20 questions

✓ 1 / 1 points

7.

How many alignments show the read’s mate unmapped?

1983

Correct Response

We use the same commands as for Q2, but for the new BAM file:

✓ 1 / 1 points

8.

How many alignments contain a deletion (D)?

31

Correct Response

We use the same commands as for Q3, but for the new BAM file:

✓ 1 / 1 points

9.

How many alignments show the read’s mate mapped to the same chromosome?

4670

Correct Response

We use the same commands as for Q4, but for the new BAM file:

Module 2 Exam

Quiz, 20 questions

20/20 points (100%)

✓ 1 / 1 points

10.
How many alignments are spliced?

0

Correct Response
We use the same commands as for Q5, but for the new BAM file:

✓ 1 / 1 points

11.
How many sequences are in the genome file?

7

Correct Response
This information can be found in the header of the BAM file. Starting with the original BAM file, we use samtools to display the header information and count the number of lines describing the sequences in the reference genome:

✓ 1 / 1 points

12.
What is the length of the first sequence in the genome file?

29923332

Correct Response
The length information is stored alongside the sequence identifier in the header (pattern ‘LN:seq_length’):

Module 2 Exam

20/20 points (100%)

Quiz, 20 questions

✓ 1 / 1 points

13.

What alignment tool was used?

stampy

Correct Response

The program name is listed in the ‘@PG’ line in the BAM header (pattern ‘ID:program_name’):

The ‘^’ sign in the search pattern tells the grep function to match the pattern ‘@PG’ at the start of the line.

✓ 1 / 1 points

14.

What is the read identifier (name) for the first alignment?

GAII05_0002:1:113:7822:3886#0

Correct Response

This information is shown in column 1 of the first alignment record in the SAM file:

✓ 1 / 1 points

15.

What is the start position of this read’s mate on the genome? Give this as ‘chrom:pos’ if the read was mapped, or ‘*” if unmapped.

Chr3:11700332

Correct Response

The location of the read’s mate is contained in column 7 (chrom) and column 8 (start position), if the mate is mapped. If the mate is unmapped, it will be marked with a ‘*’ in column 7. To answer the question, we will need to observe these fields in the first SAM record:

✓ 1 / 1 points

16.

How many overlaps (each overlap is reported on one line) are reported?

3101

Correct Response

We start by running BEDtools on the alignment set restricted to the specified region (Chr3:11777000-11794000) and the GTF annotation file listed above. To allow the input to be read directly from the BAM file, we use the option ‘-abam’; in this case we will need to also specify ‘-bed’ for the BAM alignment information to be shown in BED column format in the output:

This will create a file with the following format: Columns 1-12 : alignment information, converted to BED format Columns 13-21 : annotation (exon) information, from the GTF file Column 22 : length of the overlapAlternatively, we could first convert the BAM file to BED format using ‘bedtools bamtobed’ then use the resulting file in the ‘bedtools intersect’ command. To answer the question, the number of overlaps reported is precisely the number of lines in the file (because only entries in the first file that have overlaps in file B are reported, according to the option ‘-wo’):

✓ 1 / 1 points

17.

How many of these are 10 bases or longer?

2899

Correct Response

The size of the overlap is listed in column 22 of the ‘overlaps.bed’ file. To determine those longer than 10 bases, we extract the column, sort numerically in decreasing order, and simply determine by visual inspection of the file the number of such records. For instance, in ‘vim’ we search for the first line listing ‘9’ (‘:/9’), then determine its line number (Ctrl+g). Alternatively, one can use grep with option ‘-n’ to list the lines and corresponding line numbers:

or:

For the latter, the last ”10” line will be immediately above the first “9”, so subtract 1 from the answer.

✓ 1 / 1 points

18.

How many alignments overlap the annotations?

3101

Correct Response

Columns 1-12 define the alignments:

Potential pitfalls: Multiple reads may map at the same coordinates, so the information in columns 1-3 is insufficient. The minimum information needed to define the alignments is contained in columns 1-5, which include the read ID and the flag, specifying whether this is read 1 or read 2 in a pair with the same read ID).

✓ 1 / 1 points

19.

Conversely, how many exons have reads mapped to them?

21

Correct Response

Columns 13-21 define the exons:

✓ 1 / 1 points

20.

If you were to convert the transcript annotations in the file “athal_wu_0_A_annot.gtf” into BED format, how many BED records would be generated?

4

Correct Response

This question simply asks for the number of transcripts in the annotation file, since the BED format would represent each transcript on one line. This information can be obtained from column 9 in the GTF file: