Working with genome annotation features

1. Prepare the working directory.

1.1 Create a working directory "/workdir/$USER".

Copy all data files for this exercise from "/shared_data/epigenomics/exercise1/" into the working directory.

```
mkdir -p  /workdir/$USER/exercise1

cd /workdir/$USER/exercise1

cp /shared_data/epigenomics/exercise1/* ./

ls -l
```

1.2 Install Filezilla client on your laptop

Filezilla is a sftp client software. If you do not have any sftp client software on you laptop, download Filezilla client from this page. Double click to install.

https://filezilla-project.org/download.php?show_all=1

*** The installer might prompt you to install other additional software, e.g. virus protector, always click "no" to decline.**

1.3 Install IGV on your laptop

Go to the IGV web site (https://software.broadinstitute.org/software/igv/), click “Download”. Double click the IGV installation tool to install IGV. On Windows computer, the software is installed in the directory C:\Program Files.

1.4 Getting familiar with "screen"

If you do not know about Linux "screen" or "tmux" commands, now it is time to get familiar with it.

Most of the tools you will be using in this exercise take long time to finish. You will need to use the "screen" persistent sessions to run the software, so that you can safely detach from the session, and the job will keep running in the background on the server.

```
#Start screen. You would notice that "[screen 0 ...]" shows up at the header of
the terminal window.
screen

#Now you are in the "screen" persistent session, and you can run any software in
the session now.
ls -l

#To detach from the "screen", press "ctrl-a" "d" ("d" for "detach").
```

```
#After this, you will notice that "[screen 0 ...]" dispears from the header, you
are back to regular session.
#The screen session is still alive in the background.

#To re-attach back to the screen session
screen -r

#You can create many independent sessions within one "screen" by pressing "ctrl-
a" "c" ("c" for "create").
#After the new session is created, "[screen 1 ...]" shows up at the header of the
terminal window.
#You can create as many independent sessions as you want, they will be named
"screen 0","screen 1", "screen 2", et al.

#You can switch between these sessions by pressing "ctrl-a" "n" ("n" for next).

#To kill a screen session
#press "ctrl-d" when you are inside screen session.

#To detach from the screen
#press "ctrl-a" "d" to detach from "screen".

#Please be aware of the difference bwtween "detach" and "kill". With "detach",
you can re-attach back to the session. With "kill", the session is permanently
removed, together with any jobs running in the session.
```

2. Converting between file formats

2.1 GFF3, GTF and BED

You are provided with a gff3 formatted file: ara.gff3. Inspect the content of the file.

```
cd /workdir/$USER/exercise1

less ara.gff3
# press "space" to move to the next page; press "q" to exit
```

Generate some basic statistics of the gff3 file based on the 3rd column "Feature type". The command would tell you the number of gene features in this file, and whether this gff3 file contains non-protein-coding gene feature (e.g. transposable_element, miRNA, et al.).

```
cut -f3 ara.gff3 | sort | uniq -c
```

Converting the gff3 file to a gtf file, and then convert back to a new gff3 file.

```
gffread -E -T -o ara.gtf ara.gff3

gffread -E -G -o ara_converted.gff3 ara.gtf
```

Compare the difference between gff3 and gtf file formats, especially the last column of the two files.

```
head ara.gff3

head ara.gtf
```

By comparing the original ara.gff3 and the new ara_converted.gff3, you would find some information are lost in the ara_converted.gff3. For example, you do not see the "gene" features in the new file.

```
head ara.gff3

head ara_converted.gff3

cut -f3 ara_converted.gff3 | sort | uniq -c
```

Convert the ara.gff3 file to a "bed" formatted file, using the "awk" command.

```
awk 'BEGIN { OFS = "\t" } {if ($3=="gene") print $1, $4-1, $5, ".", ".", $7}'
ara.gff3 > ara.bed

less ara.bed
```

2.2 Extract protein and transcript sequences from the genome.

Inspect the genome sequence file "ara.fasta" and the "ara.gff3" file. You would find that,

- "ara.fasta" uses "1, 2, 3, 4, 5, mitochondria, chloroplast" as chromosome names (in the fasta sequence headers, the text after the first space character is ignored);
- "ara.gff3" uses "Chr1, Chr2, Chr3, Chr4, Chr5, ChrC, ChrM" as chromosome names.

```
grep ">" ara.fasta

cut -f1 ara.gff3|sort |uniq
```

The Linux "sed" command can be used to fix this problem, and write to a new file "ara_2.gff3"

```
sed "s/^Chr//" ara.gff3 | \
sed "s/^C/chloroplast/" | \
sed "s/^M/mitochondria/" > ara_2.gff3

cut -f1 ara_2.gff3|sort |uniq
```

Now you can use gffread to extract protein and transcript sequences. The output are two new files: transcript.fasta and protein.fasta

```
gffread ara_2.gff3 -g ara.fasta -w transcript.fasta -x protein.fasta

ls -lrt
```

2.3 BEDgraph and BigWig files

As wig file format is replaced by BigWig format now, we will work with BEDgraph and BigWig file here.

If the input file sample1.bedGraph is not sorted, use the Linux "sort" function to sort the file first (sort by column 1 and 2)

```
sort -k1,1V -k2,2n sample1.bedGraph > sorted.bedGraph
```

Run "bedGraphToBigWig", which a tool in the UCSC Kent Utilities package.

```
export PATH=/programs/kentUtils/bin:$PATH

bedGraphToBigWig sorted.bedGraph genome.txt sample1.bw
```

3. Visualize the BigWig file with IGV genome browser

4.2 Download files to your laptop.

Using Filezilla to download the "sample1.bw" files.

4.3 Launch IGV on your laptop.

Double click “igv.bat” in the directory "C:\Program Files\IGV_2.8.11" to start IGV. It might take a few seconds before you see the software starting.

Most commonly used genomes are already loaded in IGV. In this exercise, you will use the "A thaliana (TAIR 10)". In the pull-down menu at the upper-left corner, click "More" and select "A thaliana (TAIR 10)".

From menu “File” -> “Load file”, open the “sample1.bw”.

Select "1" from the "chromosome" pull down menu.