Data Interpretation of RNASeq and Bisulfite sequencing Data in Cancer Research (March 2019)

Exercise: Command line tools (lectured by Martina Fröhlich)

In the theory section you have learned how a BAM file looks like. Now you will learn how to work with it. Usually, most tasks can be solved with clever combinations of SAMtools, BEDtools and Unix tools such as wc, grep, awk.

1. Basic tasks on BAM and VCF files

The files for this exercise are located in /bfg/data/courses/deNBI2019/commandline_IGV. As BAM files and files downloaded from database can be very huge, try to avoid copying them. For practical reasons, you can for example generate a softlink in your home directory that links to the BAM file or the complete folder you are working with. For this

- go into your home directory (cd) and type

  mkdir deNBI2019_IGV # this will be your working directory for this session
  cd deNBI2019_IGV
  *ln -s /bfg/data/courses/*deNBI2019*/commandline_IGV/*

  However, you won't be able to make changes in this linked folder, so your results need to be placed in the working directory.

- Go to commandline_IGV/bam_files
  Here, you see two BAM files from DNA Whole Exome Sequencing for patient X. One contains the alignments for the tumor sample (tumor_DNA_PatientX.bam) and the other for the matching control sample (control_DNA_PatientX.bam).
  (Note: For practical reasons, the BAM and VCF files used in this exercise contain only reads mapping to and variants present in specific areas of the genome. As a result, most of the regions in the genome have no reads mapping to them.)
- The BAM files are binary versions of the SAM files. In Section 2, you will learn how to open or analyse them.

In the VCF files (located in commandline_IGV/vcf_files) you can find the identified variants (somatic and germline) for PatientX and PatientY.
Many tools you will work with (e.g. Bedtools closest or Tabix) require the files to be sorted. Most of the files we provided to you have already been sorted. However, the VCF files are unsorted. You can sort a simple VCF file with the classical UNIX command "sort" by

sort –k1,1 –k2,2n file.vcf

- As your VCF file contains a header, you need to use this command instead:

  (head -n 1 file.vcf && tail -n +2 file.vcf | sort –k1,1 –k2,2n) > file_sorted.vcf

- Sort the files snvs_PatientX.vcf and snvs_PatientY.vcf

2. Introduction to Samtools

SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format. (http://samtools.sourceforge.net, http://www.htslib.org/doc/samtools.html)

2.1 Samtools view

With Samtools view, you can read the BAM file. Usually, you don't want to read the entire BAM file at once, but you want to browse through it in a specific manner.

- Type samtools view --help to get usage information.
- Inspect the header of the BAM file
- Now, use a pager (e.g. more) in combination with Samtools view to look through the first reads of the BAM file

  Usually, the output of samtools view is written on stdout. However, you can also use a pipe "|" to feed it into another program or use the option –b to generate a new BAM file out of it. Many tools require the BAM file to be sorted and indexed. To achieve this, you can use Samtools sort and Samtools index.

2.2 Samtools idxstats and samtools flagstat

To get an overview of the reads in your BAM file, you can also use Samtools idxstats and Samtools flagstat

- Type
  samtools idxstats tumor_DNA_PatientX.bam

- Can you identify how many reads are mapped to which chromosome?

- Type

  samtools flagstat tumor_DNA_PatientX.bam

- How many reads are properly paired?
- How many reads have a mate mapped to a different chromosome?

2.3 Combining Samtools and Unix commands

awk is an interpreted programming language designed for text processing. It is a standard feature of most Unix-like operating systems and very useful for performing simple data extraction and reporting tasks. Below are some examples how to combine samtools and awk for getting information about the alignments.

Check, how many reads are in the BAM file:
samtools view file.bam | wc -l

Extract a column, e.g. column 6 (CIGAR strings):
samtools view file.bam | awk '{print $6}'

Extract by pattern matching all reads with softclipped bases in CIGAR:
samtools view file.bam | awk '($6 ~ /S/)'

3. Bedtools

Bedtools is a set of powerful tools to analyze sequencing data. Today, we will only cover one of the tools within Bedtools, but you might want to have a look at the complete tool set. (http://bedtools.readthedocs.io/en/latest/)

3.1 Preparation

Some tools in Bedtools require the input datafiles to be in BED file format. In its simplest form a bed file contains the fields

1. chrom (name of chrom)
2. chromStart (start position of the feature)
3. chromEnd (end position of the feature)

One way to generate a BED file from an existing e.g. TAB delimited file is by using AWK. Here, a simple bed file is generated from the sorted PatientX's VCF file:

```
awk -v OFS='\t' 'NR>1 {print "chr"$1,$2,$2}' file_sorted.vcf > file_sorted.bed
```

Note: There is also a tool called vcf2bed, but we won't use this here

3.2 Bedtools closest

You can now use closestBed (a.k.a. Bedtools closest) to find for example the gene or exon that is closest to the variant in your file. Here, we were interested in the gene that is closest to the variants. For this, you need a file containing the information about the start and end points of all annotated genes. You can use for example the RefSeq annotation from Annovar (http://annovar.openbioinformatics.org) (Downloaded by us for you and placed in *commandline_IGV//*databases)

You can find the closest gene with

bedtools closest -a file_sorted.bed –b RefSeq_Nov15_2011_from_annovar.bed.gz -D "b" > result.bed

- Which is the gene closest to the variant at coordinate chr3:137717991?