Searching in Biological Databases (Task 2)

Alex-Alex-Helena

7 oktober, 2021

Introduction

The purpose of this task is to play around with querying different types of biological databases using R. In the following, some questions will be answered using queries.

Question 1:

Retrieve the genome of a cat via its *scientific name* or *taxonomic identifier from NCBI Taxonomy*. Then read the file.

This information will be retrieved using the library `biomartr`. The documentation can be found here

Answer 1:

```
# Retrieval of the genome into a file. THIS CODE TAKES SOME TIME!
file.path <- getGenome(db = "refseq", organism = "Felis catus")
```

```
#> Starting genome retrieval of 'Felis catus' from refseq ...
#>
#> It seems that this is the first time you run this command for refseq.
#> Thus, 'assembly_summary.txt' files for all kingdoms will be retrieved from refseq.
#> Don't worry this has to be done only once if you don't restart your R session.
#>
#> Something went wrong when trying to access the FTP site 'ftp://ftp.ncbi.nlm.nih.gov/'. Sometimes the
#>
#> Completed!
#> Now continue with species download ...
#> File _ncbi_downloads/genomes/Felis_catus_genomic_refseq.fna.gz exists already. Thus, download has be
#> The genome of 'Felis_catus' has been downloaded to '_ncbi_downloads/genomes' and has been named 'Feli
```

```
# or

#file.path <- getGenome(db = "refseq", organism = "9685")
```

```
# Display the genome.
(cat.genome <- read_genome(file.path, format = "fasta"))
```

```
#> DNAStringSet object of length 37:
#>          width seq                                               names
#>  [1] 242100913 ATCAGGAGATCTAGATGCCTG...AAGCACCTTCATGTTCCCAA NC_018723.3 Felis...
#>  [2]     46965 CTTTCTTTTTCTAAAATTCTC...CACCAATTATATGGGACTAG NW_019365239.1 Fe...
#>  [3]     58068 AAATCGTGACACATGCTACAT...GCCTCCTGGGCCTTCTCAGC NW_019365240.1 Fe...
#>  [4]     50743 AGTTATAGTAATCTTCCTAGG...CCTGCCTTCCTTTTCTTTTC NW_019365241.1 Fe...
#>  [5]     22574 CATGATTTAGTGAAAACGTAA...TTCTATTTATCACATTTGTT NW_019365242.1 Fe...
#>  ...       ... ...
#> [33]     61658 TCTCCATCAGTCCCTGTGGAG...ACTGATATTTAAAGAAGAGT NW_019365269.1 Fe...
#> [34]     37620 AGAGCTTACTTAAAAAAAAAT...GGAGATCCACTTGGTTGCAA NW_019365270.1 Fe...
#> [35]     51987 GTCAACCGTCTCCAAAAAAAG...TAGTTCAAACGGTCCAGTCT NW_019365271.1 Fe...
#> [36]     41842 ATTTCCTAAGCGAGGTTACCA...AGGGAAAAGCATGAGCGCGA NW_019365272.1 Fe...
#> [37]   1157532 GAGGCAGCGCCGACTCTGAGC...GTATTTTCCCTGAATGGCTG NC_018725.3 Felis...
```

Question 2:

Find the allele names in the Applied Biosystem identifiler allelic ladder (from the `seqinr` library)

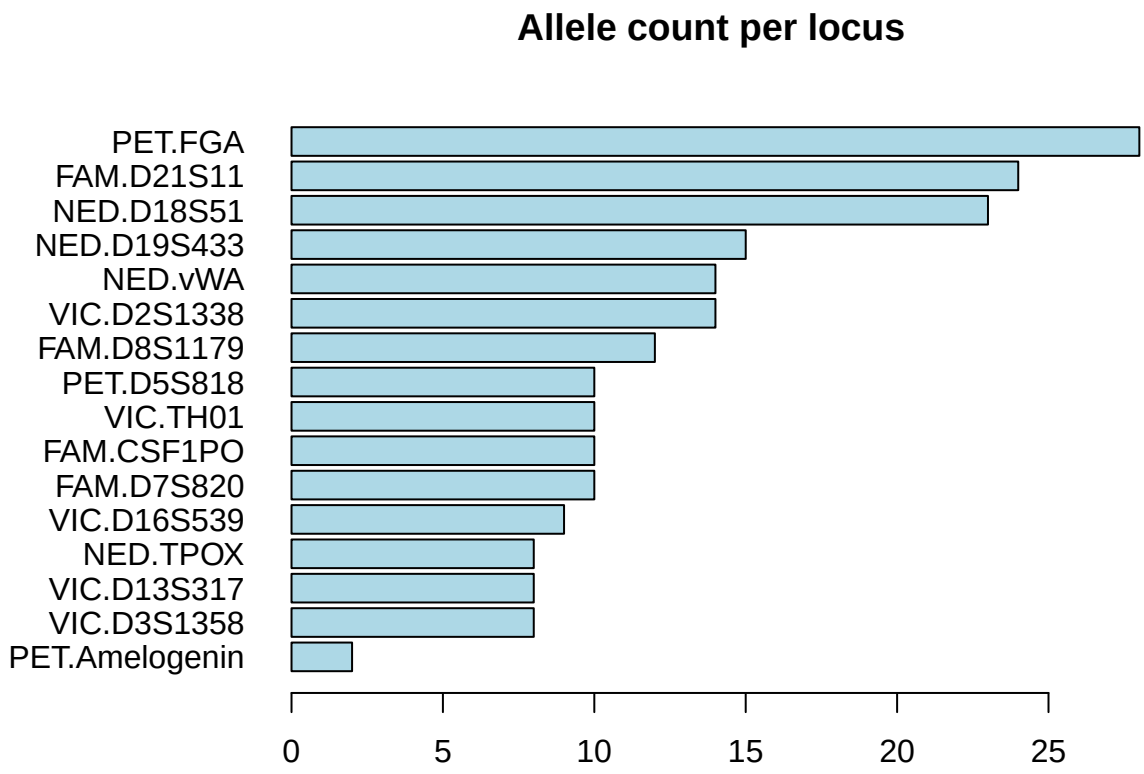
Answer 2:

The simple solution is just to use

```
df <- data(identifiler)
```

Could also make a histogram of alleles per locus (an example found in the documentation of `seqinr`)

```
op <- par(no.readonly = TRUE) # Used to reset settings later.
par(mar = c(3,8,4,2)+0.1)
allcount <- unlist(lapply(identifiler, function(x) lapply(x, length)))
barplot(allcount[order(allcount)], horiz = TRUE, las = 1,
main = "Allele count per locus", col = "lightblue")
```




```
par(op) # Reset the changed margin.
```

Question 3:

We have the Uniprot code of a human transcription factor: Q01196. We must (a) identify the name of the protein; and (b) find molecular pathways where this protein is participating in both KEGG and REACTOME and compare them.

Answer 3:

(a) identify the name of the protein.

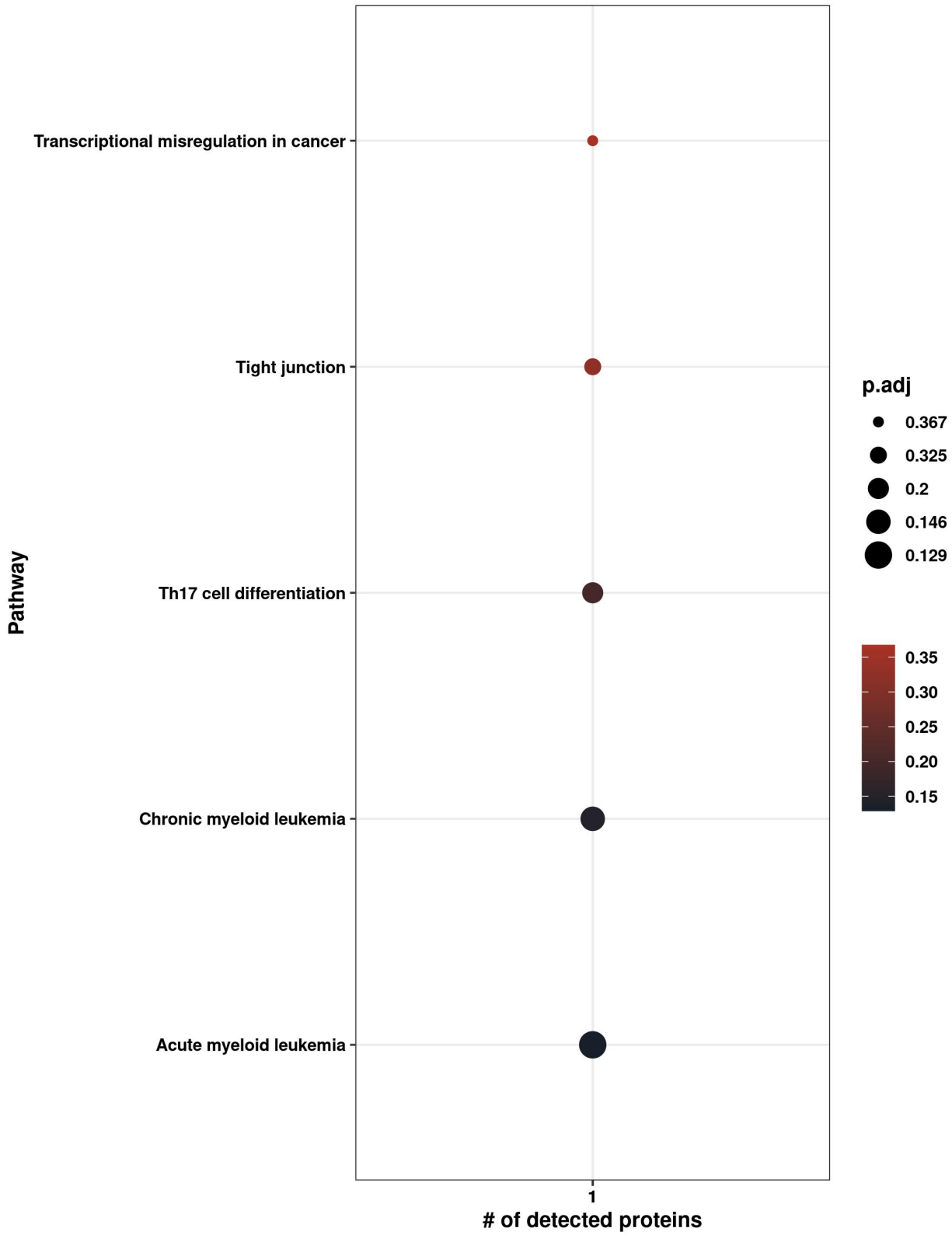
```
#Let's convert the uniprot code to the name of the protein:
prot<-c("Q01196")
ConvertID(prot, ID_from = "ACC+ID" , ID_to = "GENEWIKI_ID", directorypath = NULL)
```

The protein is named RUNX1.

(b) find molecular pathways where this protein is participating, in both KEGG and REACTOME and compare them;

First, we will search in KEGG.

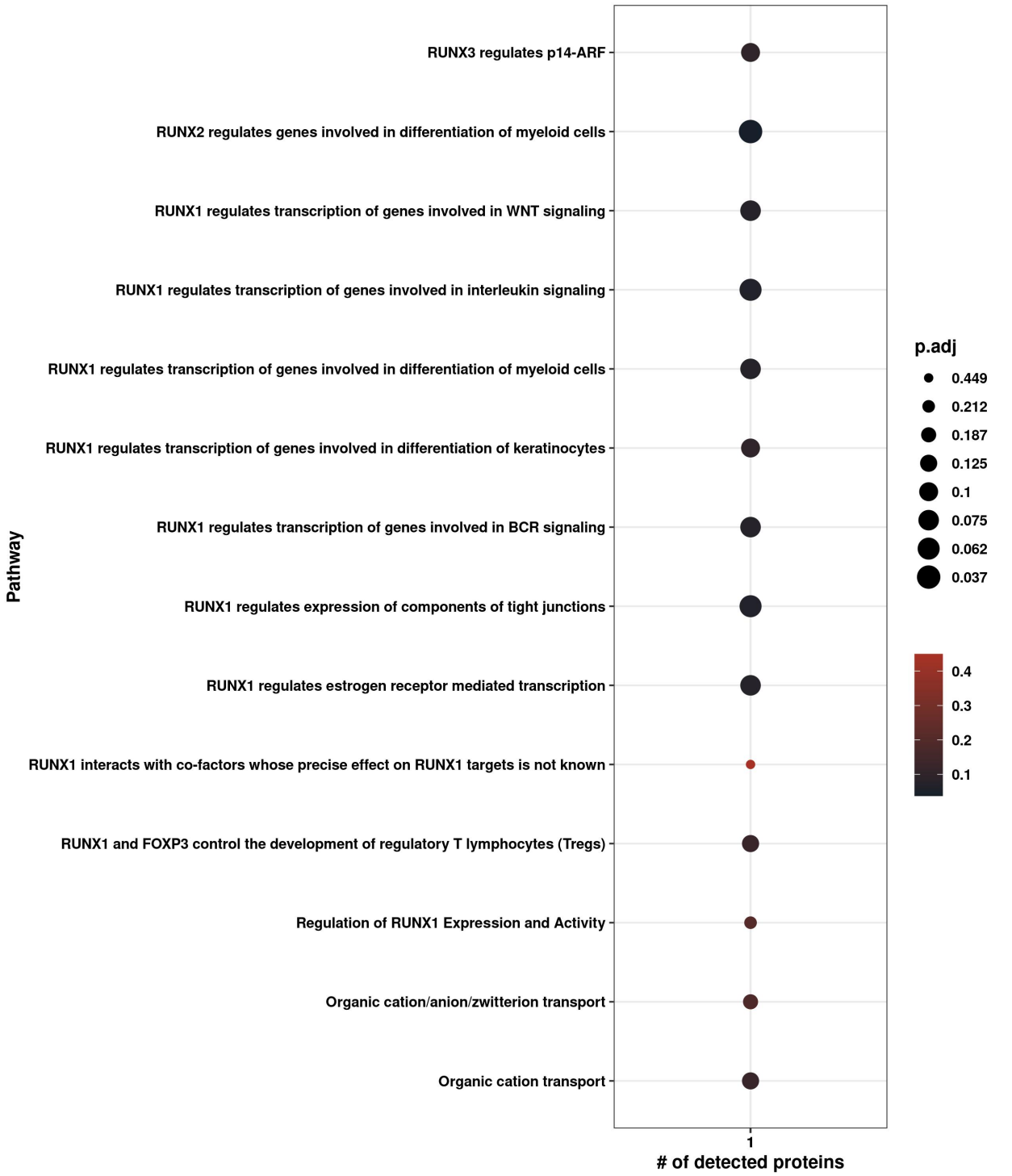
```
prot<-c("Q01196")
#First, we look in KEGG database. We set the p-value at 0.75, to obtain a higher number of processes.
Enrichment.KEGG(prot,OS="hsapiens",p_value=0.75,directorypath=".") # Saving the image and displaying in
```



RUNX1 seems to be participating in the following molecular pathways: transcriptional misregulation in cancer, tight junction, th17 cell differentiation, and specially in both chronic and acute myeloid leukemia.

Now, we will search in REACTOME.

```
#First, we look in KEGG database. We set the p-value at 0.75, to obtain a higher number of processes.
Enrichment.REAC(prot,OS="hsapiens",p_value=0.75,directorypath=".") # Saving the image and displaying in
```



It can be observed that REACTOME includes much more RUNX1-related pathways than KEGG database. Pathways included in REACTOME appear to be more specific and give us more information about the molecular process. RUNX1 is particularly contributing in cell differentation-related pathways, in some

signallings, but also in pathways associated to inflammation. KEGG pathways are general cellular processes, even whole diseases (AML).