answers

December 15, 2020

1 Answers

To run the commands alongside the answers you must first have run all of the tutorial commands for that section and generated the output files they reference.

1.1 Tutorial sections

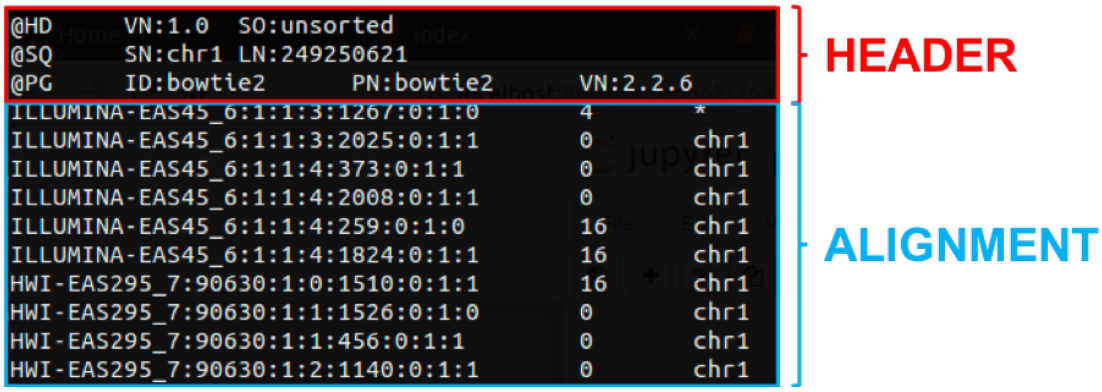
Let's go to our data directory.

```
[ ]: cd data
```

1.2 Aligning the PAX5 sample to the genome

Q1. How can you distinguish between the header of the SAM format and the actual alignments?

Each header line begins with the **@** character followed by one of the two-letter header record type codes and each alignment line starts with the read name.



Q2. What information does the header provide you with?

The header is categorised using two letter record type codes.

Record type	Description
@HD	header line with general information about the file
@SQ	reference sequence dictionary and alignment order
@PG	Programs used to generate the alignment file

Within each of these categories are a series of tags and values.

Record type	Tag	Description
@HD	VN	format version
@HD	SO	alignment sorting order
@SQ	SN	reference sequence name
@SQ	LN	reference sequence length
@PG	ID	program record identifier
@PG	PN	program name
@PG	VN	program version number
@PG	CL	command used when the program was run

Q3. Which chromosome are the reads mapped to?

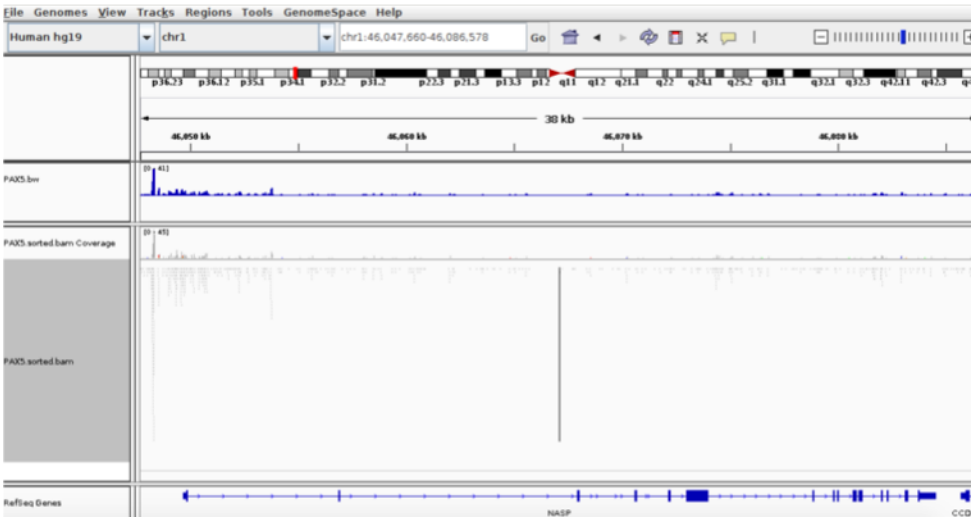
The reads are aligned to chromosome 1 (**chr1**). You're looking for this line in the header.

```
@SQ SN:chr1 LN:249250621
```

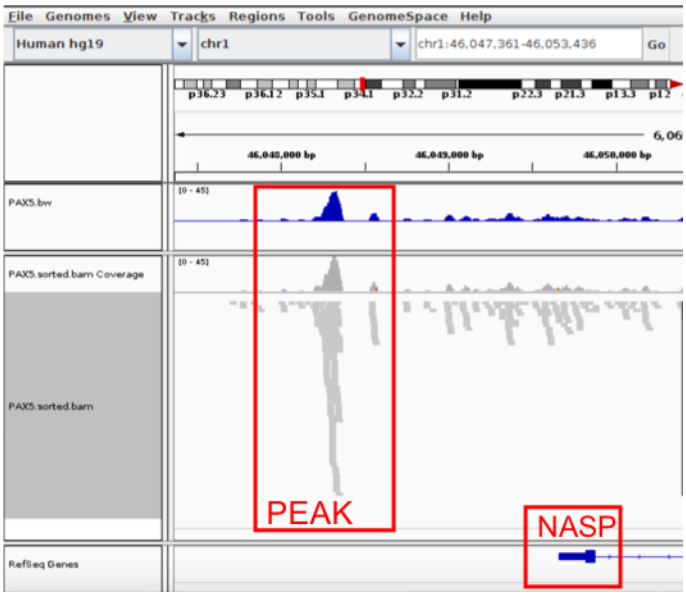
1.3 Visualising alignments in IGV

Q1. Look for gene NASP in the search box. Can you see a PAX5 binding site near the NASP gene?

When you have loaded your PAX5.sorted.bam and PAX5.bw into IGV, made the necessary adjustments and searched for NASP, you should see the following.



There is a peak representing a PAX5 binding site approximately 1kb upstream of the NASP gene, in the promoter region.



Q2. What is the main difference between the visualisation of BAM and bigWig files?

bigWig files display dense, continuous data as a graph whereas BAM files display the read alignment pileup which is discrete.

1.4 Aligning the control sample to the genome

bowtie2 -k 1 -x bowtie_index/hs19 Control.fastq.gz -S Control.sam

samtools view -bSo Control.bam Control.sam

samtools sort -T Control.temp.bam -o Control.sorted.bam Control.bam

samtools index Control.sorted.bam

1.5 File formats

Q1. The simplest bed file contains just three columns (chromosome, start, end) and is often called BED3 format. What extra columns does BED6 contain?

BED6 contains **BED3** columns with the addition of three extra columns: **name**, **score** and **strand**

Q2. In the above examples, what are the lengths of the intervals?

The intervals in the example were **50**, **500** and **200** respectively. Remember that the start coordinates are 0-based which means that you can just subtract the start position from the end position.

chromosome	start	end	length
chr1	50	100	100 - 50 = 50
chr1	500	1000	1000 - 500 = 500
chr2	600	800	800 - 600 = 200

Q3. Can you output a BED6 format with a transcript called "loc1", transcribed on the forward strand and having three exons of length 100 starting at positions 1000, 2000 and 3000?

chromosome	start	end	name	score	strand
chr1	999	1099	loc1	.	+
chr1	1999	2099	loc1	.	+
chr1	2999	3099	loc1	.	+

Q4. What additional information is given in the narrowPeak file, beside the location of the peaks?

narrowPeak files are in BED6+4 format which contains the peak locations together with **overall enrichment**, **peak summit**, **p-value** and **q-value**.

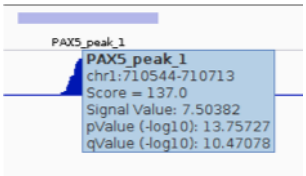
column name	description
signalValue	overall/average enrichment for the region
pValue	-log10(pvalue) for the peak summit

column name	description
qValue	-log10(qvalue) for the peak summit
peak	summit position relative to peak start

Q5. Does the first peak that was called look convincing to you?

Look at the peak in IGV by entering the coordinates in the search box (i.e. *chr1:710543-710713*).

If you hover over the peak annotation, you can see some of the information produced by MACS2 for this peak.



There are 29 reads piled up under this peak which has a q-value of 10. Remember that this is a negative log and so is equivalent to 1e-10.

chromosome	start	end	name	score	strand	signalValue	pValue	qValue	peak
chr1	710543	710713	PAX5_peak_1	137	.	7.50382	13.75727	10.47078	97

Q6. In the small example table above, why have the coordinates changed from the BED description?

From the BED file:

chromosome	start	end
chr1	50	100

From the GTF file:

chromosome	start	stop
chr1	51	100

Notice that the start values differ by 1. This is because the start coordinates in BED format are **0-based** (i.e. first base is at position 0) while the start coordinates in GTF format are **1-based** (i.e. the first base is at position 1).

1.6 Inspecting genomic regions using bedtools

Q1. Looking at the output of the `bedtools genomecov` we ran, what percentage of chromosome 1 do the peaks of PAX5 cover?

The peaks of PAX5 cover **0.28%** of chromosome 1 (chr1).

The output of `bedtools genomecov` contains 5 columns which represent:

1. chromosome (or entire genome)
2. depth of coverage from features in input file
3. number of bases on chromosome (or genome) with depth equal to column 2
4. size of chromosome (or entire genome) in base pairs
5. fraction of bases on chromosome (or entire genome) with depth equal to column 2

The output for chromosome 1 was:

```
chr1    0   248561847   249250621   0.997237
chr1    1   602947      249250621   0.00241904
chr1    2   68077       249250621   0.000273127
chr1    3   15400       249250621   6.17852e-05
chr1    4   963         249250621   3.86358e-06
chr1    7   1387        249250621   5.56468e-06
```

The important value here is in column 5 - fraction of bases with depth equal to column 2.

There are two ways to get the percentage. The first is to add together the fraction of bases which have a depth > 1 on chr1 and convert it into a percentage.

```
0.00241904 + 0.000273127 + 6.17852e-05 + 3.86358e-06 + 5.56468e-06
= 0.00276338
= 0.28%
```

Alternatively, you could subtract the fraction of bases with 0 coverage from 1.

```
1 - 0.997237 = 0.002763 = 0.28%
```

Q2. Looking at the output from `bedtools intersect`, what proportion of PAX5 peaks overlap genes?

72% of the PAX5 peaks overlap genes (a proportion of **0.72**).

You need to divide the number of peaks overlapping genes (2722) by the total number of peaks (3799).

Q3. Looking at `PAX5_closestTSS.txt`, which gene was found to be closest to MACS peak 2?

The closest gene to PAX5_peak_2 is a **lincRNA** called **RP11-206L10**.

1.7 Motif analysis

Q1. Which motif was found to be the most similar to your motif?

Using the output from Tomtom, we can see that the most similar motif is **PAX5**.