Project Bioinformatic Resources

Annalisa Xamin, Nicola Perotti

Introduction

The following analysis focus on RNA-seq count data extracted from different cancer datasets from the Cancer Genome Atlas (TCGA). From the original TCGA data 50 cases (tumor samples) and 50 controls (normal samples) were randomly selected.

Analysis

The packages needed for the analysis are the following:

```r
library(sessioninfo)
library(tidyverse)
library(biomaRt)
library(MotifDb)  # large collection of motifs across different species
library(seqLogo)
library(PWMEnrich)
library(PWMEnrich.Hsapiens.background)
library(GEOquery)
library(oligo)
library(pd.hg.u133.plus.2)
library(hgu133plus2.db)
library(genefilter)
library(limma)
library(pheatmap)
library(stringr)
library(GenomicFeatures)  # to build object with specific information such as genomic coordinates
library(ggplot2)
library(edgeR)  # designed to perform Differential Expression Analysis from RNA-Seq data
library(fgsea)  # computation of enrichment scores and other statistics
library(org.Hs.eg.db)  # database annotation human specific transcript from Ensembl
library(clusterProfiler)  # uses Fisher test, explore gene list of interest against a reference
library(enrichplot)  # build enrich map
library(ggnewscale)
library(DOSE)
library(pathview)  # contains cartoons of pathways and we project on them our genes of interest
```

For reproducibility, we include the R session information.

```r
print("Reproducibility information:")
Sys.time()
proc.time()
```

```
options(width = 120)
session_info()
```

Load data

In particular, we consider data coming from **Lung adenocarcinoma**.

```
load("./RData/Lung_adenocarcinoma.RData")
```

After loading the data, the following three data-frames are available:

- `raw_counts_df` which contains the raw RNA-seq counts;
- `c_anno_df`, which contains sample name and condition (case and control);
- `r_anno_df`, which contains the ENSEMBL genes ids, the length of the genes and the genes symbols.

Filter protein coding genes

To extract only protein coding genes from `raw_counts_df` and `r_anno_df`, we use the `biomaRt` package.

```
query <- getBM(attributes = c("ensembl_gene_id", "external_gene_name", "gene_biotype"),
    filters = c("ensembl_gene_id"), values = r_anno_df$ensembl_gene_id, mart = ensembl)

query_protein_coding <- query[which(query$gene_biotype == "protein_coding"), ]
```