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).

Most expressed transcripts and their clinical significance

Each sample is a vector of numbers of reads (as-is, not normalized) against transcript names (all 178136 transcripts). Reads per transcript were averaged for all the AML samples, and 1000 transcripts with the highest read coverage were chosen. Respective gene names were found using GENCODE_v34_hg38_comprehensive list of genes and transcripts (downloaded from the UCSC table browser). The table with 1000 most abundant transcripts and genes for the AML subset is available in the repo (./lists/ma_AML). 30 most covered transcripts and the respective genes are shown in the table below, for the full table see either the ma_AML table or the output of the respective jupyter notebook (ma_AML.ipynb).

Transcript name	Gene name	Coverage	Transcript name	Gene name	Coverage
ENST00000340368.8	INSIG1	404122	ENST00000225174.7	PPIF	61213
ENST00000263239.6	DDX18	291257	ENST00000558506.1	EIF5	60122
ENST00000361050.3	MPEG1	287857	ENST00000367103.3	MAPKAPK2	58062
ENST00000438806.5	PNISR	153418	ENST00000260810.9	TOPBP1	57533
ENST00000611208.4	ARL8B	116656	ENST00000569240.5	CNOT1	56165
ENST00000357304.8	PRRC2B	111875	ENST00000418623.1	not found	55594
ENST00000313368.7	TAF7	105589	ENST00000339594.8	BAZ1B	54218
ENST00000264161.8	DARS1	99752	ENST00000231368.9	LNPEP	49512
ENST00000526562.5	RPL27A	98084	ENST00000502553.5	TARS1	48657
ENST00000549690.1	LYZ	80331	ENST00000263805.8	ZNF106	45004
ENST00000392168.6	TCP1	78758	ENST00000361565.8	IPO9	44897
ENST00000535572.5	WNK1	78667	ENST00000233997.3	AZU1	44526
ENST00000393263.7	NAP1L1	78322	ENST00000620695.2	AZU1	43802
ENST00000426431.2	SP1	64713	ENST00000541679.7	RAN	43784
ENST00000369851.5	GNAI3	63108	ENST00000246006.4	CD93	42654

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). 12 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). 90 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). 50 out of 222 genes were found in the MA-AML set.

These intersections suggest that the most expressed transcript isoforms might not be of prognostic relevance to AML. For a comparison, if we select 1000 most abundantly expressed transcripts from 1000 randomly drawn RNA-seq samples of the same dataset, the number of intersections for prognostic lists above will be 5, 70, 37 respectively.

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

Differentially expressed transcripts and their clinical significance

Perhaps a set of the most differentially expressed genes might give a better intersection, so the "baseline" of expression was found by averaging expression from 1000 randomly drawn samples of the whole dataset, and the baseline was subtracted from the averaged AML subset expression. 1000 transcripts with the highest and the lowest result (positive and negative differential expression, respectively) were chosen. Respective gene names were found using `GENCODE_v34_hg38_comprehensive` list of genes and transcripts (downloaded from the UCSC table browser). The table with 2000 most differentially expressed transcripts and genes for the AML-rand subset is available in the repo (`./lists/diff_AML_rand`). Some 30 transcripts with the highest difference are shown below, for the full table see either the `diff_AML_rand` table or the output of the respective jupyter notebook (`diff_AML.ipynb`).

Transcript name	Gene name	Coverage	Transcript name	Gene name	Coverage
ENST00000357103.4	ADIPOR2	+251040	ENST00000251595.10	HBA2	-38068
ENST00000375048.7	DDOST	+201941	ENST00000320868.9	HBA1	-27626
ENST00000305885.2	FEN1	+123128	ENST00000307407.7	CXCL8	-20454
ENST00000535572.5	WNK1	+71733	ENST00000225964.9	COL1A1	-17604
ENST00000336095.10	RNF24	+60039	ENST00000335295.4	HBB	-16080
ENST00000394668.2	RPL34	+57898	ENST00000446046.5	FN1	-14698
ENST00000341068.7	ANAPC1	+53574	ENST00000443816.5	FN1	-14542
ENST00000454306.6	PRRC2A	+50879	ENST00000445125.2	not found	-13772
ENST00000398571.6	USP34	+47664	ENST00000390237.2	IGKC	-8567
ENST00000583389.5	EIF4A1	+45860	ENST00000620041.4	FTH1	-7987
ENST00000339594.8	BAZ1B	+43204	ENST00000260356.5	THBS1	-7887
ENST00000411531.5	EIF4G1	+40114	ENST00000304636.7	COL3A1	-7872
ENST00000378962.3	CXorf21	+36287	ENST00000361789.2	MT-CYB	-7666
ENST00000547798.1	TMBIM6	+32184	ENST00000297268.10	COL1A2	-7151
ENST00000356936.5	NCL	+31885	ENST00000636712.1	SERPINA1	-6939

The intersection of the diff expressed genes with "clinically prognostic" lists is higher (14/23, 115/716, 56/222 for the lists mentioned above).

But some genes might be expressed differentially in this particular type of cells (bone marrow hematopoietic, peripheral blood cells) in general compared to random cell type, therefore, as a baseline we might use an averaged expression profile of the respective healthy cells. Therefore, a set of samples containing "peripheral blood", "marrow" and "hematopoietic" in their name was found (13007 samples), and a differential expression profile using this baseline was calculated. The table with 2000 most differentially expressed transcripts and genes for the AML-blood subset is available in the repo (`./lists/diff_AML_blood`). Some 30 transcripts with the highest difference below, for the full table see either the `diff_AML_blood` table or the output of the respective jupyter notebook (`diff_AML.ipynb`).

Transcript name	Gene name	Coverage	Transcript name	Gene name	Coverage
ENST00000264028.4	ARCN1	+335176	ENST00000335295.4	HBB	-45423
ENST00000450554.6	U2AF2	+211583	ENST00000251595.10	HBA2	-33854
ENST00000199389.10	EIF2AK1	+174646	ENST00000320868.9	HBA1	-33306
ENST00000257829.7	NAT10	+130390	ENST00000307407.7	CXCL8	-17145
ENST00000435706.6	TOP2B	+80163	ENST00000390237.2	IGKC	-16603
ENST00000265896.9	SQLE	+72381	ENST00000546260.5	SOD2	-15389
ENST00000301785.5	HNRNPUL2	+68104	ENST00000368738.3	S100A9	-14285
ENST00000349736.9	CSNK2A1	+62803	ENST00000394936.7	PSAP	-10505
ENST00000263238.6	ACTR3	+59760	ENST00000263341.6	IL1B	-8184
ENST00000393043.5	CLTC	+58928	ENST00000582401.5	TXNIP	-7669
ENST00000565644.5	SLC7A5	+58018	ENST00000368733.3	S100A8	-7569
ENST00000534548.6	NCAPD3	+56548	ENST00000382692.2	DEFA1	-7184
ENST00000597451.5	UBA52	+51579	ENST00000631466.1	IGHG1	-6357
ENST00000356244.7	RANGAP1	+43433	ENST00000390549.6	IGHG1	-6357
ENST00000295628.3	LRRC58	+42106	ENST00000390323.2	IGLC2	-6299

The intersection of the diff expressed genes with “blood” baseline with “clinically prognostic” lists is approx. the same (14/23, 118/716, 55/222 for the lists mentioned above).

The respective jupyter notebook can be found in the repo (`diff_AML.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.