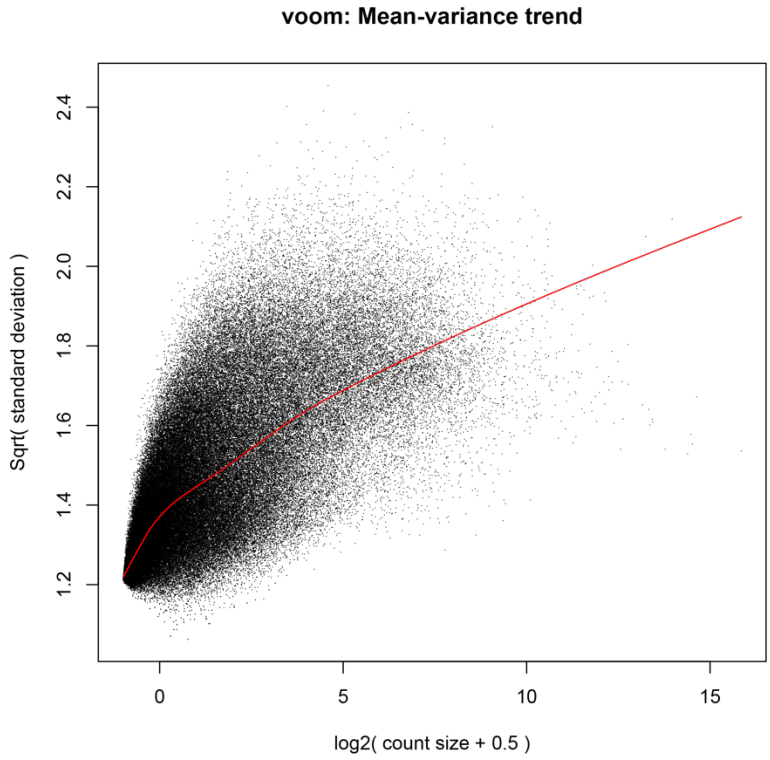
Annotated RNA-seq metadata (aligned RNA-seq data from various sources, mostly SRA and GEO, consisting of 238522 RNA-seq samples aligned towards 178136 transcripts) were downloaded from the ARCHS4 database[1]. Each sample row contains raw RNA-seq read counts (as-is, not normalised)

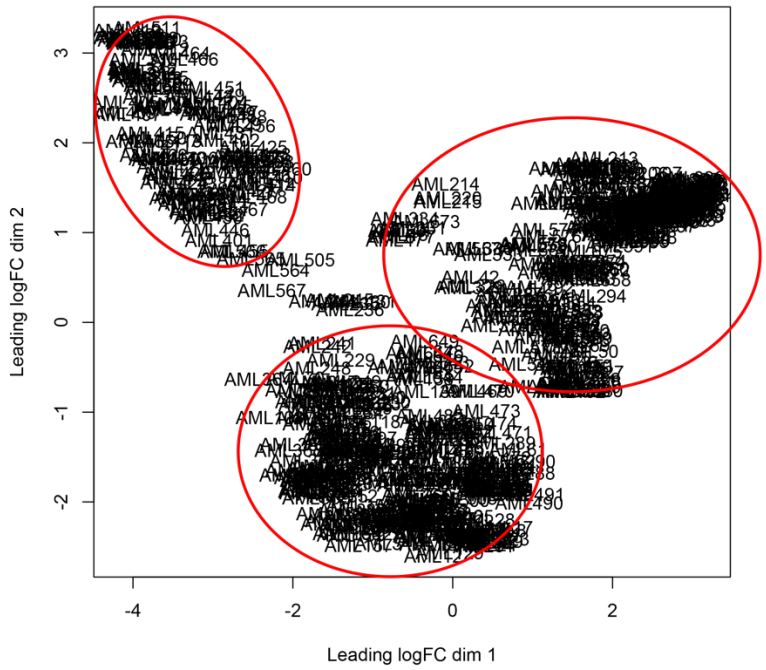
From all the samples, a subset of samples containing 'AML' in '`Sample_source_name_ch1`' field was selected (678 samples), and 1000 samples were randomly drawn for a baseline.

Differentially expressed transcripts

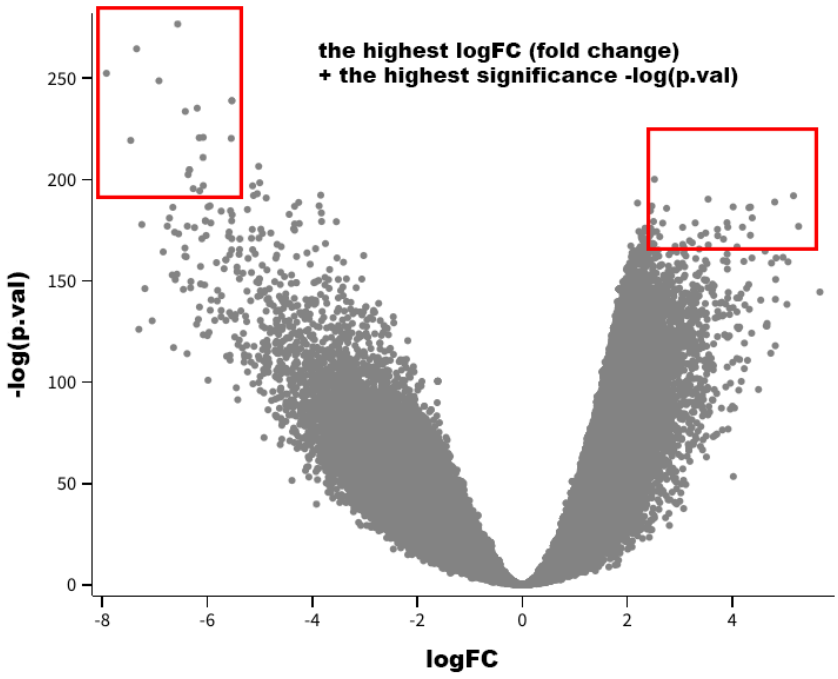
Differentially expressed transcripts were obtained using limma-voom R packages for RNA-seq data analysis. voom estimates the mean-variance relationship of the RNA-seq counts and generates a weight for each observation; limma uses empirical Bayes algorithm to assess the most statistically significant differentially expressed genes. The gene count number - variance relationship for the AML dataset is shown below. The general trend is the higher the count number → the higher the variance, expectedly.



A simple clustering analysis (based on the distance between samples) shows that the samples seem to have three distinct clusters. This might be investigated further. The clustering graph is shown below.



For the differential expression analysis, a contrast matrix was created based on the difference in expression in two datasets (AML/rand), where the variance and other effects have been accounted for with empirical Bayes algorithm (see DE_analysis.r). As a result, a fold of change in expression and p-values for statistical significance are calculated. A volcano graph for visualization of the result is shown below. We're interested in the top left and right quarters of the graph below: the highest fold of change (both positive and negative) and the lowest possible p-val (or highest -10 log(p-val)).



Therefore, 50 top-DE (differentially expressed) transcripts between AML-rand samples were found and saved into the DE_AML_transcripts.fa file: exonic sequences and the respective identifiers -- transcript names.

R script with the analysis pipeline is saved in the repository as DE_analysis.r.

Then, the DE transcripts from the file DE_AML_transcripts.fa are imported with the jupyter notebook auxiliary_inputs.ipynb. The transcript counts are imported from the archs4 matrix anew (can improve – no need to import anew, can save into a sep file; plus random samples are different from the random samples used for the analysis described above, therefore, the differentially expressed genes might be different). Then the counts are normalised to get counts per million (see the jupyter notebook), and further normalised by the transcript length. Then a table with normalised counts for each AML/random sample is created for the N differentially expressed transcripts of interest (in this case, 50, but one of the transcript annotations was not found in the hg38, so 49).

The table can now be used for the auxiliary NN input: transcript sequences can be obtained from a separate dictionary by the transcript name as a key, and expression levels can be obtained from the matrix.

Example of the table output:

-	ENST00000477988.1	ENST00000561385.5	ENST00000284509.10	ENST00000262262.4
AML1	0.0	4.693	96.64	44.663
AML2	0.0	20.249	252.841	4.7
AML3	0.0	6.1	142.217	43.449
AML4	0.0	97.211	238.677	77.985

Respective limma-voom tutorial: https://ucdavis-bioinformatics-training.github.io/2018-June-RNA-Seq-Workshop/thursday/DE.html