Bi188 2013

Computational exercise 2

Due:
Part 1: midnight on May 26th, 2013
Part 2: midnight on June 2nd, 2013

1 The case

You are presented with a case of a cancer patient suffering from leukemia. You decide to sequence the exome of the tumor and of a matched normal sample in order to figure out what the driver mutations are. You carry out SNV and indel analysis in a way similar to what you did in the first computational exercise. You compare the variants in the tumor with the somatic variants from the matched normal sample. You identify novel SNVs and indels in the tumor, however none of them have an effect on protein sequence, splicing or some other obvious genomic feature that could give you reason to think they are driver mutations. You hypothesize that chromosomal rearrangements may be responsible and since your SNV and indel analysis was not designed to capture those, you have failed to see them. You therefore generated whole genome sequencing data (2x75bp reads from fragments with an average length of ∼500bp), with a total coverage of ∼5X.

The sequencing data in FASTQ format can be found here:

```
/woldlab/bostau/data00/pub/georgi/Bi188/Exercise2/normal.end1.fastq
/woldlab/bostau/data00/pub/georgi/Bi188/Exercise2/normal.end2.fastq
/woldlab/bostau/data00/pub/georgi/Bi188/Exercise2/tumor.end1.fastq
/woldlab/bostau/data00/pub/georgi/Bi188/Exercise2/tumor.end2.fastq
```

Note that these are uncompressed reads (they are uncompressed in order to carry out paired-end alignment with Bowtie2).

2 Questions

2.1 Part I. Identifying chromosomal rearrangements

In this part of the exercise you will have to develop a way to identify chromosomal rearrangements in both samples. You will then have to examine the effect such translocations may have on the gene(s) affected, and, if you find multiple genes affected, figure out which ones are most likely to be involved in the development of the tumor. As usual, show both your code and reasoning (hint: there is no need for you to try to do anything complicated, simply looking at where reads (or the different ends of reads) map should be sufficient in this case; look very carefully at the SAM format specifications and the bowtie SAM output and you will easily figure out what you need to do).

Useful bioinformatics tips

At this point you have some experience running Bowtie so instead of telling you what exactly to do, I will point you to the manual (which you can find at `http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml`) and advise you to read on how to align paired-end reads. Pay attention to your fragment length (I would allow a maximum fragment length of double the average fragment length), and note that for what you will be doing here, it may be better to only look at really solid alignments, i.e. the `--very-fast` mode is probably better than the `--very-sensitive` one you used last time.

Refer to the first exercise for the other informatics infrastructure components you might need and to the class website for the SAM format specifications (`http://woldlab.caltech.edu/bi188/private/PSets/SAM1.pdf`). You can convert BAM to SAM (and stream the alignments within your python scripts) using `samtools view`.

You may (but not absolutely necessarily will) find the FLAG field in SAM alignments useful. I have provided you with a function for parsing it: `/woldlab/bostau/data00/pub/georgi/Bi188/Exercise2/ExerciseFunctions2.py`

2.2 Part II. Functional genomics analysis

Having identified the putative driver mutation and taking into account the functions of the proteins involved, you carry out RNA-seq on normal and tumor cells. The results (for your convenience, in FPKMs, i.e. the final product of an RNA-seq analysis) are provided here:

`/woldlab/bostau/data00/pub/georgi/Bi188/Exercise2/normal_tumor.RNA-seq.FPKM.txt`

You also carry out a ChIP-seq experiment against Znf384 in normal white blood cells. We have provided the set of binding sites identified from it here, in BED format:

`/woldlab/bostau/data00/pub/georgi/Bi188/Exercise2/Znf384.peaks.bed`

Finally, you carry out ChIP-seq against the H3K27ac histone modification in both normal white blood cells and tumor cells. A combined set of sites enriched for H3K27ac and the strentgh of H3K27ac (measured in RPM, do not worry about cross-sample normalization in this case) are provided here:

`/woldlab/bostau/data00/pub/georgi/Bi188/Exercise2/normal_tumor_H3K27ac_RPM.txt`

Based on these data, what is the biochemical mechanism driving the tumor?