Annotated RNA-seq metadata (aligned RNA-seq data from various sources, mostly SRA and GEO, consisting of **238522 RNA-seq samples** aligned towards **178136 transcripts**) were downloaded from the ARCHS4 database[1].

From all the samples, a subset of samples containing 'AML' in 'Sample_source_name_ch1' field was selected - **678 samples**. Each sample is a vector of numbers of reads (as-is, not normalized) against transcript names (all 178136 transcripts). Only 100 transcripts with the highest read coverage per sample were chosen, and a set of **unique 3186 transcripts** was found, accounting for all the top-100 most abundantly expressed transcript isoforms from all the chosen samples. Then respective gene names were found using GENCODE_v34_hg38_comprehensive list of genes and transcripts (downloaded from the UCSC table browser), constituting **2198 unique genes** (let's call it a **MA-AML set,** most abundant - AML). While several transcript isoforms might belong to one gene, we're only focusing on unique gene list for now. The lists of both unique transcripts and genes are available in the repo (./lists/most_common_transcripts_AML, ./lists/most_common_genes_AML).

To clarify on the clinical importance of the most abundantly expressed genes in the cancer of interest (AML), the intersections between the most abundantly expressed genes and clinically significant genes in AML were found:

1) a list of **23 genes** that were shown to exhibit the highest values of **prognostic significance** with respect to **somatic alterations** in AML[2] (derived from ClinSEQ cohort) was downloaded. The list is included in the repo (./lists/clinSEQ_AML_23). Only **17 out of 23 genes were found** in the MA-AML set;
2) a list of **716 genes** with significant **differential splicing events** between the **FAV and ADV TCGA-AML** cohort patients was retrieved[3]. The list is included in the repo (./lists/TCGA-AML_AS_716). Only **186 out of 716 genes were found** in the MA-AML set;
3) a list of **222 genes** constituting **commonly differentially spliced genes between the TCGA and Clinseq AML cohorts** was retrieved[3]. The list is included in the repo (./lists/TCGA-AML_clinSEQ_AS_common_222). Only **80 out of 222 genes were found** in the MA-AML set.

The numbers of these intersections suggest that the most expressed transcript isoforms might not be of prognostic relevance to AML, and we need to curate the set of genes manually.

The respective jupyter notebook can be found in the repo (get_expr. ipynb).

[1] A. Lachmann *et al.*, "Massive mining of publicly available RNA-seq data from human and mouse," *Nat. Commun.*, 2018, doi: 10.1038/s41467-018-03751-6.

[2] M. Wang *et al.*, "Validation of risk stratification models in acute myeloid leukemia using sequencing-based molecular profiling," *Leukemia*, 2017, doi: 10.1038/leu.2017.48.

[3] G. Anande *et al.*, "RNA Splicing Alterations Induce a Cellular Stress Response Associated with Poor Prognosis in Acute Myeloid Leukemia," *Clin. Cancer Res.*, 2020, doi: 10.1158/1078-0432.ccr-20-0184.