# A Re-exploration of Microarray Tumor Classification

## Introduction

Colorectal cancer (CRC) is the fourth-leading cause of death from cancer in the world, and the only method for prognostic classification for chemotherapy treatment is through pathological staging. Despite advances in cancer screening, pathological staging often fails to accurately predict recurrence in patients undergoing treatment. As a result, 10-40% of patients between stages II and III of CRC develop recurrence.

Previous research has shown microsatellite instability as the only molecular marker useful for early prognosis of CRC. Although the method of using microarrays to study gene expression profiles was considered, it was deemed poorly reproducible due to the multiple pathways associated with the cancer. Recent gene expression profile studies using unsupervised hierarchical clustering, integrated genetic/epigenetic analysis, and classification have identified at least three distinct subtypes of colon cancer. With this discovery, there was a transition to no longer viewing colon cancer as a homogenous identity. In this study, Marisa, et al. demonstrate a reproducible molecular classification method based on genome-wide mRNA expression analysis to diagnose prognostic biomarkers of CRC patients.

## Data

Samples from this study were obtained from the CIT cohort based in France, where patients ranged from stage I to IV of colon cancer. 750 frozen tumor samples were taken from patients that had undergone surgery between 1987 and 2007. Of the 750 samples, 566 passed the RNA quality requirements for gene expression analysis. Affymetrix U133 Plus 2.0 chips were used to run these samples, and the resulting data was split into two sets: discovery and validation. The discovery set had 443 samples and the validation set had 123. The validation dataset was further supplemented with 906 samples from seven public datasets (GSE13067, GSE13294, GSE14333, GSE17536/17537, GSE18088, GSE26682, and GSE33113). All of these samples used a similar chip, and had raw data and additional information available.

An additional dataset was added to the validation set from The Cancer Genome Atlas (TCGA), accounting for another 152 samples. Although the TCGA data was obtained through a non-Affymetrix chip, it provided insightful annotations on the DNA alterations in the colon cancer samples. The analysis was done using CEL files located in a central location in the shared computing cluster. There was a missing file which was obtained through downloading from the NCBI Gene Expression Omnibus (sample ID: GSM971958). All of the samples were linked to the same folder for access. Each sample goes through RNA purification, quality

control, fluorescent probe production, hybridization, and raw data processing, and the data can be found under the accession number GSE39582.

## Methods

To analyze the data, packages from the Bioconductor repositories were used, specifically the libraries affy, affyPLM, and sva. ReadAffy was used to read all the CEL files from the folder entitled "Samples" in order to produce an AffyBatch. The affy function rma(), standing for Robust Multi-Array Average expression measure, was used to normalize these samples. This function converted the data read in the previous step into an expression set. It took quite a few minutes to run as it ran through multiple steps.

Then, the AffyBatch was converted into a PLM dataset using FitPLM. From there, Normalized Unscaled Standard Errors (NUSE) and Relative log expression (RLE) graphs were produced. NUSE produced the standard error estimates for each probe set. During this process outliers were identified. The RLE analyzed the level of variation for each probeset. The median values for each were then plotted in the form of a histogram against frequencies.

The project metadata file was then read into a dataframe and the "Normalization Combat Mod" and "Normalization Combat Batch" column values were stored as variables. Inputting these two variables along with the rma() output data into the function ComBat() resulted in a probe x sample genomic measure matrix which had been adjusted for batch effects. The purpose of this step was to adjust for batch effects and produce a model matrix based on the form of "Normalization Combat Mod". Outliers were removed and batch effects were removed using "Normalization Combat Batch". The output of this function was then saved as a .csv file called "expression_data.csv".

The final step was to produce the PCA graph and analyze principal components. The dataframe produced in the previous step was transposed, scaled, and transposed again. The resulting matrix was inputted into the function prcomp() and the arguments center and scale were both set to false since scaling and centering had been performed in the previous steps. Principal Components 1 and 2 were then plotted against each other with the variance explained in the axis labels.

Clustering analysis was performed again on the expression data subject only to the chi-square filter. Expression level fold change (FC) between the clusters was used to assess regulation changes. Disregarding significance of differential expression, the top 1000 up- and downregulated genes by FC (2 sets of 1000 each) were used in enrichment analysis. With the Bioconductor gene annotation package corresponding to the similar chip platform used across the public data sets, hgu133plus2 from Affymetrix [4], the probe IDs were mapped to gene symbols for comparison against major gene sets. When different probe IDs mapped to the same gene symbol, the one with the greatest FC was preserved for the sake of enrichment analysis.

The gene set collections were obtained from MSigDB and accessed via Bioconductor package GSEABase [5,6]. They included cancer hallmark gene sets (50 sets), KEGG curated sets (186 sets), and all GO

ontology sets (10402 sets) [6-8]. Fisher's exact test was used for the comparisons, each pairing one of the gene collections with either the up- or downregulated gene sets (6 comparisons total). After adjusting p-values (Benjamini-Hochberg) for multiple hypotheses, significantly enriched sets ($p < 0.05$) were enumerated and the top 3 (by significance) were taken from each comparison for reflection to Marisa, et. al. [3].

## Results

**Figure 1.**

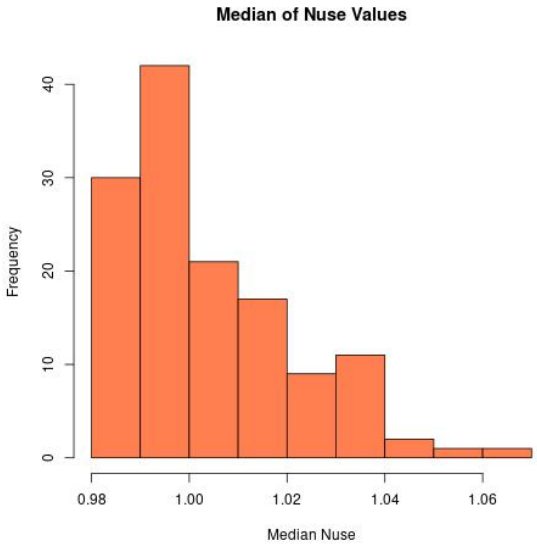



**Figure 2.**

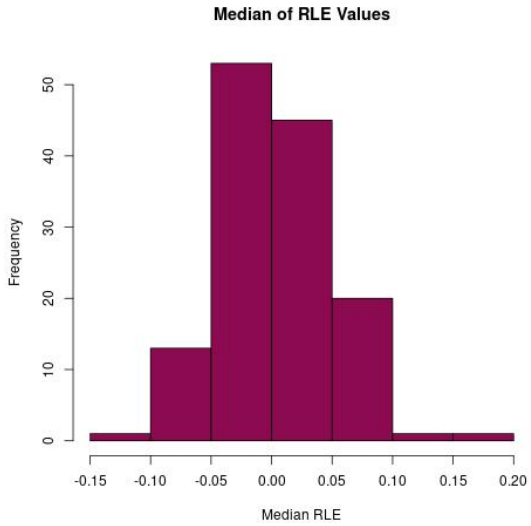



**Figure 1** The histogram of Median Normalized Standard Error (NUSE) values displays the median NUSE standard errors and the frequency at which they appeared. In the median NUSE values histogram shown, it is clear the most frequent median NUSE value was 0.9. Toward the higher end of the median NUSE values, the frequency decreased proving that not many of the samples had increased standard errors and there were not many low quality samples.

**Figure 2** The Median Relative Log Expression (RLE) histogram shows the frequency of the median values of RLE. The highest frequency is close to 0 as expected, since an RLE of 0 would indicate that there are no differentially expressed genes.

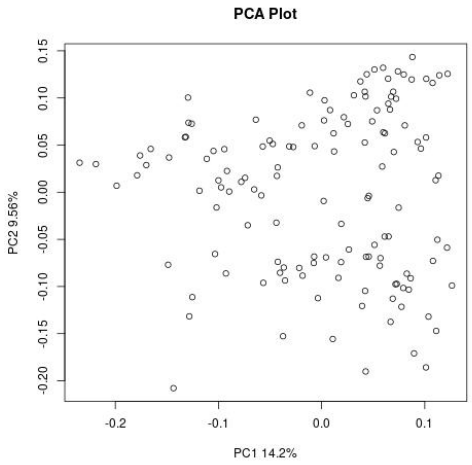



**Figure 3** In Figure 3 a PCA plot was created using PC1 and PC2 produced after running PCA on the expression dataframe. PC1 had a variance of 14.2% while PC2 had a variance of 9.56%.

Noise Filtering and Dimensionality Reduction: In order to process and explore the data effectively, noise filtering and dimensionality reduction performed first. An Rscript was used to achieve this, the dimensionality of the probes was reduced based on three metrics defined by Marissa et al. First, the expression data was read by reading csv expression matrix file which returned a total of 54,675 probes. The first filtration step involved a logarithmic test which involved returning the genes where at least 20% of the gene-expression values were greater than log2(15), through which all the probe sets passed this criteria. Second filtration involved chi square test to filter the data according to median variance which left the data with 24,239. Finally, the variation filter was used which cleared the matrix of probes having a coefficient of variation > 0.186.

**Figure 4.1**

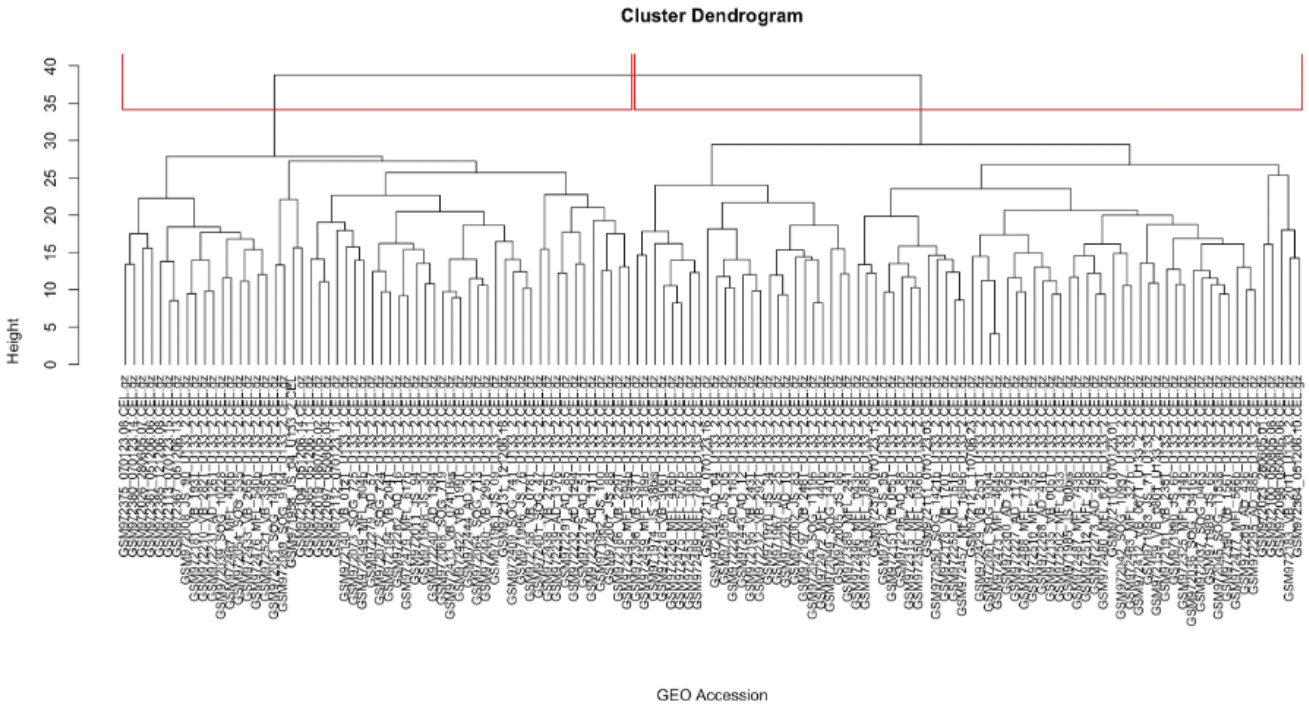
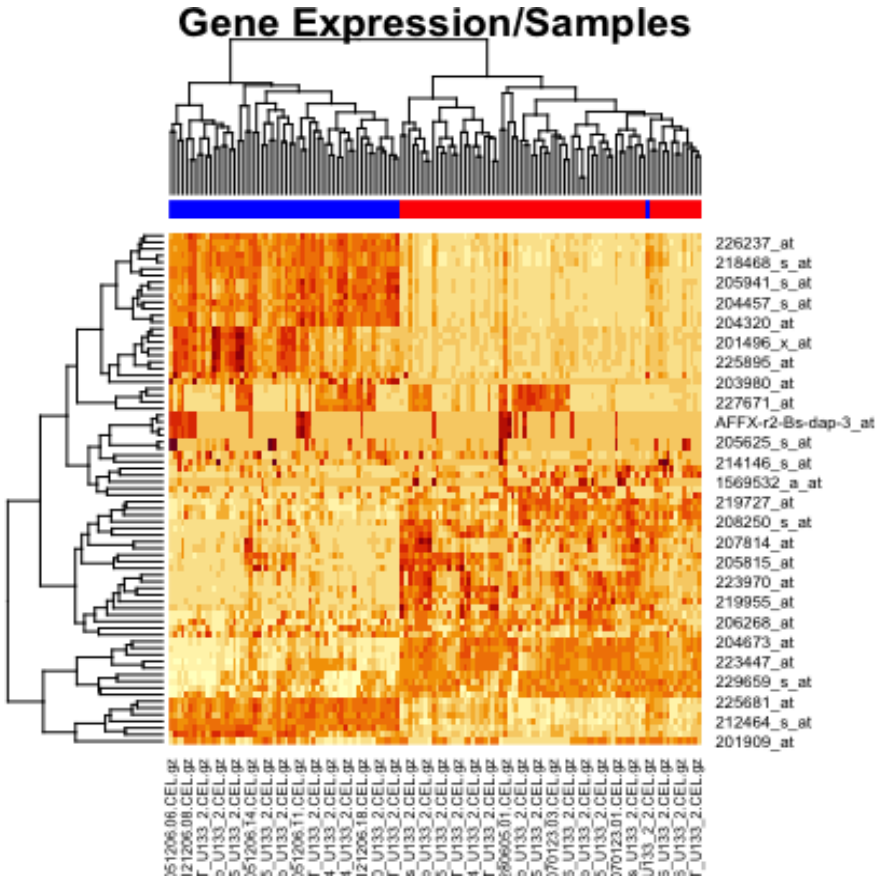



**Figure 4** Hierarchical clustering and Subtype: After having filtered the data that passed above thresholds were clustered by using an unsupervised method of hierarchical clustering to discover associations and patterns within the probe sets. The data was divided into two clusters with cluster1 having a size of 58 and cluster2 76 probes based on similarity of gene expression. A heatmap was produced to observe the colon molecular subtype into two categories C3 being red and others being blue. The classification for subtype was collected from the metadata csv file along with the respective geo accession IDs. A gene is considered differentially expressed if a difference or change observed in read counts or expression levels/index between two experimental conditions is statistically significant. 77 out of all were observed to be differentially expressed in this dataset. To further analyze the clusters a welch t-test was performed where differentially expressed genes had adjusted p value $< 0.05$. Additionally, a t test was performed on a dataset that passed the second variance filter from noise filtering steps and a data frame was created with p value, t statistic value and an adjusted p value.

**Figure 4.2**




After chi-square and symbol duplicate filtering, there were 13,708 genes left in the expression data set. For gene enrichment analysis, the 1000 most up- and downregulated genes were selected; the greatest 10 are given in Table 1. Statistics are from evaluating differential expression according to Welch's t-test, with p values adjusted for false discovery rate (Benjamini-Hochberg). The log base 2 of FC between the clusters is also given.

**Table 1.** The top 10 up- and down regulated genes between the clusters, highlighted red and green respectively.

| Gene Symbol | Diff. Exp. t | Diff. Exp. p | Adj. p | $Log_2FC$ |
|---|---|---|---|---|
| FCGBP | -18.52 | 4.93e-37 | 2.92e-34 | 0.5804 |
| CLCA1 | -13.93 | 1.67e-27 | 1.82e-25 | 0.5636 |
| SPINK4 | -13.06 | 4.43e-25 | 3.17e-23 | 0.5388 |
| MUC2 | -14.80 | 2.57e-29 | 3.64e-27 | 0.4507 |
| REG4 | -8.98 | 1.67e-14 | 2.18e-13 | 0.4491 |
| OLFM4 | -7.23 | 8.16e-11 | 6.35e-10 | 0.4306 |
| ST6GALNAC1 | -15.54 | 3.26e-28 | 3.93e-26 | 0.4292 |
| ITLN1 | -10.70 | 5.37e-19 | 1.41e-17 | 0.4204 |
| AGR3 | -11.96 | 5.37e-19 | 1.41e-17 | 0.4181 |
| HEPACAM2 | -11.98 | 3.52e-22 | 1.56e-20 | 0.4084 |

| | | | | |
|---|---|---|---|---|
| C3 | 14.94 | 5.22e-28 | 6.02e-26 | -0.4065 |
| SPOCK1 | 20.94 | 8.16e-36 | 3.80e-33 | -0.4264 |
| SULF1 | 20.39 | 1.53e-40 | 1.46e-37 | -0.4282 |
| COL10A1 | 14.54 | 5.02e-23 | 2.60e-21 | -0.4568 |
| MYL9 | 18.50 | 2.19e-33 | 5.90e-31 | -0.4589 |
| CTHRC1 | 18.57 | 1.56e-38 | 1.08e-35 | -0.4679 |
| FNDC1 | 21.10 | 1.08e-35 | 4.75e-33 | -0.4933 |
| GREM1 | 17.44 | 4.65e-36 | 2.25e-33 | -0.4990 |
| THBS2 | 20.38 | 1.45e-41 | 1.84e-38 | -0.5329 |
| SFRP2 | 18.33 | 9.03e-29 | 1.18e-26 | -0.5810 |

The enumerations of significantly enriched gene sets are recorded in Table 2, with the top 3 from each comparison compiled in Table 3.

**Table 2.** Enumeration of enriched gene sets from three major collections in up and down regulated genes.

| Gene Set Collections | Enriched in 1000 Upregulated | Enriched in 1000 Downregulated | Total in Collection |
|---|---|---|---|
| Cancer Hallmark | 7 | 17 | 50 |
| KEGG Curated | 21 | 17 | 186 |
| All GO Ontology | 24 | 412 | 10402 |

**Table 3.** Top 3 enriched gene sets of 3 collections by Fisher's exact test significance adjusted for multiple hypotheses with the Benjamini-Hochberg method. Enrichment in up- or downregulated genes is indicated with red and green highlights, respectively.

| Gene Set | Hypergeometric Statistic (Odds Ratio) | p | Adj. p |
|---|---|---|---|
| HALLMARK_ESTROGEN_RESPONSE_LATE | 3.815 | 6.18e-12 | 3.09e-10 |
| HALLMARK_GLYCOLYSIS | 2.562 | 7.84e-06 | 1.96e-4 |
| HALLMARK_G2M_CHECKPOINT | 2.467 | 2.04e-05 | 3.39e-4 |
| HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION | 12.593 | 7.26e-55 | 3.63e-53 |
| HALLMARK_KRAS_SIGNALING_UP | 4.284 | 2.16e-14 | 5.41e-13 |
| HALLMARK_COAGULATION | 5.166 | 7.24e-14 | 1.21e-12 |
| KEGG_ASCORBATE_AND_ALDARATE_METABOLISM | 13.928 | 3.50e-09 | 3.80e-07 |
| KEGG_DRUG_METABOLISM_CYTOCHROME_P450 | 5.637 | 4.09e-09 | 3.80e-07 |

| | | | |
|---|---|---|---|
| KEGG_DRUG_METABOLISM_OTHER_ENZYMES | 7.038 | 8.17e-09 | 5.06e-07 |
| KEGG_COMPLEMENT_AND_COAGULATION_CASCADES | 6.478 | 2.59e-10 | 4.82e-08 |
| KEGG_FOCAL_ADHESION | 2.773 | 9.25e-07 | 8.60e-05 |
| KEGG_ECM_RECEPTOR_INTERACTION | 4.031 | 1.82e-06 | 1.12e-4 |
| GOBP_XENOBIOTIC_GLUCURONIDATION | 128.346 | 4.21e-11 | 2.94e-07 |
| GOBP_FLAVONOID_GLUCURONIDATION | Inf | 5.66e-11 | 2.94e-07 |
| GOBP_SMALL_MOLECULE_METABOLIC_PROCESS | 1.686 | 2.05e-09 | 7.10e-06 |
| GOCC_COLLAGEN_CONTAINING_EXTRACELLULAR_MATRIX | 6.015 | 1.12e-45 | 1.16e-41 |
| GOCC_EXTERNAL_ENCAPSULATING_STRUCTURE | 5.056 | 2.84e-45 | 1.48e-41 |
| GOMF_EXTRACELLULAR_MATRIX_STRUCTURAL_CONSTITUENT | 8.398 | 2.43e-31 | 8.42e-28 |

**Discussion**

From the public data used in Marisa et. al. gene expression analysis was carried out *a la* study. After retrieving the data from GEO, normalizing, and reducing it with quality filters, PCA and clustering were performed to observe trends for categorizing CC. Then, gene enrichment was performed against 3 major gene set collections.

The three filtering metrics that were used to reduce the dimensionality in the normalized expression matrix data ultimately filtered probe sets down to 77 compared to Marisa et al which could be due to different filtration technique while they filtered for 5% with log2(15) we filtered ours by 20% but we still have a majority of probe sets in common with Marisa et al. For the subtype classification with unsupervised method, hierarchical clustering was able to classify the 134 patients into two clusters, with 58 samples for cluster 1 and 76 samples for cluster 2 different results were obtained as our clustering involved only two categories one for C3 and second being others instead of 6 subtype.

In the top ten up- and down- regulated genes, we observed some known and suspect oncogenes. In the upregulated set, OLFM4 and AGR3 are implicated in tumor growth, particularly OLFM4 which is an extracellular protein that promotes cell adhesion [9,10]. Among the downregulated genes, SULF1 is known to be downregulated in many varieties of cancer [11]. The gene THBS2 is a known tumor suppressor, and its downregulation is typical for growing tumors [12].

The gene enrichment analysis recapitulated several of the findings displayed in Figure 2 of Marisa et. al. In particular, we reproduced three of their reported enrichment results: enrichment in KEGG metabolism of xenobiotics by cytochrome P450, KEGG focal adhesion, and KEGG ECM receptor interaction. The up- vs. downregulation assignment matches their results for cluster 3 (C3). Furthermore, in C3 Marisa et. al.

reported several enriched pathways amongst the upregulated genes for simple sugar metabolism (pentose, fructose, sucrose) and in our results we see the cancer hallmark gene set for glycolysis also enriched amongst the upregulated genes. This occurs again with our result for the cancer hallmark gene set for epithelial mesenchymal transition; Marisa et. al report a few GO pathways of the same variety enriched in the downregulated genes.

We limited our results to the top three from each comparison. From these, we see further subtle similarities that suggest a closer relationship to Marisa et. al's result, including our report of enrichment in cell cycle G2M checkpoint genes and their report of enrichment in cell cycle genes in general.

## Conclusion

Overall, the results of the replication agree well with the findings of Marisa et. al. Many patterns observed in their cluster 3 are reflected by our clustering, indicating the reliability of these unsupervised learning techniques for classifying CC subtypes and the robustness of the data against varied filters and other qualifications. Introducing the cancer hallmark gene set collection into the analysis provided a new perspective on the utility of the molecular subtype classification presented by the study.

## References


1. Marisa L., de Reynies A., Duval A., Selves J., et al. (2013). Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. PLoS Med 10(5): e1001453. doi:10.1371/journal.pmed/1001453
2. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/
3. Aravind S., Pablo T., Vamsi K. M., et. al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43), 15545-15550. DOI: 10.1073/pnas.0506580102
4. Carlson M. (2016). hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2). R package version 3.2.3.
5. Morgan M, Falcon S, Gentleman R (2021). GSEABase: Gene set enrichment data structures and methods. R package version 1.56.0.
6. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell systems, 1(6), 417–425. https://doi.org/10.1016/j.cels.2015.12.004
7. Ashburner et al. (2000). Gene ontology: tool for the unification of biology. Nature Genetics, 25(1), 25-9.
8. Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30.
9. Wei, Y., Song, Q., Zhang, F., & Yuan, T. (2021). Olfm4 Is Highly Expressed in HCC Patients and as a Biomarker and Therapeutic Target for HCC. Canadian journal of gastroenterology & hepatology. https://doi.org/10.1155/2021/5601678
10. Joanna O., Martina T., Veronika B., et. al. (2015). The role of AGR2 and AGR3 in cancer: Similar but not identical. European Journal of Cell Biology, 94(3–4), 139-147. https://doi.org/10.1016/j.ejcb.2015.01.002.
11. Lai, J. P., Sandhu, D. S., Shire, A. M., & Roberts, L. R. (2008). The tumor suppressor function of human sulfatase 1 (SULF1) in carcinogenesis. Journal of gastrointestinal cancer, 39(1-4), 149–158. https://doi.org/10.1007/s12029-009-9058-y
12. Jack L., Michael D. (2004). Tumor progression: the effects of thrombospondin-1 and -2. The International Journal of Biochemistry & Cell Biology, 36(6), 1038-1045, https://doi.org/10.1016/j.biocel.2004.01.008.