Meta information

Data: Cancer gene expression RNA-Seq dataset
From: https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq
Source: Samuele Fiorini, samuele.fiorini '@' dibris.unige.it, University of Genoa, redistributed under Creative Commons license (http://creativecommons.org/licenses/by/3.0/legalcode) from https://www.synapse.org/#!Synapse:syn4301332.

Method: bulk RNA sequencing
Number of samples: 801
Number of numerical features: 20531
Number of categorical features: 1

Categories are based on cancer type: BRCA – Breast invasive carcinoma; COAD – Colon adenocarcinoma; KIRC - Kidney renal clear cell carcinoma; PRAD – Prostate adenocarcinoma; LUAD – Lung adenocarcinoma.

Introduction

A cancer gene expression dataset acquired by bulk RNA-seq was used for this project (ref. above). It's composed of 801 different cancer samples, which are grouped into 5 groups: BRCA – Breast invasive carcinoma; COAD – Colon adenocarcinoma; KIRC - Kidney renal clear cell carcinoma; PRAD – Prostate adenocarcinoma; LUAD – Lung adenocarcinoma. Cancer types are stored in a different dataset and were added during data analysis to a single dataframe. The dataset has 20531 attribute which correspond to different genes. Gene expression was measured using Illumina HiSeq sequencing. The expression values are expressed as in log2(1+expression). Originally, the dummy gene names (Gene_1; Gene_2…) were given in the dataset. However, the repository has a link to an original redistribution directory. In there, dataset with original gene names is located. During analysis gene names were imported from one of the datasets. The source of analyzed data stated, that no rearrangement of data was performed and dummy names correspond to real gene names in the same position.

This dataset was chosen to gain experience in RNA-seq data analysis and there were no prior biases on which dataset to choose specifically. The data has some missing gene names represented as "?". This should not pose any troubles unless this type of genes comes up as differently expressed one. Additionally, some of the genes have zero expression. This is also normal, as not all genes in the cell are expressed. However, it's known that RNA-seq methods are zero inflated, meaning not all expressed genes are captured due to biases in sequencing methods, RNA capture type and enzymes used in the reaction. It is important to note, that Bulk RNA-seq methods have a higher capture rate, therefore data is not so scare as in single-cell methods. In single-cell RNA-seq dataset gene expression can be imputted based on cell localization in gene expression. Closely localized cells can be assumed identical. In bulk RNA-seq methods, imputing is not advised as averaging of cellular population to a single representative sample has a higher variation due to analyzed tissue size and acquisition techniques. Some samples might have cell which other samples do not. In this case, cancer samples might have immune cells in the population, which might skew the gene expression.

Table 1. Representation of analyzed dataframe.
Data is ordered in an 801x20531 matrix. At the end there is a column representing cancer type.

	gene_0	gene_1	gene_2	gene_3	gene_4	gene_5	gene_6	gene_7	gene_8	gene_9	...	gene_20529	gene_20530	Cancer Type
sample_0	0.0	2.017209	3.265527	5.478487	10.431999	0.0	7.175175	0.591871	0.0	0.000000	...	5.286759	0.0	PRAD
sample_1	0.0	0.592732	1.588421	7.586157	9.623011	0.0	6.816049	0.000000	0.0	0.000000	...	2.094168	0.0	LUAD
sample_2	0.0	3.511759	4.327199	6.881787	9.870730	0.0	6.972130	0.452595	0.0	0.000000	...	1.683023	0.0	PRAD
sample_3	0.0	3.663618	4.507649	6.659068	10.196184	0.0	7.843375	0.434882	0.0	0.000000	...	3.292001	0.0	PRAD
sample_4	0.0	2.655741	2.821547	6.539454	9.738265	0.0	6.566967	0.360982	0.0	0.000000	...	5.110372	0.0	BRCA
sample_5	0.0	3.467853	3.581918	6.620243	9.706829	0.0	7.758510	0.000000	0.0	0.000000	...	5.355133	0.0	PRAD
sample_6	0.0	1.224966	1.691177	6.572007	9.640511	0.0	6.754888	0.531868	0.0	0.000000	...	8.330912	0.0	KIRC
sample_7	0.0	2.854853	1.750478	7.226720	9.758691	0.0	5.952103	0.000000	0.0	0.000000	...	6.551490	0.0	PRAD
sample_8	0.0	3.992125	2.772730	6.546692	10.488252	0.0	7.690222	0.352307	0.0	4.067604	...	7.828321	0.0	BRCA
sample_9	0.0	3.642494	4.423558	6.849511	9.464466	0.0	7.947216	0.724214	0.0	0.000000	...	4.759151	0.0	PRAD

Biases of the data might arise during sequencing. It's not known if the samples were prepared at the same time and were they sequenced on the same run. Processing the samples at different time points might introduce technical noise into the data-set. Due to it, gene that is not differently expressed might come up as one. However, a large number of samples for a single cancer type should reduce the noise.

Descriptive data analysis

The data set was loaded as pandas dataframe with Cancer Type labels at the end as showed in table 1.

First, the sum count number for each sample was analyzed to assess the coverage of the samples (**Fig. 1 A**) (counts represent uniquely mapping reads). As seen from the histogram, the count number per samples has a mean of 132,287.85, with. The count number

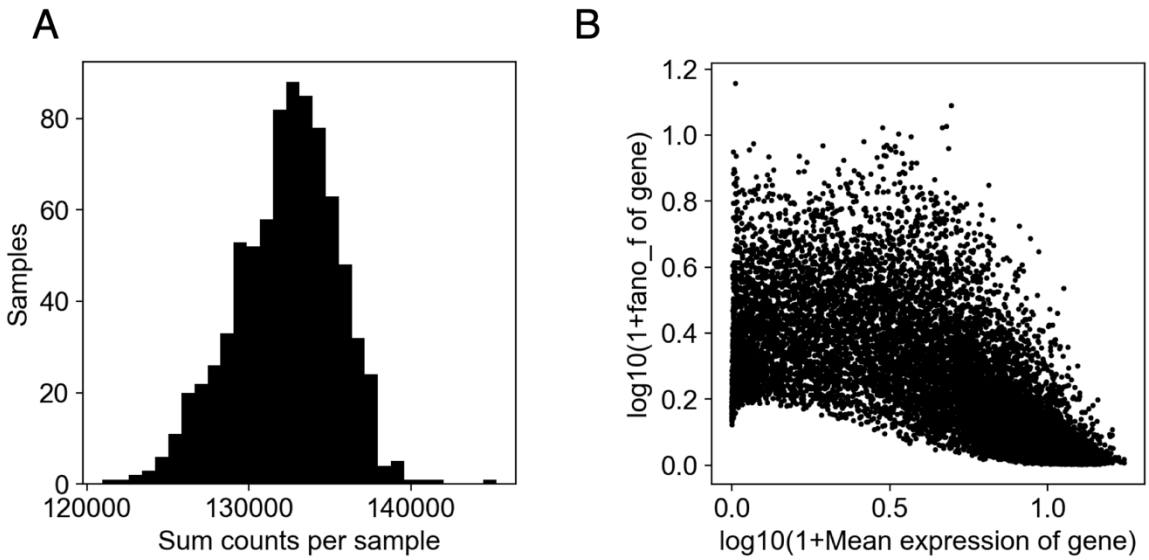

Figure 2. Gene count and expression descriptive analysis
A. Distribution of sum count number per sample represent a normal distribution (p-value = 0.0004) with a mean centered at 132,287.85 counts per sample with a small variance (SD = 3,183.43). **B**. Mean gene expression plotted against the variance of that gene expressed as fano-factor. Units are represented in log10(1+value).

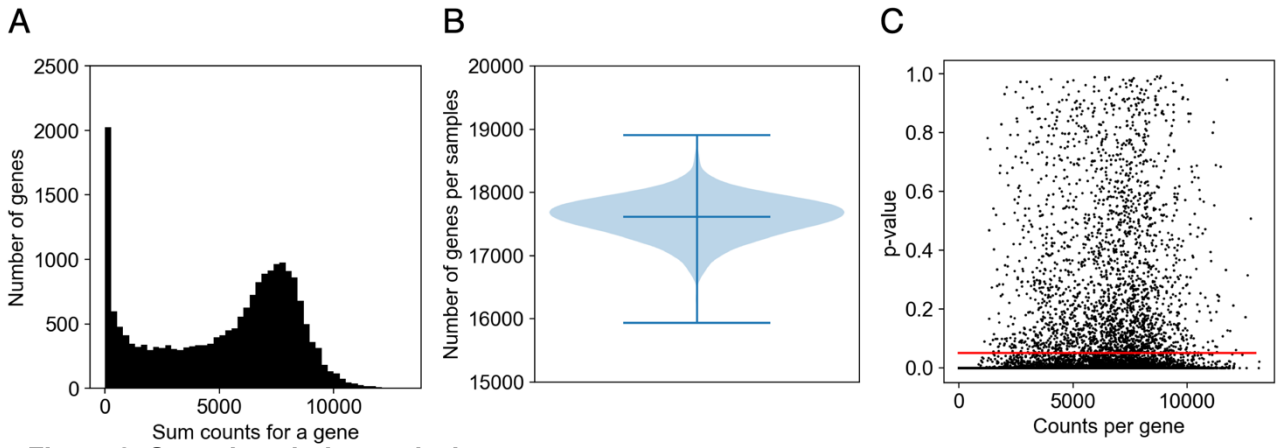

Figure 2. Gene descriptive analysis

A. Distribution of sum count number per a gene with majority of genes having 7500. Additionally, ~2000 genes have an expression of zero. **B**. Violin plot displaying the number of genes detected per sample. Majority of the samples have around 17,800 genes detected and the distribution of genes detected is close to normal. **C**. Scatter plot displaying normality test p-values against counts per genes. 88.7% of genes are have p-value <0.05.

per samples has a normal distribution (p-value = 0.0004). Samples have a similar sequencing depth as the variance of count number is relatively small (standard deviation is 3183.43). However, for downstream analysis, it it's advised to normalized the data to an equal count number per samples to avoid biases arising from different sequencing depth. The data set has a lot of variables (20531), therefore a simple descriptive table of it is hard to understand. To avoid this, mean gene expression and variance distribution was analyzed in scatter plot showed in **Fig. 1 B**. As seen from the data, the variance of gene expression reduces as the mean expression value increases. This is a typical thing seen in gene expression analysis, as signal noise reduced with it' intensity. Except for the highly expressed genes, majority of them have a variance of 0.2 – 0.6 (expressed as log10(1+fanfo factor). A fraction of genes is highly variable and have variance of 0.6-1.0.

Next, an investigation of sum counts per a gene showed majority of genes have a sum count around 7500 (**Fig. 2A**). Additionally, ~2000 genes have an expression of zero. This might arise due to technical aspects of RNA-seq or low expression of those genes. There is no way to verify, however bulk RNA-seq samples have a higher capture rate than single cell methods. As a result, this shouldn't pose any problems. All samples have around 17,800 genes detected, and the overall coverage is normally distributed (**Fig 2B**). This points out that there were no biases for different cancer types and there shouldn't be any technical biases arising due to unsuccessfully detected genes sequencing. To evaluate each gene separately, it's expression distribution in all the samples was evaluated and comparted to normal distribution. **Fig. 3C** shows a scatter plot of each gene comparison to normal distribution p-value against its count number. 88.7% of genes are distributed normally (p-value < 0.05) among the samples.

Overall, descriptive analysis shows that the dataset is high quality and there are no abnormalities in detected gene number and their expression. All the analyzed samples similar distribution and thus can be compared to each other.

Gene expression correlation analysis

After the evaluation of technical variation in gene detection and count distribution, gene expression program correlation in each sample was analyzed. It's known that samples, belonging to the same cancer type should have a high gene expression correlation.

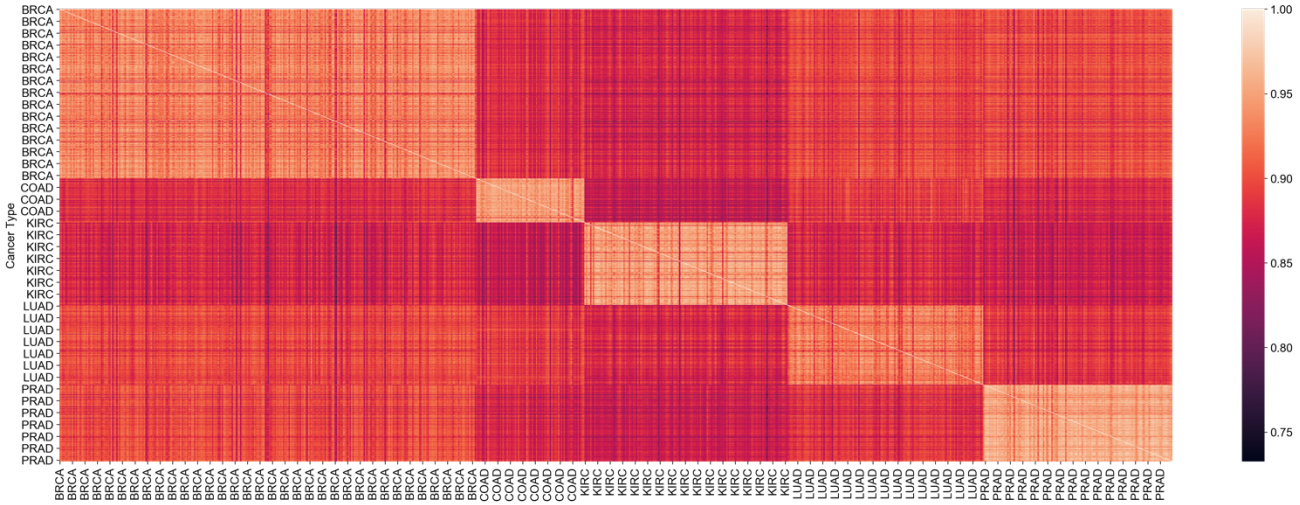

Figure 3. Gene expression correlation heatmap
Heatmap represent gene expression correlation matrix coefficients for each sample in the dataframe. Sample were sorted based on cancer type; each sample's gene expression was compared to all other sample's gene expression. Color bar represents Pearson correlation coefficients. Lightly shaded clusters correspond to highly correlated samples belonging to the same cancer type

Additionally, in a successful experiment, samples from the same species belonging to different tissue should still maintain a high correlation of gene expression, as a lot of detected genes are house-keeping and have a low variance. To do this, a pair-wise Pearson correlation test was performed for gene expression among all the samples. A 801x801 correlation matrix was plotted as a heatmap of correlation coefficients (**Fig 3**)

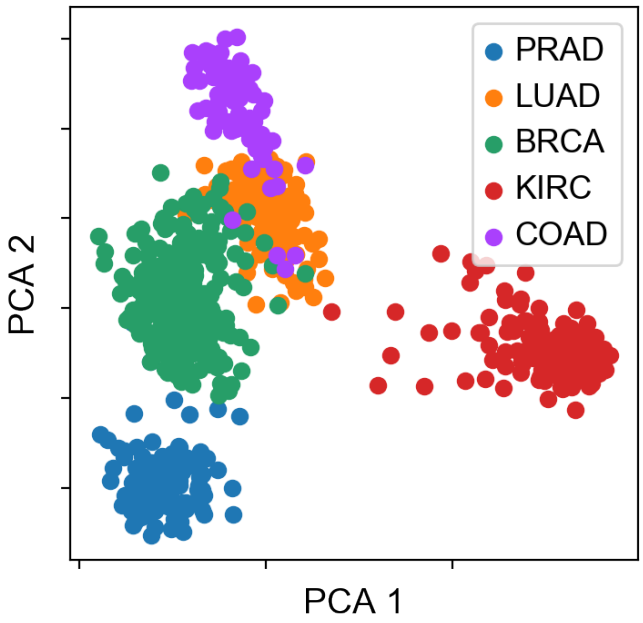

Figure 4. Principal component analysis
Visualization of all 20531 gene expression values in 2D space via two principal component analysis. Cancer type labels were removed before doing PCA and then were mapped back on to the newly projected values. **BRCA** – Breast invasive carcinoma; **COAD** – Colon adenocarcinoma; **KIRC** - Kidney renal clear cell carcinoma; **PRAD** – Prostate adenocarcinoma; **LUAD** – Lung adenocarcinoma.

As seen from the data, different cancer types form 5 distinct cluster of higher gene expression correlation. Additionally, overall correlation among all the samples is high (> 0.7). This shows that dataset is high quality, not shewed and there are no technical variations.

Principal component analysis and data clustering

To better analyze the high dimensional data-set, a principal component analysis was carried out. First of all, raw data was projected into 2D space. Additionally, cancer type labels were assigned based on the initial dataset.

As depicted in **fig. 4**, different samples form almost separate clusters in 2D gene expression space. Most separated are prostate adenocarcinoma (PRAD) and kidney renal carcinoma samples (KIRC). Breast, lung and colon cancer samples form clusters that are closer to each other. This might be due to nature of the cancer (adenocarcinomas or etc.) or due to cell-type. In conclusion, projection into 2D gene expression space cluster the data into different groups, showing that there are underlying correlating expression patterns that allow the data to be differentiated.

For further analysis, each sample was normalized to 120,000 total count per sample to make the comparison of the data more reliable. To perform clustering of the data, Principal Component Analysis is performed only on highly variable genes. To filter out non-variable genes, threshold of expression was set to [0.125 – 13]. A threshold for minimum variance was chosen empirically. It was set to -0.5; 0; 0.5; 1 and clustering in 2D PCA plots was analyzed (**Fig 6**)

Based on results, a threshold of minimum variance was chosen to be -0.5 (**Fig 5**), as these filtering parameters maintain the identity of the clusters. Increasing the variance threshold lead to merging of BRCA and PRAD samples, meaning meaningful data is lost.

Next, UMAP (has increased speed and better preservation of the data's global structure then t-SNE) and Leiden clustering were performed on the samples with neighbor number set to 50. The dataset was moved from Pandas dataframe to Scanpy's AnnData object designed to store RNA-seq gene expression matrices. The label identity was

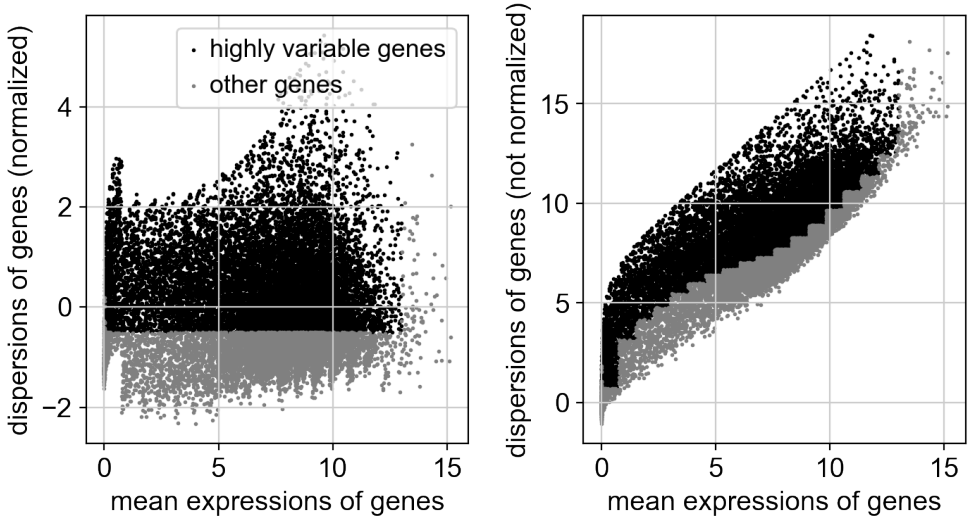

Figure 5. Identification of highly variable genes.
Highly variable genes were identified by removing all genes, having higher than mean 13 expression (high expression of few genes arise due to technical problems). Additionally, minimal expression was set to 0.125. For dispersion cutoff, few different values were tested: -0.5; 0; 0.5; 1 and PCA was performed. Based on results, a cutoff of -0.5 was chosen.

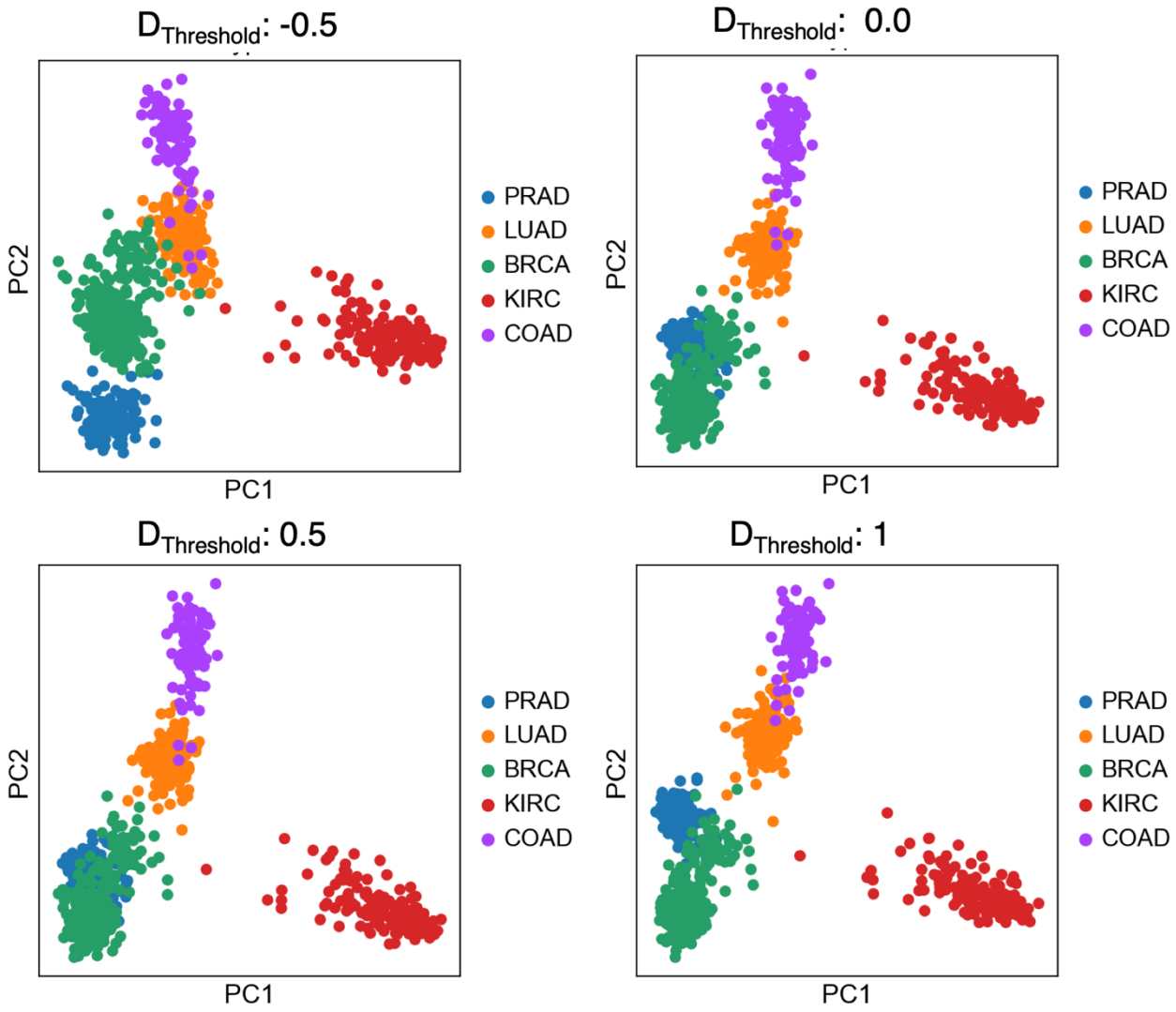

Figure 6. Principal component analysis with different dispersion cutoffs
Principal component analysis performed on gene expression data with different dispersion cutoff thresholds: -0.5; 0.0; 0.5; 1. Based on embedding of the data in 2D space threshold of -0.5 was chosen for further analysis. Cancer type labels were removed before doing PCA and then were mapped back on to the newly projected values. **BRCA** – Breast invasive carcinoma; **COAD** – Colon adenocarcinoma; **KIRC** - Kidney renal clear cell carcinoma; **PRAD** – Prostate adenocarcinoma; **LUAD** – Lung adenocarcinoma.

preserved as different *batch* name. To assess the optimal number of principal components, clustering was performed with 10, 20, 30, 40 and 50 PCs (principal component explained variance plots are not helpful in high-dimensional gene expression datasets). Main results with 20 and 30 PCs are shown in **Fig. 7**. All of the samples localize into discrete clusters. Notably, the only things that PC number affects is clustering of BRCA sample. For lower number of PC (10 and 20) BRCA cancer sample is segregated into 3 clusters (only 2 distinct). When higher number of PCs is utilized (30, 40 or 50), it only forms two clusters. Interestingly, BRCA sample visually is still composed out of two different populations. This separation is maintained for large number of nearest neighbors. Overall, each cancer type is assigned to a different, autonomous cluster (except for few cells that are miss assigned). Next, Differential Gene Expression (DGE) analysis is carried using a function from https://github.com/AllonKleinLab/klunctions/tree/master/sam/Analysis/scBasics/helper_funct

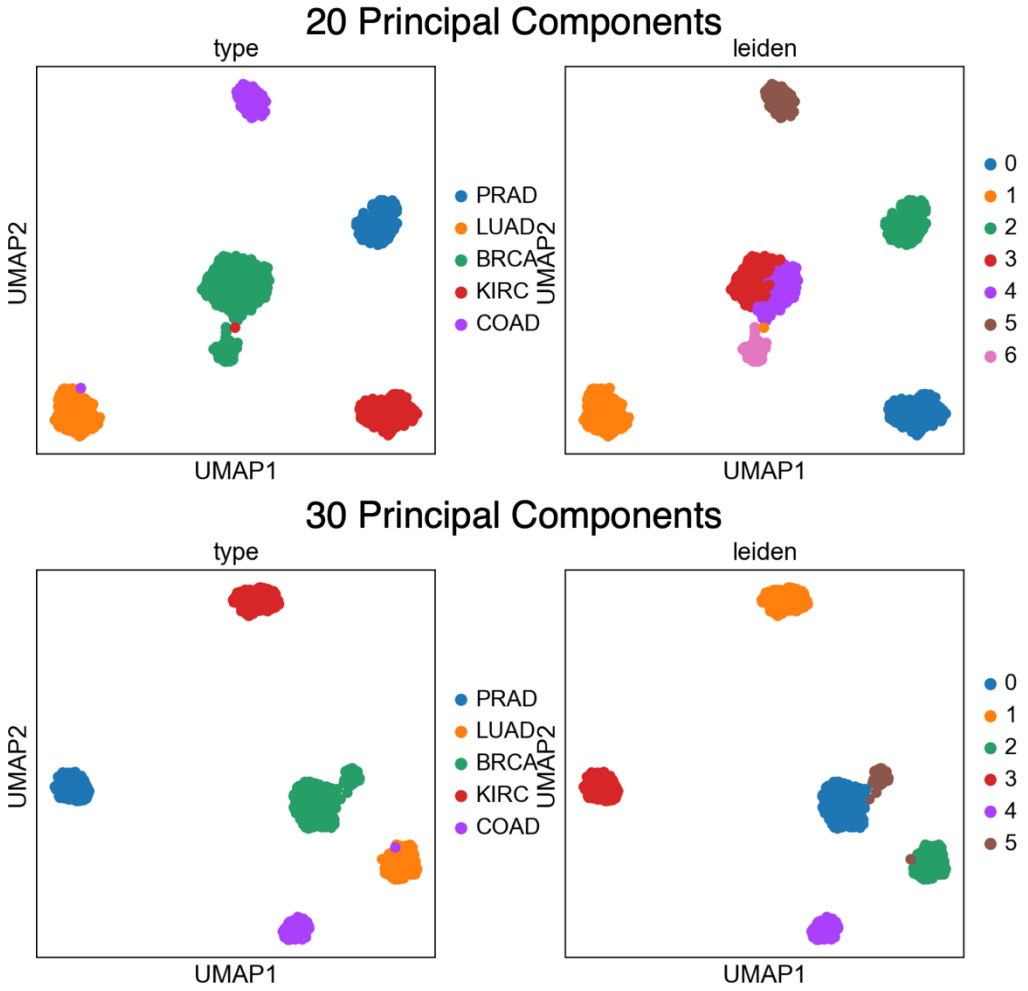

Figure 7. Leiden clustering with different number of principal components
To assess the optimal number of principal components to be used for Leiden clustering, clustering was performed with 10; 20; 30; 40 and 50 PCs. Graph only represents UMAP embedding with 20 and 30 PCs – point where BRCA sample forms 2 clusters instead of 3. **BRCA** – Breast invasive carcinoma; **COAD** – Colon adenocarcinoma; **KIRC** - Kidney renal clear cell carcinoma; **PRAD** – Prostate adenocarcinoma; **LUAD** – Lung adenocarcinoma.

ion.py. The function performs Rank Sums test between specified clusters gene expression and correct p-values with Benjamini/Hochberg method to remove false-positive cases.

DGE analysis of differently clustered BRCA cancer samples

First, DGE analysis was performed on BRCA sample clustered with 20 PCs to evaluate if there is any meaningful gene expression causing the splitting of the 3rd and 4th cluster. The expression of top 4 differently expressed genes in cluster 3 and 4 are shown in **fig 8**. There is no large difference in gene expression, except for *CASP14* (cluster 4, ratio – 3.90, p-value < 0.001) and *CIDEC* (cluster 3, ratio – 9.43, p-value < 0.001). *CASP14* is a caspase which is expressed during keratinocyte differentiation. Visualization of keratinocyte marker genes (*KRT5*, *KRT15*) show enrichment in both 3rd and 4th clusters (**Fig. 8**). This may arise due to different number of keratocytes in analyzed samples as this is bulk RNA-seq dataset. On the other hand, extracellular matrix degrading metaloproteinase1 (*MMP1*) expression is higher in *CASP14* expressing cells (ratio 3.6, p-value < 0.001). *CIDEC* expression plays important role in apoptosis via induced DNA fragmentation, however

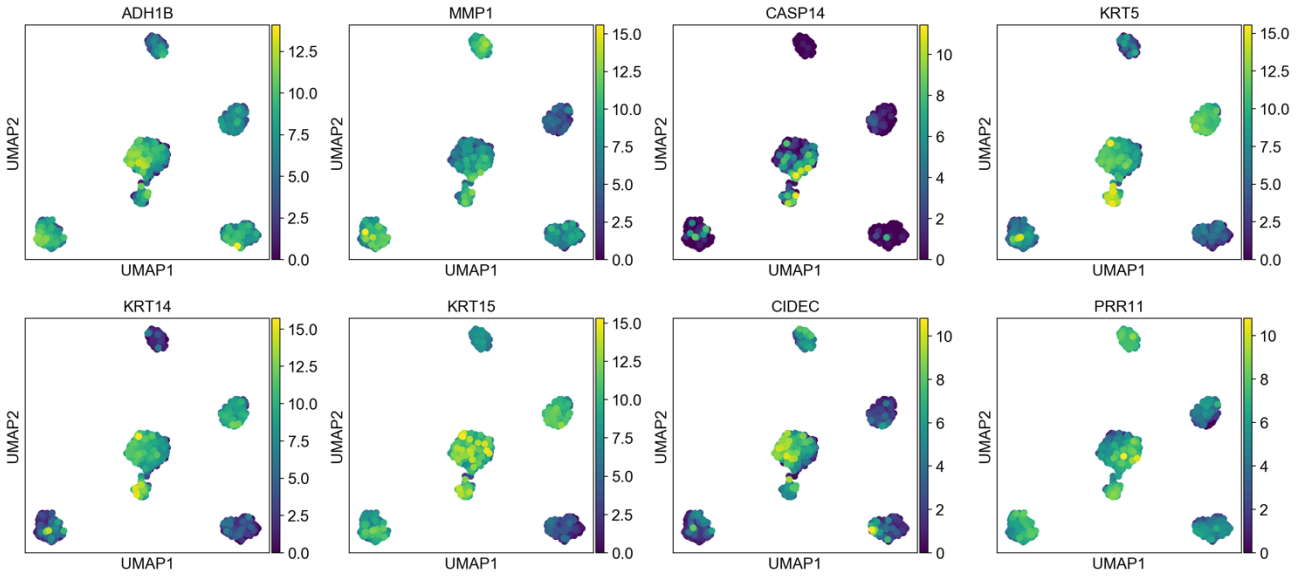

Figure 8. Differential gene expression of 3rd and 4th clusters in BRCA samples
Usage of 20 PCs for sample clustering lead to segregation of single unified BRCA cluster into two parts (Fig. 7 – 20PCs for reference). Graph shows visualization of differential gene expression on the UMAP embedding. Color bar represents gene expression values in log2(1+expression). Genes: Alcohol dehydrogenase 1B (***ADH1B)*** differently expressed in 3rd cluster; Matrix metalloproteinase-1 (***MMP1***) - 4th cluster; Caspase 14 (***CASP14***) – 4th cluster; Keratin5 (***KRT5***) - 3rd cluster; Keratin14 (***KRT15***) - 3rd cluster; Keratin15 (***KRT15***) – both clusters, Cell Death Inducing DFFA Like Effector C (***CIDEC***) - cluster 3; Proline rich 11 (***PRR11***) – cluster 4.

enrichment of other apoptosis genes is not seen. Overall, no biological difference in gene expression can be said from a simple DGE analysis of cluster 3 and 4. Performing single-cell RNA-seq would allow to visualize cell cycle and differentiation caused noise, which might be averaged in current sample due to bulk procedures. Based on this data, clustering of 30 PCs was analyzed further.

Differential gene expression analysis of data clustered using 30 principal components

Next, DGE analysis was performed on two remaining 30PCs BRCA sample clusters (0th and 5th – **fig. 7** for reference). Analysis showed, that most differently expressed gene in cluster 5 are *HORMAD1* (ratio - 64.4, p-value < 0.0001) and *ART3* (ratio -57.4, p-value < 0.0001). Interestingly, these genes are highly expressed in male testis. However, literature review and analysis showed that *HORMAD1* and *ART3* are highly expressed in basal-like breast cancer cells (Wang et al., 2018; Tan et al., 2016). Additionally, this type of breast cancer is defined by lack of expressions of estrogen (*GPER*), progesterone (*PGR*), and ERBB2 receptors (Tan et al., 2016). To test this hypothesis, gene expression of all three receptors was visualized on the UMAP embedding (**Fig 9**).

Indeed, most of the cluster 5 BRCA samples have a reduced expression of *GPER* (ratio – 1.98, p-value < 0.0001), *PGR* (ratio – 31.15, p-value < 0.0001) and *ERBB2* ratio – 2.27, p-value < 0.0001) receptors. Additionally, AGR3, highly expressed in 0th cluster (ratio - 257.69, p-value < 0.0001) is related to metastatic breast cancer properties (Obacz et al., 2019). Metastatic properties of cluster 0 cells are further supported by high expression of FOXA1 (ratio - 57.7, p-value <0.0001) and Androgen receptor (ratio - 38.4, p-value <0.0001) (Fig. 10) (Rangel et al., 2018). The correlation of two expressed genes in BRCA samples is

0.76 while correlation with the whole dataset is 0.35. These results could show that two clusters of BRCA sample are separated into basal-like and metastatic-like cancer samples

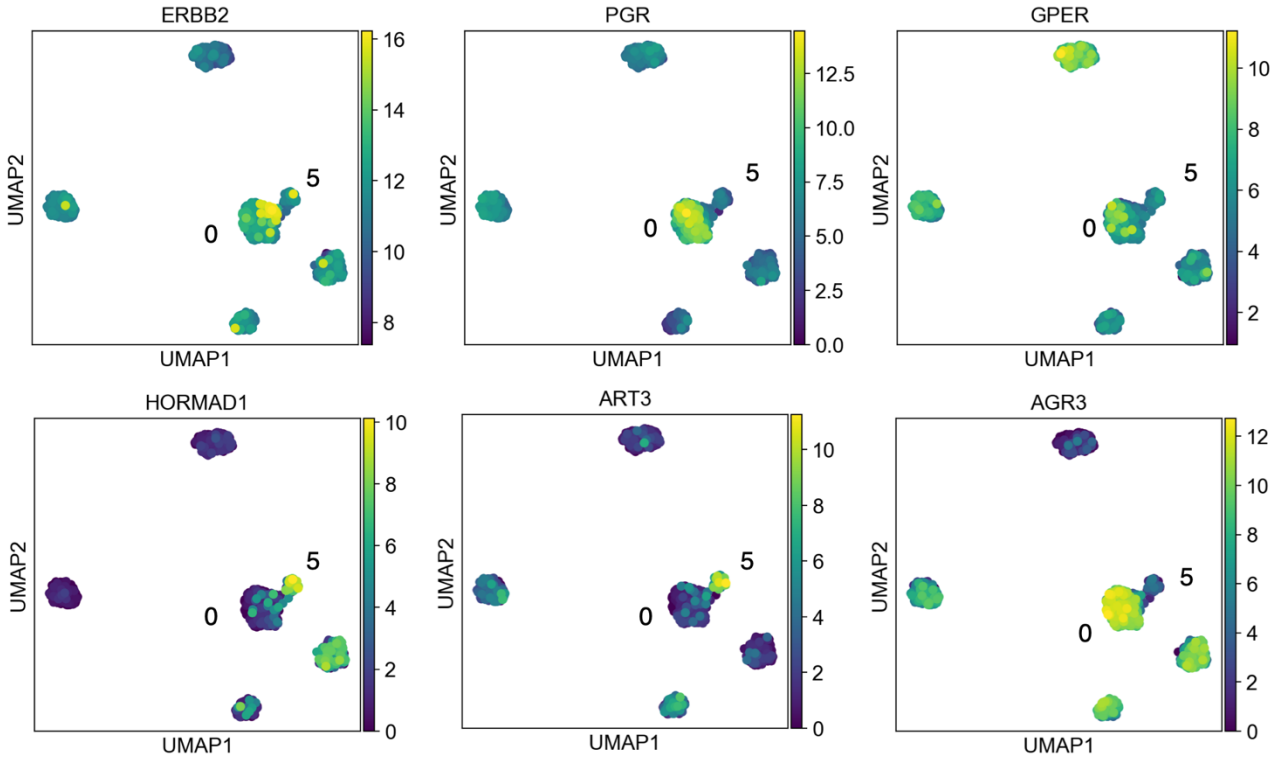

Figure 9. Visualization of differential gene expression on the UMAP embedding
Usage of 30 PCs for sample clustering causes the formation of two distinct BRCA sample cluster (Fig. 7 – 30PCs for reference). Visualization of differential gene expression on the UMAP embedding. Is show in the graph. Color bar represents gene expression values in log2(1+exprsion). Top row represents expression of triple-negative breast cancer marker genes: Erb-B2 Receptor Tyrosine Kinase 2 (***ERBB2***); Progesterone receptor (***PGR***) and G protein-coupled estrogen receptor 1(***GPER***). basal-like breast cancer cells markers HORMA domain-containing protein 1 (***HORMAD1***) and ADP-Ribosyltransferase 3 (***ART3***).

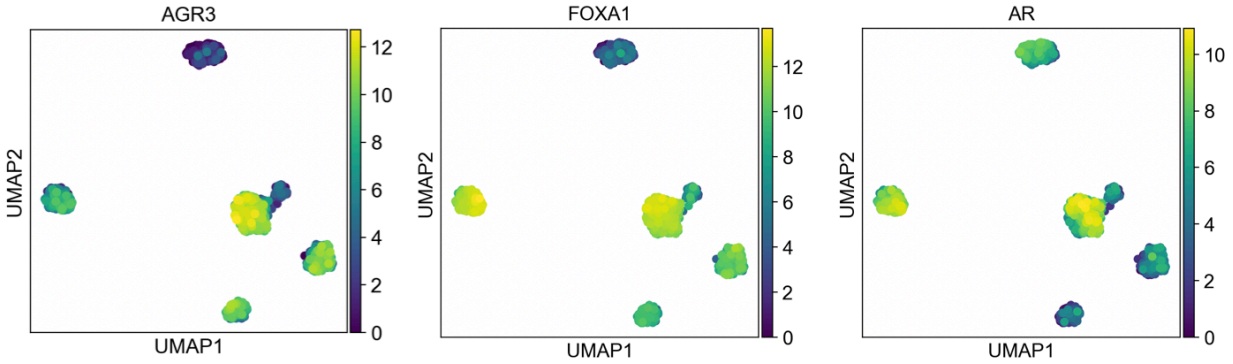

Figure 10. Visualization of metastatic BRCA cancer gene expression on the UMAP embedding
Metastatic properties inducing gene expression of *AGR3*, *FOXA1* and *AR* is enriched in cluster 0 of BRCA sample. Color bar represents gene expression values in log2(1+exprsion). ***AGR3*** - Anterior Gradient 3; ***FOXA1*** - Forkhead box protein A1; ***AR*** – androgen receptor

DGE analysis of the remaining cancer samples

Due to the nature of bulk RNA-seq, intra-sample resolution is lost. As expected, all the samples have differentially expressed genes related to their specific tissue and not a lot of things can be said from comparing them. For example, in cluster 1, corresponding to Kidney cancer (KIRC) samples, all top five (ACSM2A, CDH16, ACSM2B, SLC17A3 and SLC22A2) differently expressed genes have 1000-fold enrichment (p-value < 0.0001). All of these genes are specific to Kidney tissue. ACSM2 paralogs participate in the glycine conjugation pathway in the detoxification of xenobiotics such as benzoate and ibuprofen. Expression levels of this gene in the kidney may be correlated with kidney function. SLC encoded transporters participate in uric acid and other compound transportation in kidneys.

The same is seen in Lung adenocarcinoma sample – most differently expressed genes are lung surfactant associated genes. These encoded proteins bind specific carbohydrate moieties found on lipids and on the surface of microorganisms. *NKX2–1*, specifically expressed in LUAD sample (almost 1000-fold enrichment, p-value < 0.0001) is key lung development gene and is utilized as prognostic biomarker in adenocarcinomas (Moises et al., 2017). Interestingly, 7th differently expressed NAPSA gene can be used to distinguish adenocarcinomas from other forms of lung cancer (Uteno et al., 2004).

Most differently expressed genes in prostate cancer samples (*NYP, PRAC, PCA3*) belong to a group of proteins, named prostate-cancer specific antigens PSA (Lenka et al., 2013). Yet again they're detected as highly expressed (~1000-fold enrichment).

Colon cancer samples show the same trend again. KRT20 appears as a differently expressed gene and is used as a prognostic marker in cancer progression. Additionally, its

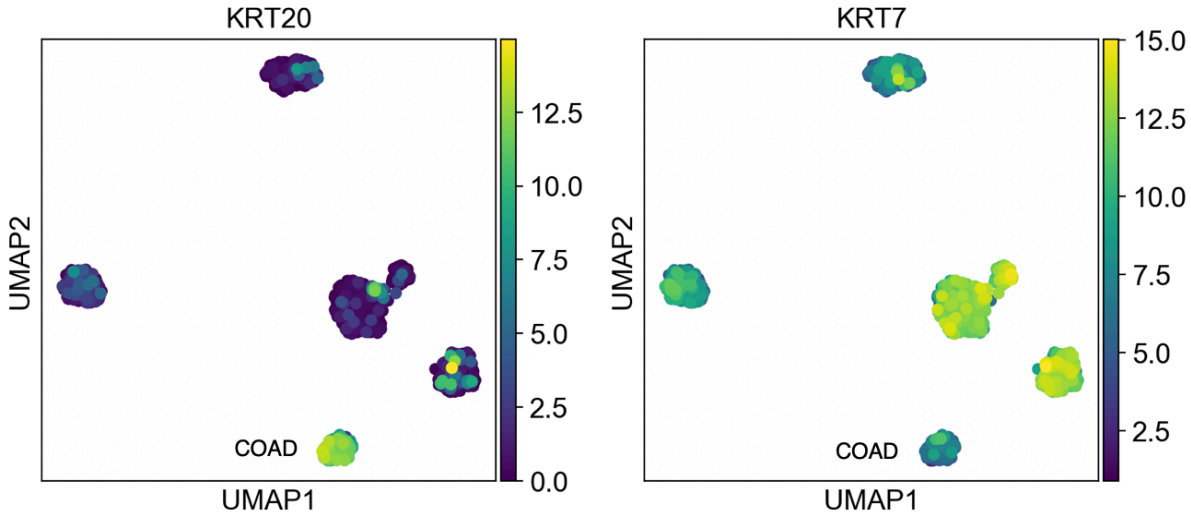

Figure 11. Visualization of prognostic marker KRT20 – KRT7 expression in Colon Cancer samples
KRT20 appears as a differently expressed gene in COAD sample and is used as a prognostic marker in cancer progression. Its high expression is followed by low expression of *KRT7.* Color bar represents gene expression values in log2(1+exprsion). COAD – Colon adenocarcinoma samples.

high expression is followed by low expression of KRT7 (Harbaum et al., 2012). Visualization of gene expression on UMAP embedding depicts this correlation perfectly (Fig. 10)

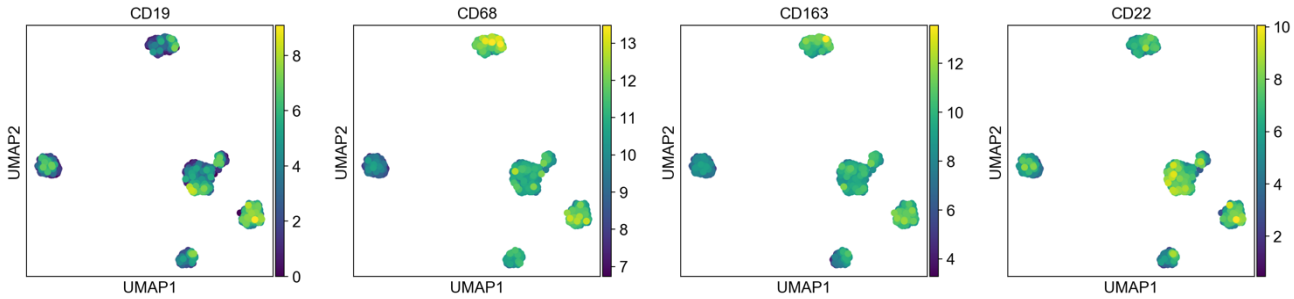

Figure 12. Visualization of immune cell marker gene expression
Visualization of expression of B-cell marker ***CD19***; macrophage markers ***CD68*** and ***CD163*** and Basophil markers ***CD22*** on the UMAP embedding.

Lastly, main immune cell markers were visualized on the UMAP embedding to investigate if there is any heterogeneity in cancel cell immune environments (Fig. 11). B cell marker CD19 is enriched in the LUAD cancer sample (2nd cluster) (~7-fold enrichment, compared to other clusters). CD68 and CD168 macrophage markers has a higher expression in prostate cancer samples (2.8-fold enrichment). Basophil maker CD22 is detected more in BRCA cancer samples (2-fold). It is important to note, that even though samples have different enrichment of immune cells, each dot represents an average gene expression of a patient sample. Size of the tissue sample, way of obtaining it and other technical variations can cause heterogeneity of immune cell markers and such data should be addressed with care.

Conclusions

The data set provides a large number of features and samples (801 x 20531) and therefore has a strong statistical significance. Descriptive data analysis revealed that all the samples have a similar coverage. On average, each sample has 132,287 counts with a minimal variation (standard deviation is 3183.43) and are distributed normally (p-value = 0.0004). The variance of gene expression expressed as log10(1+fano-factor) falls in between 0.2 – 0.6 for majority the genes and decreases with increased expression. Majority of genes have a sum count around 7500 with ~2000 genes having undetected expression. All samples have around 17,800 genes detected, and the overall gene detection is normally distributed. Descriptive analysis concluded that technical aspects of the dataset are high quality and there shouldn't be any biases arising due to it as variations are small and most the measurements are normally distributed.

Gene expression correlation revealed five district sample cluster of high gene expression correlation (> 0.9) corresponding to five cancer types. Additionally, all samples have > 0.7 correlation, showing that there are no technical or sequencing biases.

Principal component analysis verified gene expression correlation results as 5 different cancer samples are clustered into 5 different groups, with adenocarcinoma grouping more closely.

Leiden clustering performed with 20 and 30 principal components leads to different representation of BRCA samples (2 vs 3 clusters). Investigation of higher graduality clustering (3 clusters) leads to no additional gain of information as gene expression difference are small and should be addressed with caution.

Analysis of two distinct BRCA sample clusters, formed by 30 principal component Leiden clustering segregates the samples into invasive breast carcinoma cells and cells, having basal-like cancer properties.

Differential gene expression analysis of remaining cancer samples leads to detection of main, literature described prognostic markers of specified cancer types. Additionally, tissue specific genes are detected due to comparison of different tissues.

Analysis of distinct immune cell marker expression in the samples might hint to different cancer immune cell surrounding. However, a higher resolution analysis is needed to address this analysis with significance.

Literature

- Harbaum L., Pollheimer MJ., Kornprat P., Lindtner RA., Schlemmer A., Rehak P., Langner C. Keratin 20 - a diagnostic and prognostic marker in colorectal cancer? Histol. Histopathol. 2012 Mar;27(3):347-56.
- Lenka G., Weng WH., Chuang CK., Ng KF., Pang ST. Aberrant expression of the PRAC gene in prostate cancer. Int. J. Oncol. 2013 Dec;43(6):1960-6.
- Moisés J., Navarr A., et al., NKX2–1 expression as a prognostic marker in early-stage non-small-cell lung cancer. BMC Pulmonary Medicine (2017) 17:197.
- Obacz, J., Sommerova, L., Sicari, D., Durech, M., Avril, T., Iuliano, F. ... Fessart, D. (2019). Extracellular AGR3 regulates breast cancer cells migration via Src signaling. Oncology Letters, 18, 4449-4456.
- Rangel N., Fortunati N., Osella-Abate S. et al., FOXA1 and AR in invasive breast cancer: new findings on their co-expression and impact on prognosis in ER-positive patients. BMC Cancer. 2018 Jul 3;18(1):703.
- Tan L., Song X., et al., ART3 regulates triple-negative breast cancer cell function via activation of Akt and ERK pathways. Oncotarget. 2016 Jul 19;7(29):46589-46602.
- Ueno T, Linder S, Elmberger G. "Aspartic proteinase napsin is a useful marker for diagnosis of primary lung adenocarcinoma". Br. J. Cancer. 88 (8): 1229–33, 2004
- Wang X., Tan Y. et al., Epigenetic activation of HORMAD1 in basal-like breast cancer: role in Rucaparib sensitivity. Oncotarget, 2018, Vol. 9, (No. 53), pp: 30115-30127.