LAB#6 – Applied Omics

The goal of this laboratory is to analyse a clinically annotated breast cancer transcriptomics dataset from The Cancer Genome Atlas (TCGA), namely to estimate gene expression in a cohort of tumours and some matched normal samples and use that information to find molecular features potentially relevant for understanding the biology of the disease and for its clinical management.

The data needed for this laboratory can be downloaded from http://imm.medicina.ulisboa.pt/group/distrans/SharedFiles/CompBioISTLab6.zip. After decompressing the downloaded ZIP bundle, you will find the following files:

- `TCGA-C8-A138-01_1.fastq` and `TCGA-C8-A138-01_2.fastq`: FASTQ files with the output of high-throughput paired-end sequencing of mRNA from TCGA breast tumour sample TCGA-C8-A138-01.
- `TCGA_BRCA_Gene_ReadCounts.txt`: tab-delimited text file with the number of RNA-seq reads mapping to each gene (row) in each tissue sample (column) in the TCGA breast invasive carcinoma (BRCA) cohort. More on how those read counts are obtained can be found here: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/. Please note that samples are named according to TCGA barcode identifiers (https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/), so make sure you are able to identify the patient (study participant) and the sample type (https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes). Note that there may samples of different types biopsied from the same patient.
- `TCGA_BRCA_ClinicalAnnotation.txt`: tab-delimited text file with personal, clinical and molecular information (columns) about patients (rows) and their primary tumours. You can find information on breast cancer staging here: https://www.cancer.net/cancer-types/breast-cancer/stages. Information on molecular markers and subtypes is discussed below.

Please note that, for several questions below, there can be many alternative good answers. Grading of this laboratory will therefore have a special focus on the **rationale** guiding your decisions, so make sure you explain it clearly and illustrate well (namely with eloquent plots, when appropriate) the supporting statements. Nevertheless, please be succinct, as the reports are limited to **10 pages**.

Group Ø (2 points)

R Quiz (done on November 20th).

Group I (4 points)

a) Provide your brief quality assessment of the raw sequencing data in the FASTQ file. Any problem one should be aware of?

b) Use your favourite aligner (e.g. Kallisto, Salmon, STAR, RSEM, Whippet) and annotated human transcriptome (explicitly state where you got it from) to estimate gene expression in that sample. How do your estimates compare with those in the provided read count table for the same sample? If you find differences, explain what their sources may be.

(Aligning may take several minutes, so it is suggested that you start working on the next questions while you wait for it to be completed.)

Group II (8 points)

a) Based on the provided read count table, summarise the distributions of read coverage and library complexity across TCGA samples. Any sample(s) raising concerns?

b) Use your favourite data transformation and normalisation method to make gene expression profiles comparable between samples. Justify your choice and show that it worked. Any sample(s) raising concerns?

c) Which phenotypic traits dominate data variance? Which genes are associated with the main axes of variance? Did you spot any non-biological batch effect worth acting on?

d) What are the main differences in expressed genes and activated pathways between primary tumours and normal breast samples? How does the age of patients affect those differences?

Group III (6 points)

Breast cancer biopsies are tested, usually by immunohistochemistry, for the presence of proteins that are estrogen and progesterone receptors, encoded by genes *ESR1* and *ESR2*, and *PGR*, respectively. The hormone receptor status is key in deciding treatment options. Similarly, breast tumours are tested for a protein called human epidermal growth factor receptor 2 (HER2), encoded by gene *ERBB2*, which promotes the growth of cancer cells. HER2-positive breast cancer is treated differently. You can find the results of such tests for the TCGA primary tumour samples in the provided patient annotation table.

a) How good is the cognate genes' mRNA expression at recapitulating the binary classifications that result from those immunohistochemistry-based tests?

PAM50 is a 50-gene signature that classifies breast cancer into five molecular intrinsic subtypes: Luminal A, Luminal B, HER2-enriched, Basal-like and Normal-like (v. corresponding column in the provided patient annotation table). This classification is also used in the clinic to support therapeutic decisions.

b) Based on an overall survival analysis, how would you rank those five subtypes in terms of prognosis?

c) Based on gene expression, find the gene signature that best classifies the molecular subtypes, explicitly assessing its performance. How many of the genes in your signature are part of the PAM50 gene list[1]? How does the classifying performance of your gene signature compare to that of one involving the PAM50 genes?

[1] *UBE2T, BIRC5, NUF2, CDC6, CCNB1, TYMS, MYBL2, CEP55, MELK, NDC80, RRM2, UBE2C, CENPF, PTTG1, EXO1, ORC6L, ANLN, CCNE1, CDC20, MKI67, KIF2C, ACTR3B, MYC, EGFR, KRT5, PHGDH, CDH3, MIA, KRT17, FOXC1, SFRP1, KRT14, ESR1, SLC39A6, BAG1, MAPT, PGR, CXXC5, MLPH, BCL2, MDM2, NAT1, FOXA1, BLVRA, MMP11, GPR160, FGFR4, GRB7, TMEM45B, ERBB2*