Julie Garcia
April 9, 2018

Lab 9

This dataset is an already mapped BAM file using DNA-seq reads across the entire exome that were generated from a normal adolescent. We are going to apply some of the BAM file manipulations using samtools to prepare the alignment file with necessary processing steps of sorting the BAM, removing PCR duplicates, indexing the BAM, and summarizing the mapped content. Then we will use this post-processed BAM file to identify structural variants using delly and format with svprops. For this lab, we will be only concerned with germline structural variants (tumor with no matched normal) of translocations, which would be akin to gene fusion events if using RNA-seq reads. Each of the pre-processing steps may take up to 30 minutes each base on the file size, so plan accordingly.

1) First download the BAM file from DropBox as well as the hg19 fasta reference sequence. This file is >3GB, so be patient.
2) Using samtools, sort the BAM file based on coordinate and then calculate the mapping statistics. How many total reads are properly paired and what percentage is this?

```
samtools sort -@ 8 –T /tmp/subj1 –o subj1.sorted.bam
subj1_dnaseq.bam
samtools flagstat subj1.sorted.bam
```

Output:

```
44129985 + 0 in total (QC-passed reads + QC-failed reads)
2639689 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
44129985 + 0 mapped (100.00% : N/A)
41490296 + 0 paired in sequencing
20807236 + 0 read1
20683060 + 0 read2
37423982 + 0 properly paired (90.20% : N/A)
38623742 + 0 with itself and mate mapped
2866554 + 0 singletons (6.91% : N/A)
78200 + 0 with mate mapped to a different chr
41110 + 0 with mate mapped to a different chr (mapQ>=5)
```

37423982 reads were properly paired at 90.20%.

3) Calculate how many PCR duplicates there are in the BAM file (use –S argument) using samtools, remove them, and index this reduced file.

The samtools documentation said that rmdup is obsolete and that we should use markdup instead. To use this I needed to run fixmate first to add ms and MC tags before running markdup, so I had to resort by name, run fixmate, then sort by coordinate again and then run markdup, then index. Running markdup, the –S option was not available, but the –s option was.

```
samtools sort -n -@ 8 -o subj1.namesort.bam subj1.sorted.bam
samtools fixmate -m subj1.namesort.bam subj1.fixmate.bam
samtools sort -@ 8 -T /tmp/subj1 -o subj1.positionsort.bam
subj1.fixmate.bam
samtools markdup -s -r subj1.positiionsort.bam subj1.markdup.bam
samtools index sub1.sorted.bam
```

```
Output:
READ 44129985 WRITTEN 30393423
EXCLUDED 2639689 EXAMINED 41490296
PAIRED 38623742 SINGLE 2866554
DULPICATE PAIR 11680056 DUPLICATE SINGLE 2056506
DUPLICATE TOTAL 13736562
```

4) Using Delly, calculate germline structural variants from the final BAM file you created in question 3 using the hg19 fasta sequence. Next convert the bcf output file to a tab file with svprops.

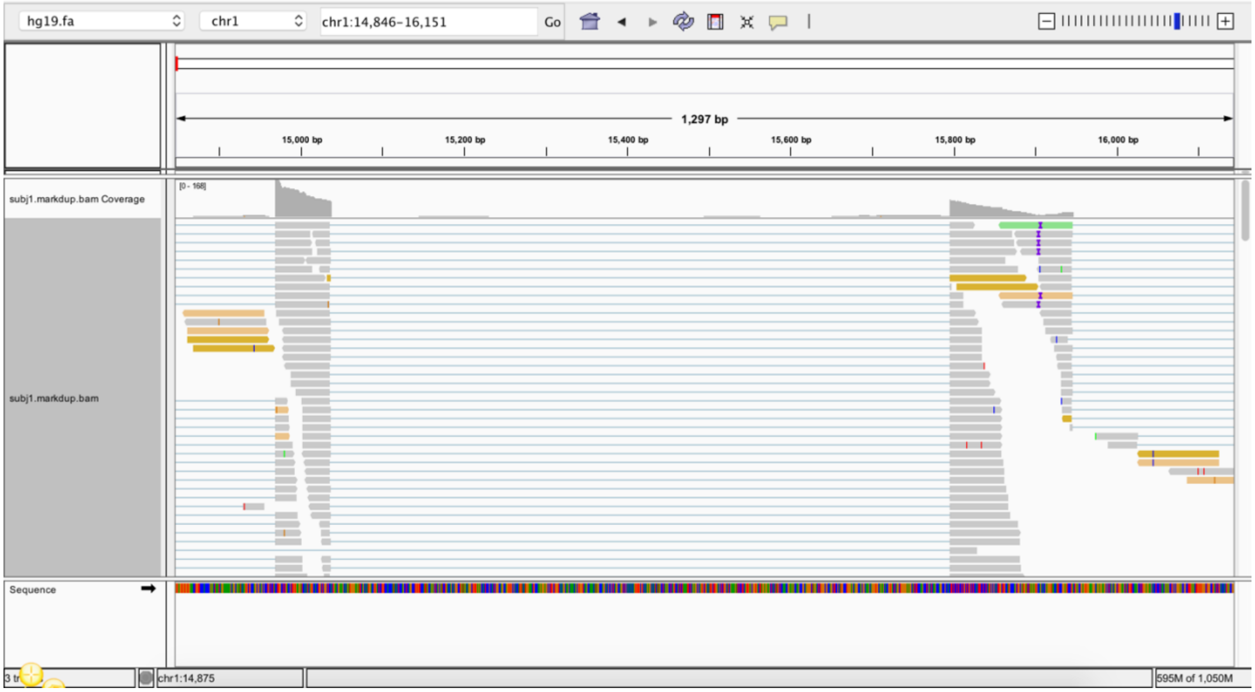
```
delly call -g hg19.fa -o subj1.bcf subj1.markdup.bam
./src/svprops subj1.bcf > subj1.tab
```

5) Read the converted file into R and report the germline translocation (BND) with the most paired end support (PEsupport).

```
data = read.table("/usr/local/bin/svprops/subj1.tab", header=T)
colnames(data)

data.sorted = data[order(-data$PEsupport),]
data.sorted[1,]
chr16 226793 chr16 227367 DEL00035270 with PESupport = 50072
```

6) Now use IGV to pull up this translocation. First load the BAM file (has to be indexed to view in IGV) and go to the first coordinate location from the Delly output. Next, right click and select to color alignments by insert size. The colors of certain reads will tell you which chromosome the mate for a pair is mapping (see color legend below). You can also right click on the track and collapse the reads to view all in a single window. Finally, make sure in Preferences that mapping quality values >1 are being shown and right click on one of the colored reads (should be light green if looking at chr 2) and select to view mate region in split screen. This will provide the 2 breakpoints for the translocation in a single viewing window. Take a screenshot of this translocation event in IGV (or export as image) and paste into your lab.


hg19.fa
chr1
chr1:14,846-16,151
Go
1,297 bp
15,000 bp
15,200 bp
15,400 bp
15,600 bp
15,800 bp
16,000 bp
subj1.markdup.bam Coverage
[0 - 168]
subj1.markdup.bam
Sequence
chr1:14,875
595M of 1,050M