UNIX command line tools for manipulation and analysis of genomic data (II)

This practice is focused in Samtools and Bedtools, two command line tools for the manipulation of SAM/BAM and BED files respectively, two file formats that are widely used in genomic data analysis.

INSTALLATION

1. Install libraries
   - MobaXterm

```
apt-get install make
apt-get install gcc-g++
apt-get zlib-devel
apt-get libbz2-devel
apt-get liblzma-devel
apt-get libncurses-devel
```

   - Ubuntu

```
sudo apt-get install make
sudo apt-get install g++
    sudo apt-get libncurses5-dev
sudo apt-get zlib1g-dev
sudo apt-get libbz2-dev
sudo apt-get liblzma-dev
```

2. Download Samtools source (samtools-1.10.tar.bz2) from http://www.htslib.org/download/ into “software” directory

3. Navigate to downloaded samtools directory

```
cd software
```

4. Unzip and compile Samtools

```
bunzip2 samtools-1.10.tar.bz2
tar -xvf samtools-1.10.tar
cd samtools-1.10
./configure
make
```

5. Download Bedtools source code (zip) into “software” directory
https://github.com/jchenpku/bedtools2-cygwin/releases (MobaXterm)
https://github.com/arq5x/bedtools2/archive/v2.29.2.zip (Others)

6. Navigate to downloaded bedtools directory

```
cd ../../software
```

7. Unzip and compile Bedtools

MobaXterm

```
unzip bedtools2-cygwin-2.29.2.zip
cd bedtools2-cygwin-2.29.2
make static
```

Others

```
unzip bedtools2-2.29.2.zip
cd bedtools2-2.29.2
make static
```

Note: Both Samtools and Bedtools can be installed using package managers for Linux (apt-get) and OS X (brew or macports)

Linux:

```
apt-get install samtools
apt-get install bedtools
```

OSX:

```
/usr/bin/ruby -e "$(curl -fsSL
https://raw.githubusercontent.com/Homebrew/install/master
/install)"
brew install samtools
brew install bedtools
```

SAMTOOLS:

- Visualize alignment file in BAM format
  - `samtools view bam/1M68_pH5_0.04CO2_R1.bam | less`

Discussion about SAM flag combinations in this post:
https://ppotato.wordpress.com/2010/08/25/samtool-bitwise-flag-paired-reads/

- Get flag summary of bam alignment
  - `samtools flagstats bam/1M68_pH5_0.04CO2_R1.bam`
- Count reads that are properly aligned (compare with stats)
  - `samtools view -f 2 bam/1M68_pH5_0.04CO2_R1.bam | wc -l`
- Extract reads from a given chromosome (e.g. it can be useful to extract retroviral sequences)
  - `samtools view bam/1M68_pH5_0.04CO2_R1.bam I | less`
- Index file (File must be already sorted by coordinate)

- `samtools sort -o bam/1M68_pH5_0.04CO2_R1.sorted.bam bam/1M68_pH5_0.04CO2_R1.bam`
- `samtools index bam/1M68_pH5_0.04CO2_R1.sorted.bam`

This biostars thread includes figures and an interesting discussion about the origin of duplicates in NGS experiments:
https://www.biostars.org/p/229842/

- Remove PCR or Optical duplicates
  - `samtools sort -n bam/1M68_pH5_0.04CO2_R1.bam | samtools fixmate -m - - | samtools sort -o bam/1M68_pH5_0.04CO2_R1.sorted.bam -`
  - `samtools markdup bam/1M68_pH5_0.04CO2_R1.sorted.bam bamdup/1M68_pH5_0.04CO2_R1.bam`
  - `samtools flagstats bamdup/1M68_pH5_0.04CO2_R1.bam`
  - `samtools view -f 1024 bamdup/1M68_pH5_0.04CO2_R1.bam | head`

Other useful commands included in Samtools

- Extract reads in fastq format (print first 10)
  `samtools fastq bam/1M68_pH5_0.04CO2_R1.bam | head`
- Visualize coverage
  `samtools coverage -A bam/1M68_pH5_0.04CO2_R1.bam`
- For genomic variants detection see mpileup, call

BEDTOOLS

INTERSECT

- Find intersecting regions of two bed files
  - `bedtools intersect -a bed/regions_example.bed -b bed/regions_example2.bed > bed/intersect_1_2.bed`
- Find intersecting regions one file against multiple bed files
  - `bedtools intersect -a bed/regions_example.bed -b bed/regions_example2.bed bed/regions_example3.bed > bed/intersect_1_2-3.bed`
- Count number of overlaps of regions of one bed file which regions in another file (Note that -b can receive BAM, VCF, GFF or BED files
  - `bedtools intersect -c -a bed/regions_example.bed -b bed/regions_example2.bed > bed/regions_exampleIn2.bed`

CONCATENATE

The next two commands were already used in the previous session:

- Concatenate BED files into a
  - `cat bed/regions_example.bed bed/regions_example2.bed bed/regions_example3.bed > bed/all_regions.bed`

- Sort BED files (two methods)
  - `sort -k 1,1 -k2,2n bed/all_regions.bed > bed/all_regions.sorted.bed`
  - `bedtools sort -i bed/all_regions.bed > bed/all_regions.sorted.bed`

MERGE

- Merge regions that overlap in a single bed file
  - `bedtools merge -i bed/all_regions.sorted.bed > bed/all_regions_example_merged.bed`

- Merge regions at less than 20 nt in a single bed file
  - `bedtools merge -d 20 -i bed/all_regions.sorted.bed > bed/all_regions_example_mergedAt20nt.bed`

- Merge regions forcing same strand
  - `bedtools merge -s -i bed/all_regions.sorted.bed > bed/regions_example_merged_samestrand.bed`

FLANK

- Create chromosome size file
  - `grep -P '\tchromosome\t' genes/Saccharomyces_cerevisiae.R64-1-1.99.gff3 | cut -f 1,5 > genes/Saccharomyces_cerevisiae.chromSizes`

- Get 500bp upstream regions of genes
  - `bedtools flank -i bed/s_cerevisiae_genes.bed -g genes/Saccharomyces_cerevisiae.chromSizes -s -l 500 > bed/s_cerevisiae_upstream500.bed`

- Get 500bp downstream regions of genes
  - `bedtools flank -i bed/s_cerevisiae_genes.bed -g genes/Saccharomyces_cerevisiae.chromSizes -s -r 500 > bed/s_cerevisiae_downstream500.bed`

COVERAGE FUNCTIONS

- Compute coverage of regions (genes)
  - `bedtools coverage -a bed/s_cerevisiae_genes.bed -b bam/1M68_pH5_0.04CO2_R1.bam > bed/s_cerevisiae_genes.coverage.tsv`

- Compute coverage of multiple BAM files in separated columns (genes)
  - `bedtools multicov -bed bed/s_cerevisiae_genes.bed -bams bam/1M68_pH5_0.04CO2_R1.bam bam/1M69_pH5_0.04CO2_R2.bam > bed/s_cerevisiae_genes.multicov.tsv`

- Create genome coverage file
  - `bedtools genomecov -bg -i bam/1M68_pH5_0.04CO2_R1.bam > bed/ 1M68_pH5_0.04CO2_R1.bedgraph`

Take a look to other useful Bedtools commands such as annotate, closest, subtract, window, cluster, maskfasta, and shuffle.

Obtain BED file of genes from S. cerevisiae GFF

- Extract the desired fields from gene features (chromosome, start, stop, description, 0, strand)
  - `grep -P '\tgene\t' genes/Saccharomyces_cerevisiae.R64-1-1.99.gff3 | awk '{print $1"\t"$4"\t"$5"\t"$9"\t0\t"$7}' > bed/genes_1_index_strand.bed`

- Shift all regions 1 nucleotide towards 5' end (correct coordinates)
  - `bedtools shift -i bed/genes_1_index_strand.bed -s -1 -g genes/Saccharomyces_cerevisiae.chromSizes > bed/genes_0_index_strand.bed`

- Replace description by Gene ID in genes_0_index_strand.bed
  - `perl -p -e 's/\tID=gene:([^;]+);[^\t]+\t/\t\1\t/' bed/genes_0_index_strand.bed > bed/s_cerevisiae_genes.bed`