Genomic data visualization

Work locally on your laptop

The following instructions have been tested on Ubuntu. Mac and Windows users can use Ubuntu in the Virtual Machine (setup distributed by Stefan Wyder)

1. IGV:
   - The distribution of the VM already contains an IGV under /home/student/software
   - Download and install IGV:
     ```
     wget
     http://data.broadinstitute.org/igv/projects/downloads/IGV_2.
     3.72.zip
     unzip IGV_2.3.72.zip
     ```
   - Run IGV: execute igv.sh within the IGV distribution
     ```
     ./igv.sh
     ```

2. OmicCircos:
   - Download and install:
     ```
     R
     source("https://bioconductor.org/biocLite.R")
     biocLite("BiocUpgrade")
     biocLite("OmicCircos")
     ```

3. Circos (we will not use Circos):
   - Download Circos:
     ```
     wget http://circos.ca/distribution/circos-0.69-2.tgz
     tar zxvf circos-0.69-2.tgz       #unzip folder
     ```
   - List missing Perl modules:
     ```
     bin/circus –modules
     ```
   - Install missing Perl modules:
     ```
     sudo perl –MCPAN -e shell
     install Config::General
     …
     exit
     ```

     If installation of GD perl modul fails:
     ```
     sudo apt-get install libgd-perl
     ```
   - Run Circos: execute bin/circos within your Circos distribution
     ```
     bin/circos
     ```

Exercise 1: IGV

In this exercise you will have a look at human genome data which is available from the IGV server.

1. Human alignment data
    a. Load the human hg18 genome
    b. Load files from server: Tutorial → SNP Validation → SNP Calls and WGS Alignments
    c. Explore the data
    d. Jump to position chr1:159,467,422-159,469,014 and try to understand coverage plot and the meaning of the colored reads

2. Human expression data
    a. Load human hg19 genome
    b. Load files from server: Body Map 2.0 (Illumina HiSeq)
        → Transcripts Assemblies → Cufflinks → Cufflinks_brain
        → Coverage → brain
        → Alignments → Merged 50 bp and 75 bp (hiSeq) → brain
    c. Explore the data
    d. Change the track display type of the brain coverage track
    e. Display the splice junction track and try to understand it
    f. Create a Sashimi Plot
    g. Display the HOXA1, HOXA2 and HOXA3 in parallel

Exercise 2: OmicCircos

In this exercise you will get to know the R package OmicCircos which can be used to easily produce Circos like plots. You need to run R via the terminal or Rstudio. We will work with a human data set that comes along with the package.

1. Load data
    a. OmicCircos package and the hg19 genome:
    `library(OmicCircos)`
    `data(UCSC.hg19.chr)`
    → the genome contains informations of the chromosomes including segmentation information. Have a look at the data by using `head(UCSC.hg19.chr)`

b. Mapping data: data frame which includes gene expression data of samples
`data(TCGA.BC.gene.exp.2k.60)`
→ have a look at it with `head(TCGA.BC.gene.exp.2k.60)` and try to understand the data

c. Load link data: contains link information between positions.
`data(TCGA.BC.fus)`
→ have a look at it with `head(TCGA.BC.fus)` and try to understand it

2. We will now produce a circus graphic step by step (you can run the code after each step to see what happens).
   a. To get a perfect circle the plot area needs to be a square and we need to setup an "empty" plot:

```
old.par <- par(no.readonly = TRUE)            #store old par
settings that we can reset it
par(mar=c(0.5, 0.5, 0.5, 0.5))
plot(c(1,800), c(1,800), type="n", axes=FALSE,xlab="",
ylab="")
```

   b. Now we add the segment data (ideogram of human). The track is added at the radius=400, with a circle width of 6 and is of type "chr" (segment data):

```
circos(R=400, type="chr", cir="hg19", print.chr.lab=TRUE,
W=6, scale=TRUE)
```

   c. Add a further mapping track with expression data of the first sample (TCGA.A1.A0SK.01A) as a line plot (type="l"):

```
circos(R=350, type="l", cir="hg19", W=40,
mapping=TCGA.BC.gene.exp.2k.60[, c("chr", "po",
"TCGA.A1.A0SK.01A")], B=TRUE, col="red3", lwd=1, scale=TRUE)
```

   d. Try to add the same data in additional tracks as different plot types (e.g. line in stair steps, horizontal lines and bar charts). Use the help (?circos) to check the options.

   e. It is also possible to plot data of several samples into the same plot. Let's plot the data of the first 3 samples into one line plot

```
circos(R=190, type="ml2", cir="hg19", W=40,
mapping=TCGA.BC.gene.exp.2k.60[, c("chr", "po",
"TCGA.A1.A0SK.01A",   "TCGA.A1.A0SO.01A",
"TCGA.A2.A04W.01A")], B=TRUE, col=c("red3", "green3",
"blue"), lwd=1, scale=TRUE)
```

This will produce an error, as the function contains a small bug. In the multiple line type it expects data for each chromosome, thus it crashes as we do not have data on Y. However, we can simply solve this by adding NA values for chr 24 (Y):

```
circos(R=190, type="ml2", cir="hg19", W=40,
mapping=rbind(TCGA.BC.gene.exp.2k.60[, c("chr", "po",
"TCGA.A1.A0SK.01A",   "TCGA.A1.A0SO.01A",
"TCGA.A2.A04W.01A")], c("24", NA, NA, NA, NA)), B=TRUE,
col=c("red3", "green3", "blue"), lwd=1, scale=TRUE)
```

f. Now we add some links:
circos(R=180, type="link", cir="hg19" , W=40, mapping=TCGA.BC.fus, lwd=1, col="red3")

One can also add polygon links (connects two segments with a polygon graph):
For this you have to provide a data frame with chromosome, start and end position of the first segment in the first 3 columns and chromosome, start and end position of the second segment in the colums 4-6. Example

```
linksPg <- data.frame(chr1=c(1, 1, 5), start1=c(100000,
30000000, 10), end1=c(55000000, 80000000, 10000000),
chr2=c(4,22,6), start2=c(5000000, 50000, 10),
end2=c(100000000, 10500000, 90000000))
circos(R=180, type="link.pg", cir="hg19" , W=40,
mapping=linksPg, lwd=1, col=rgb(0,0,1,0.4))
```

g. Highlight some parts:
To highlight some parts you have to create a vector with:
1: inner radius, 2: outher radius, 3: start chromosome, 4: star position, 5: end chromosome, 6: end position, 7: fill color, 8: border color

```
highlight <- c(180, 410, 2, 63456333, 2, 131988682,
rgb(1,0,0,0.3), "red3")
circos(cir="hg19", W=40, mapping=highlight, type="hl",
lwd=1)
```

h. Try to add some labels (use the TCGA.BC.fus data) on the outside of the ideogram. Use the help (?circos) to check the options or https://www.bioconductor.org/packages/release/bioc/vignettes/OmicCircos/inst/doc/OmicCircos_vignette.pdf

Solutions:

You can find the solutions under following link:
https://www.dropbox.com/s/ervsqhz4jvda664/OmicCircos_solutions.r?dl=0