Bi188 2013

Computational exercise 2. Part 1. Answers

1 Questions

1.1 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).

The goal of the exercise is to find large-scale chromosomal rearrangements that might be driving the tumor. To do that, the logical approach is to find such rearrangements in both the tumor and the matched normal sample and then see what has changed during the evolution of the tumor.

It was not stated in the exercise but I said it to each of you in person that you should not worry about inversions for this exercise (though it would not have been too complicated to account for those), thus the main classes of chromosomal rearrangements you had to identify were large deletions and translocations (I said "keep it simple" therefore there was no need to worry about large insertions either). A deletion would show up in the data you had as the two reads in a pair mapping in the correct orientation but at a very large distance from each other, on the ***same chromosome***. A translocation, in contrast, would have the two ends on different chromosomes. There is a lot more to finding the precise breakpoints from high-throughput sequencing data, but this is not a simple problem and people have spent a lot of time writing sophisticated software to do that. For the purposes of this exercise you did not have to worry about that either, just looking at the discordant read pairs was sufficient.

There were several technical problems you had to solve in order to complete the exercise. First, you had to find the discordant read pairs. This should have been very easy as simply looking for alignment entries in the BAM file for which the two ends map to different chromosomes or the two ends are very far apart from each other would have given you those. Second, you had to find the candidate rearrangements from the discordant read pairs, which involves some sort of clustering procedure to identify localized regions on one chromosomes that connect to another localized region somewhere far apart on it or on another chromosomes. You then had to count the reads supporting each such candidate event. Due to mapping errors, you might have found more than 200,000 such events, but fortunately for you, only a handful of them would have had support by a large number of read pairs (and there should have been a clear separation between them and the noise). Finally, you would have had to take the candidate translocations (or deletions) and figured out how they affect genes. The possibilities are endless, but what you would have specifically looked for are events inactivating tumor suppressor genes and events creating new oncogenes (such as the classic BCR-ABL example, fusions involving MLL proteins, etc.). Of course, the latter case is only possible if the fusion puts the two genes ***in frame*** and also preserves enough of their domain structure for the new protein to functions. There is also the possibility of a rearrangements bringing a gene under the control of new regulatory elements, as well as other exotic scenarios, but none of that is going on here.

Here is what I did:

```
/woldlab/bostau/data00/pub/georgi/Bi188/Exercise2/DiscordantReadPairs.py
```

```
python DiscordantReadPairs.py max_fragment_length merge_radius outfile
```

Which takes a stream of SAM alignments (from `samtools`), finds discordant read pairs and groups them into candidate regions of translocations and deletions (the "maximum fragment length" specifies the maximum allowed distance between the two ends when they map on the same chromosome and the "merge radius" refers to the radius for merging reads into a bigger region around a candidate translocation or deletion site)

```
samtools view normal.bam | python DiscordantReadPairs.py 1000 1000 Discordant_Pair_Clusters_normal
samtools view tumor.bam | python DiscordantReadPairs.py 1000 1000 Discordant_Pair_Clusters_tumor
```

If you did everything correctly, you should have found the following well-supported rearrangements:

- The normal sample contains a translocation creating a NANOGP1:ZMYM5 fusion
- The tumor contains the following fusions: P300:Znf384, TNNI2:ANKRD65, MYOG:chr13, intergenic , TNNC1:CYP26C1, NRXN3:HDAC9, SEMA4D:DOK2, NANOGP1:ZMYM5, ABHD2:SHANK3.

You may have written some code to tell you what is in frame and what is not, but since the number of rearrangements is small, you could have also done it by hand. In any case, only one of these makes sense as a driver mutation and it is the P300:Znf384 fusion.

Scoring: 6 points total for Part I

1.2 Part II. Functional Genomic Analyssis

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. You also carry out a ChIP-seq experiment against Znf384 in normal white blood cells. Finally, you carry out ChIP-seq against the H3K27ac histone modification in both normal white blood cells and tumor cells. Based on these data, what is the biochemical mechanism driving the tumor?).

The obvious hypothesis here is that the P300:ZNF384 is causing the tethering of P300 to locations in the genome where it is not supposed to be normally found (at least in white blood cells). Little is known about ZNF384 at present, but in general ZNF transcription factors usually function as negative regulators of transcription. In contrast, P300 is a histone acetyltransferase (HAT), which generally functions in the positive regulation of transcription. It is a well-known mark of active enhancers, as is the H3K27ac histone modification (though both can be found on promoters too). Thus the working model here is that the P300:ZNF384 fusions is causing the upregulation of genes that are supposed to be repressed or not so highly expressed through some combination of shifting the chromatin state of enhancers from repressed to active (and possibly, permissive to binding by other positive regulators of transcription, which then contribute further to activation) and direct activation at promoters. Parsing these further was not necessary for this exercise. A sufficient test of that hypothesis here consists of the following (not necessarily in this order):

1. Intersecting the ZNF384 binding sites with regions of H3K27ac enrichment (you would have had to write some code to that)

2. Showing that ZNF384 binding sites show increased H3K27ac levels while other H3K27ac sites do not (you would have probably had to write some code to that too)

3. Showing that the closest genes to ZNF384 binding sites (which is a rough if imperfect proxy for identifying the potential regulatory targets of those sites) have increased expression leves as measured by RNA-seq (you would have probably had to write some code to that too, using the annotation GTF file provided in previous exercises).

4. Identifying genes differentially expressed in the tumor (mostly upregulated), and comparing to genes targeted by ZNF384 (most of them are). This is a simple spreadsheet operation (no need to write code really except for comparing lists of genes) involving taking the ratios between the FPKM values in the two samples, with the usual caveats about not dividing by zero and not paying too much attention to big differences between small numbers, then setting some reasonable thresholds (say, at least 1 FPKM in one of the samples, and 2-fold change) to call things as differentially expressed. The precise parameters should not make too much of a difference. In particular, a gene the expression of which changes in an especially drastic way is KIT (from close to 0 to more than 3500 FPKM), which is also well known to have a viral oncogene homolog (v-kit) and to be mutated in many cancers, and happens to have a ZNF384 binding site a little bit downstream of its promoter.

Admittedly, we did not discuss histone modifications in class, but there is plenty of literature on the subject in PubMed providing the background, and in real life, finding genes and processes one knows little about to be important to the topic studied then having to learn about them is quite common.

Scoring: 4 points total for Part II

In order to receive all 4 points, you would have to present a sufficiently detailed and coherent outline of the reasoning described above, and ideally, some plots showing the differences in expression, and overlap between ZNF384 sites and the union of H3K27ac sites in the two conditions, and the changes in H3K27ac between the two conditions for ZNF384-positive and ZNF384-negative H3K27ac sites.

P.S. This question is based on the real-life case of the scientist whose life was saved by sequencing the genome and transcriptome of his tumor that was discussed in class. The P300:ZNF384 fusion and the overexpression of KIT are the actual most significant features of that tumor and the ZFN384 ChIP-seq peak calls are also real, from the ENCODE project.