Homework on Bioconductor / GenomeGraphs

Instructions

Below you will find a series of questions that you should provide answers. To answer completely the following questions you will need a working version of R, Internet and some research skills. The provided R-code had been done in such a way that should run without any issues. One just needs to "copy & paste" the R-code to see relevant information to answer a specific question.

Author: Kyrylo Bessonov
kbessonov@ulg.ac.be
+32 4366 **9544** (office phone)
Office number: 1/16 (section D)
Institute Montefiore (Building B37)

Help: Please contact me in person or via the phone from Mon to Fri from 10am til 6pm for additional help. Alternatively contact me via email at any time.

Questions

1) Find the total number of packages available in the Bioconductor **2.11**. Select one of the packages we did not cover and describe its use and benefits in 5 to 10 sentences.
2) What is the sequence of commands required to load any Bioconductor library?
3) Load library called "*affy*" and provide the correct citation suggested by the authors of the *affy* library (**hint:** look at help file of the library and find the required R command). Briefly describe one of the uses/purposes of the *affy* library.
4) Execute the following R-code and answer the questions below:

```
library("GenomeGraphs");
ensembl_Human_Genes = useMart("ensembl",dataset="hsapiens_gene_ensembl");

gene <- makeGene(id = "ENSG00000115145", type="ensembl_gene_id", biomart =
ensembl_Human_Genes )

transcript <- makeTranscript(id = "ENSG00000115145", type="ensembl_gene_id",
biomart= ensembl_Human_Genes)

gdPlot ( list("gene"=gene, "transcripts"=transcript))
```

- What is the gene name (i.e. hgnc_symbol) and function represented by the Ensembl ID - ENSG00000115145? What is the name of the biomaRt database (object name)? What species does it represents (Mouse, Chicken, Human)? Execute the command below to answer some of these questions

```
getBM(c("ensembl_gene_id", "hgnc_symbol", "description"),
filter=c("with_exon_transcript", "with_protein_id",
"with_transcript_variation"),values=list(TRUE, TRUE, TRUE), ensembl_Human_Genes
) [1:25,]
```

- How many exons does the `gene` object has? Execute the following command

```
attributes(gene)
```

- How many transcript variants does the gene has (refer to the plot)?
- Add additional axis to the existing plot mapping exons to the the chromosomal locations (i.e. genome axis). The label of the axis should be "position (nt)". What command should be used to create an axis object? Copy & paste your resulting plot

5) Extracting information on introns / exons :

- Execute the following command. How many chromosomes do you see? Why the number of chromosomes in this Ensembl dataset is greater than 23 chromosome pairs? What does "MT", "X" and "Y" refer to?

```
getBM("chromosome_name","","", ensembl_Human_Genes)[c(1:22,432:434),1]
```

- Execute the following commands below.
    - How many genes are on Y chromosomes? Roughly which biological function is the most abundant?

```
report = getBM(attributes=c"ensembl_gene_id", "hgnc_symbol", "chromosome_name",
"strand", "start_position", "description"), filters=c("chromosome_name"),
values=c("Y"), mart = ensembl_Human_Genes)
cat("The output file is written in the following DIR:", getwd(), "\n")
write.table(report, "report_on_Y_chromosome.txt", sep = "\t", row.names=FALSE)
```

Note: Please open the written report with Excel or NotePad. The file name is `report_on_Y_chromosome.txt` written in the current directory specified in your R terminal

    - Which of the getBM() function **parameters** allowed to only select only Y chromosome data from the rest?
- Plot a region of chromosome Y by executing the following commands

```
plusStrand <- makeGeneRegion(15195604, 15350000 , chr = "Y", strand = "+",
biomart = ensembl_Human_Genes, dp = DisplayPars(plotId = TRUE, idRotation = 90,
cex = 1.5, idColor = "green3"))

#replace the Ensemble IDs with the gene names
plusStrand@ens[,1]=c("SFPQP1", "SFPQP1", "DPPA2P1")

negStrand <- makeGeneRegion(15195604, 15350000 , chr = "Y", strand = "-",
biomart = ensembl_Human_Genes, dp = DisplayPars(plotId = TRUE, idRotation = 0,
cex = 2, idColor = "green3"))

negStrand@ens[,1]=c("TAB3P1",rep("",4))

gdPlot(list("plus"=plusStrand, "minus"=negStrand, "position"=makeGenomeAxis(),
makeTitle("position (nt)", cex=2,"black",0.1) ), 15185604, 15350000  )
```

    - How many genes are located between the 15195604 and 15350000 positions of the Y chromosome considering both + and – strand? What are their names? Does this

result corresponds to the information retrieved previously by `getBM()`? What are their functions?

- Judging from the generated plot, how many exons does the SFPQP1 genes has? What is their Ensembl Exon IDs? What are their locations on the Y chromosome?
- Change the color of gene names of the previous plot from green to blue and past the resulting plot below
- Without changing the freshly generated objects `plusStrand` and `negStrand` plot region from 15100000 to 15350000 and paste your resulting plot (**hint:** look at the `gdPlot` documentation)
- Produce another plot of a different genomic location using the previous plot as a reference example (i.e. you are not limited by chromosome Y, any location will do)