|  |
| --- |
| **Differential gene expression analysis and machine learning model classification for pancreatic ductal adenocarcinoma[[1]](#footnote-2)**  Sara Chung1, Khang Duong1, Colette Trouillot 1  1 School of Biomedical Engineering, Drexel University, USA  Course : Bmes543 Quantitative Systems Biology  Instructor: Ahmet Sacan  Date : 2022-06-07  Original Paper(s) :  <https://pubmed.ncbi.nlm.nih.gov/19260470/>  Dataset(s) :  <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse15471>  <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16515>  <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22780> |

abstract

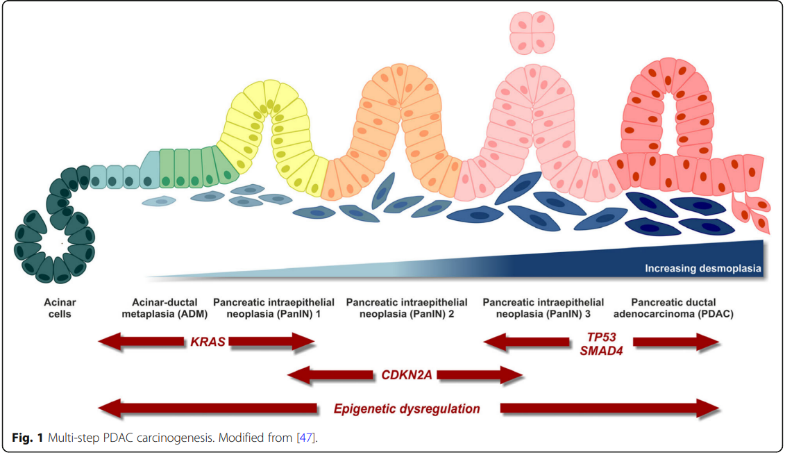
Pancreatic ductal adenocarcinoma is a highly aggressive cancer with an average 5-year survival of less than 10%1. The high lethality is due to the lack of early diagnosis, extensive heterogeneity, rapid metastasis and limited response to treatments. Due to strong desmoplastic (growth of fibrous tissue around cancer) reaction and neoplastic epithelial cells representing minority population of cancer cells, current PDAC microarray studies have been limited2. In this study, we identified overlapping differentially expressed genes in 3 datasets and determine their biological significance using python and developed a machine learning model to predict the tumor vs normal cells. Our analysis identified overlapping significant genes where greater than 66 percent the genes have been captured in previous studies or are known precursors in oncogenic signaling pathways. Our machine learning models resulted in greater than 85 percent cross validation accuracy for higher volume sample sets. Our findings with this model may address the limitations in characterizing PDAC to improve early diagnosis, therapeutic efficacy and clinical outcomes.

# introduction

The most common type of pancreatic cancer (PC), Pancreatic Ductal Adenocarcinoma (PDAC), is a highly aggressive cancer, with a median survival of <6 months and an average 5-year survival rate of 3-5%1. It accounts for >90% of all PC diagnoses and boasts a high lethality due to lack of early diagnosis, extensive heterogeneity, rapid metastasis and limited response to treatments2. The extensive heterogeneity of genetic mutations causing this cancer leads to a wide range of tumor microenvironments that make it very difficult to treat3. With classical treatments such as chemotherapy, surgery and radiation showing limited improvements in clinical outcomes and more effective systemic chemotherapy treatments being inaccessible to older or more frail individuals, there are few effective treatments for PC2. Just over 60,000 people will be diagnosed with pancreatic cancer this yearand nearly 50,000 of them will die making it the 4th largest cause of cancer death in the United States2. Survival rates of those with PC have not improved in 40 years indicating a strong need for more research and improved treatment methodologies3.

The pancreas is a large gland lying along the stomach and small bowel4. Its main functions are to aid in digestion and help control blood sugar4. The pancreas comprises exocrine, epithelial, and endocrine cells. PDAC is caused by a process called acinar (exocrine) to ductal metaplasia where the acinar cells differentiate to epithelial (ductal-like) phenotypes when experiencing macro- and microtumor environments (e.g. tissue damage, inflammatory, or stress conditions) during which these cells acquire ‘progenitor cell-like’ characteristics and become susceptible to pro-oncogenic hits such as activating mutations in the proto-oncogene KRAS, eventually transforming these cells into pancreatic intra-epithelial neoplasias (PanINs)5.

A sequential progression following this involves additional genetic hits in several tumor suppressor genes as demonstrated in **Figure 1**5,6. For example, KRAS, TP53, CDKN2A, and SMAD4 all are frequent mutations/alterations in PDAC6,7. The hallmark of PDAC is the extensive desmoplastic stroma that is caused by tumor microenvironment traits such as hypoxia5. This in turn causes strong accumulation of myeloid cells that could block CD4 and CD8 t-cells into the PDAC tumor microenvironment5. Additionally immunosuppressive T and B cell subpopulations have been detected in the microenvironment that can block T cell activation and infiltrate effector T cells5.



**Figure 1 PDAC Cancer Progression:** This image demonstrates the progression from acinar cells to pancreatic ductal carcinoma and highlights several key pathways in the path5.

This study focuses on the gene expression profiles found in tumorigenic pancreatic cancer tissue which several authors have previously looked at as well. Iacobuzio-Donahue et al and Grutzman et al have previously looked at gene expression profiles for PDAC although these were mostly on unpaired datasets unlike this research that focuses primarily on paired data8,9. There is also the paper that this study is in part based on that focused on combining laser microdissection studies and microarray data10. An additional very recent paper by Santosh & Chandrasekar published in 2022 used machine learning methodologies to predict; however, they were focused on ensemble approaches and analyzed primarily biomarkers from urine unlike the pancreatic cancer tissue we focus on11.

Dure to the strong demoplastic (growth of fibrous tissue around the cancer) reaction and the low population of neoplastic epithelial cells, current PDAC microarray studies have been limited5. The results of this combination study aims to provide a more precise determination of overexpressed genes in tumor epithelia. This identification of over-expressed genes in cell types that are too difficult to isolate with microdissection. Our findings using this model may address the limitations in characterizing PDAC to improve early diagnosis, therapeutic efficacy and clinical outcomes.

# Dataset

**2.1 Dataset #1: Training Set**

The training dataset used came from *Whole-Tissue Gene Expression Study of Pancreatic Ductal Adenocarcinoma.* The dataset included expression analysis using pairs of normal and tumor tissue samples obtained from 36 pancreatic cancer patients, 3 of which were carried out in replicate to gauge technical errors. It included a total of 78 samples and was analyzed using whole genome microarrays10. The tumor and normal tissue data was paired—both sets of samples came from the same patients.

**2.1 Dataset #2: Testing Set 1**

The first testing set was obtained from *Expression data from Mayo Clinic Pancreatic Tumor and Normal Samples* samples12*.* The dataset included microarray analysis using RNA isolated from 36 tumor and 16 normal samples. It was originally used to identify expression differences of FKBP5 genes between tumor and normal samples. It included a total of 52 samples12. Of the data, 16 were paired normal & tumor samples from the same patient while 20 of the tumor samples were not paired.

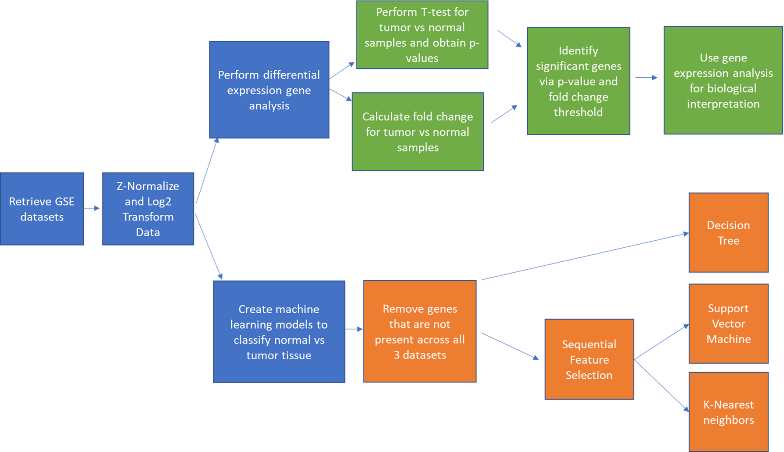
**2.2 Dataset #3: Testing Set 2**

The second testing set came from *Affymetrix Arrays Interrogated with Tumor/Normal Pancreatic Samples.* The dataset included microarray analysis of 8 frozen tumor and adjacent normal tissue from the same patient used to identify genes in the chromosome 3p12 pathway to tumorigenesis. It included a total of 16 paired samples13.

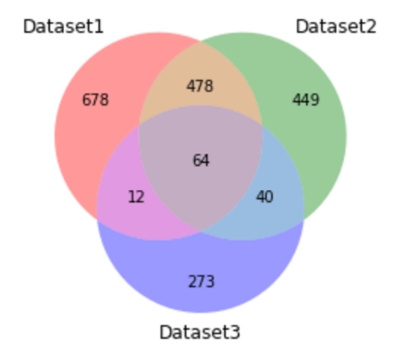
# methods

The software used for this analysis included Python implemented in Visual Studios & Jupyter Notebook ran using both Microsoft Windows and macOS. Modules used in Python included pandas, numpy, math, matplotlib, and scipy as well as machine learning models primarily from the sklearn toolset. The Database for Annotation, Visualization, and Integrated Discovery (DAVID), a comprehensive set of functional annotation tools, was used to relate the functional terms with gene lists. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis were performed using DAVID14,15.

The basic workflow used is described in the workflow diagram in **Figure 3**.



**Figure 2: Methods - Analysis Workflow:** This workflow demonstrates the analysis pathway used to identify differentially expressed genes for biological interpretation and the creation of machine learning models for predictions of tumor vs normal data.



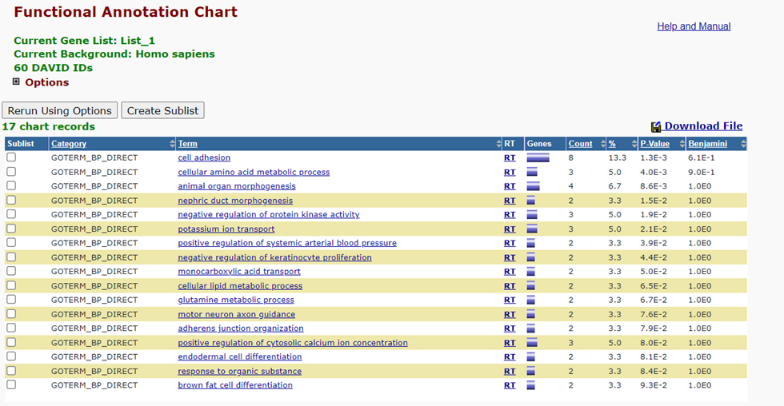
**Figure 3 Venn Diagram:** This venn diagram demonstrates the overlap of significant genes within the 3 datasets.

Raw GSE data from 3 different studies were taken for analysis in Python. All data was normalized and log transformed to ensure relevancy and the removal of duplicated and unnecessary data. From there, differential expression gene analysis was performed to determine the significant genes. Significant genes were analyzed using a 2 sample (normal v tumor) t-test and fold changes. A minimum p-value of 0.02 and fold change value of 1.5 was used to ensure that only the most significant genes were considered. This data was then taken into DAVID Bioinformatics tool to determine the biological interpretation of the genes that were found.



**Figure 4 KEGG Pathway Analysis:** The results of the KEGG pathway analysis from DAVID. DAVID was unable to match 60 of the overlapping genes and was only able to determine this pathway.

For the machine learning analysis, genes that were not present across all 3 datasets were removed to ensure that the models are trained on the same number of features. Three machine learning models were used in this analysis: decision trees, support vector machine, and k-nearest neighbors. The models were trained on dataset 1, and a 5-fold cross validation was performed to measure model performance and determine model parameters. Sequential forward feature selection was applied to the datasets for the support vector machine and k-nearest neighbors models in order to reduce irrelevant genes and improve model accuracy.



**Figure 5 Functional Annotation Chart:** Gene Ontology (GO) enrichment analysis from DAVID yielded these GOterms and the number of genes associated with each.

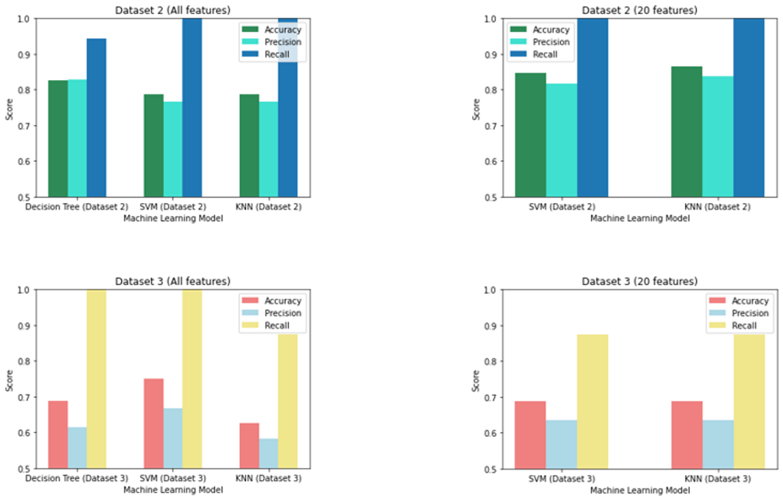
# Experiments and REsults

The experiment was performed in two parts—a differentially expressed genes analysis and a machine learning model. As seen in the Venn diagram in **Figure 3**,there were 64 genes found to be present in all 3 datasets. These are the genes that were further selected for gene set enrichment analysis. The KEGG pathway analysis results for this dataset are shown in **Figure 4**. DAVID was unable to match the majority of genes in the dataset; however, DAVID did find genes associated with the p53 pathway which is a well-known important pathway in PDAC with previous studies with a mutated version being found in 75% of case16. Mutant p53 also promoted more metastases compared to those without it and more research needs to be done to evaluate p53 as a potential therapeutic target16.

**Figure 5** shows the results of the gene ontology enrichment analysis which includes ion transport, cell adhesion, and morphogenesis which previous papers have described as significant terms for PDAC16,17.

Additional terms of interest include cellular amino acid metabolic process which is expected as amino acid metabolism has been linked to many cancers and metabolic reprogramming are such hallmarks of cancer they are being considered as treatment targets for cancer therapy18,19.

For the machine learning analysis, the cross-validation accuracies for the three models are shown in **Figure 6**. The cross-validation accuracies for the support vector machine and k-nearest neighbors models after feature selection are shown as well. The three models had similar performances as their cross-validation accuracies are comparable.



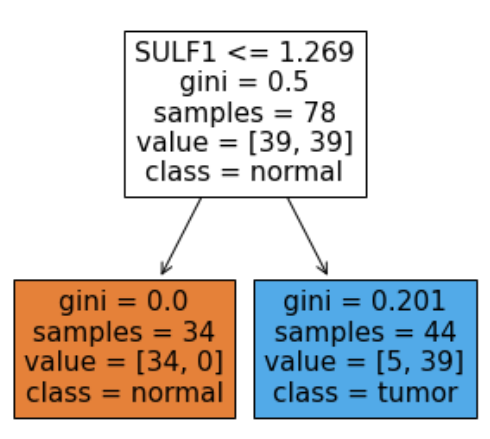
**Figure 7 Model Performance:** The model’s performance on the other 2 datasets. The accuracy, precision, and recall metrics are reported for each model and dataset.



**Figure 6 Machine Learning Model:** This figure demonstrates the cross-validation accuracies of the machine learning models for dataset 1, including the accuracies when the models are trained with 20 features from feature selection.

After cross validating and training all 3 models on dataset 1, models’ performance is tested on datasets 2 and 3, and the accuracy, precision, and recall metrics are reported in **Figure 7**. Each bar graph represents a dataset and the number of features that were used. ​From the graphs, all three models performed better on dataset 2 than dataset 3. Also, the support vector machine and k-nearest neighbors models performed just as well as the decision tree model when trained on the selected 20 features from feature selection. However, they performed worse when trained on all of the features. The three models had similar performances for a given dataset.

​The decision tree model is shown in **Figure 8** and it was found that cutting the decision tree at one level had the best performance across all three datasets. Furthermore, as noted in **Figure 8**, the SULF1-gene was the feature that was determined to be the best feature to split the tissue samples by malignancy20. This gene had one of the lowest p-values and one of the highest fold changes in dataset 1, which explains why the model might have chosen this gene to split the data. It is interesting to note that the overexpression of SULF1 has been found to be associated with PDAC20. Despite its known association with some patients with PDAC, it did not show up in the differentially expressed gene analysis earlier as it was not found to be significant in the third dataset.



**Figure 8 Decision Tree:** The decision tree model that is generated based on dataset 1. It shows that SULF1 was the selected gene/feature that the model used to split the samples into normal or tumor tissue.

# DISCUSSION

Of the 64 overlapping genes found in all 3 datasets, >66% have been identified and published in previous studies. Known biomarkers for PDAC such as CEACAM1, CEACAM5, CTNND2, and HOXC6 were captured in this analysis9. Many of these are precursors or regulators in the PI3K / Akt and RAS/MAPK signaling pathways which are hallmarks of pancreatic cancers 6,21. The machine learning models resulted in >85% cross validation accuracy for high volume samples. The machine learning models had similar performances at classifying malignant vs nonmalignant tissues and was good at classifying samples from different datasets. The feature selection was also found to improve the machine learning models indicating its importance when creating models.

Although the differentially expressed analysis yielded great results and the machine learning models were fairly accurate, there were several limitations to this dataset and improvements that can be made. There was high accuracy for GSE datasets with paired tumor vs normal sample data but the differentially expressed analysis may be skewed for unpaired, random datasets. The machine learning model parameters could be further optimized for improved accuracy, and it is recommended that future studies improve upon these parameters. The greatest limitation in this study, however, is due to the DAVID enrichment analysis. Although the overlapping genes have been shown in previous studies to be associated with PDAC, DAVID was unable to find additional pathways. Although DAVID is a continually updated database, it is incomplete and at times some user inputs cannot be mapped to any known genes in the DAVID Knowledgebase. It is thus recommended that future analyses use additional functional annotation tools such as GSEA, GOstat or Onto-Express to bridge the gap.

The most important follow up study for the analyses describe here would be to do a full meta-analysis including as much of the available data from PDAC microarray analyses. Including more datasets would yield more accurately determine the accuracy of the machine learning models and could be used to further refine the significant genes for PDAC. A better understanding of these genes could aid in the creation of new tests or therapeutic compounds for the treatment of PDAC.

# References

1. Sarantis, P., Koustas, E., Papadimitropoulou, A., Papavassiliou, A. G., & Karamouzis, M. V. (2020). Pancreatic ductal adenocarcinoma: Treatment hurdles, tumor microenvironment and immunotherapy. *World Journal of Gastrointestinal Oncology*, *12*(2), 173–181. https://doi.org/10.4251/WJGO.V12.I2.173

2. Wang, Y., Zhong, X., Zhou, L., Lu, J., Jiang, B., Liu, C., & Guo, J. (2020). Prognostic Biomarkers for Pancreatic Ductal Adenocarcinoma: An Umbrella Review. *Frontiers in Oncology*, *10*, 1466. https://doi.org/10.3389/FONC.2020.01466/BIBTEX

3. Hall, B. R., Cannon, A., Atri, P., Wichman, C. S., Smith, L. M., Ganti, A. K., Are, C., Sasson, A. R., Kumar, S., & Batra, S. K. (2018). *Advanced pancreatic cancer: a meta-analysis of clinical trials over thirty years*. www.oncotarget.com

4. Board, P. A. T. E. (2020). Pancreatic Cancer Treatment (Adult) (PDQ®). *PDQ Cancer Information Summaries*. https://www.ncbi.nlm.nih.gov/books/NBK65959/

5. Orth, M., Metzger, P., Gerum, S., Mayerle, J., Schneider, G., Belka, C., Schnurr, M., & Lauber, K. (2019). Pancreatic ductal adenocarcinoma: biological hallmarks, current status, and future perspectives of combined modality treatment approaches. *Radiation Oncology 2019 14:1*, *14*(1), 1–20. https://doi.org/10.1186/S13014-019-1345-6

6. Mehra, S., Deshpande, N., & Nagathihalli, N. (2021). Targeting PI3K Pathway in Pancreatic Ductal Adenocarcinoma: Rationale and Progress. *Cancers*, *13*(17). https://doi.org/10.3390/CANCERS13174434

7. Javadrashid, D., Baghbanzadeh, A., Derakhshani, A., Leone, P., Silvestris, N., Racanelli, V., Solimando, A. G., & Baradaran, B. (2021). Pancreatic Cancer Signaling Pathways, Genetic Alterations, and Tumor Microenvironment: The Barriers Affecting the Method of Treatment. *Biomedicines*, *9*(4). https://doi.org/10.3390/BIOMEDICINES9040373

8. Campagna, D., Cope, L., Lakkur, S. S., Henderson, C., Laheru, D., & Iacobuzio-Donahue, C. A. (2008). Gene Expression Profiles Associated with Advanced Pancreatic Cancer. *International Journal of Clinical and Experimental Pathology*, *1*(1), 32. /pmc/articles/PMC2480531/

9. Grützmann, R., Pilarsky, C., Ammerpohl, O., Lüttges, J., Böhme, A., Sipos, B., Foerder, M., Alldinger, I., Jahnke, B., Schackert, H. K., Kalthoff, H., Kremer, B., Klöppel, G., & Saeger, H. D. (2004). Gene Expression Profiling of Microdissected Pancreatic Ductal Carcinomas Using High-Density DNA Microarrays. *Neoplasia*, *6*(5), 611–622. https://doi.org/10.1593/NEO.04295

10. Badea, L., Herlea, V., Dima, S. O., Dumitrascu, T., & Popescu, I. (2008). Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepato-Gastroenterology*, *55*(88), 2016–2027.

11. Reddy, P. S., & Chandrasekar, M. (2022). PAD: A Pancreatic Cancer Detection based on Extracted Medical Data through Ensemble Methods in Machine Learning. *International Journal of Advanced Computer Science and Applications*, *13*(2), 149–156. https://doi.org/10.14569/IJACSA.2022.0130218

12. Pei, H., Li, L., Fridley, B. L., Jenkins, G. D., Kalari, K. R., Lingle, W., Petersen, G., Lou, Z., & Wang, L. (2009). FKBP51 Affects Cancer Cell Response to Chemotherapy by Negatively Regulating Akt. *Cancer Cell*, *16*(3), 259–266. https://doi.org/10.1016/j.ccr.2009.07.016

13. *GEO Accession viewer*. (n.d.). Retrieved June 7, 2022, from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22780

14. Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. https://doi.org/10.1038/NPROT.2008.211

15. Sherman, B. T., Hao, M., Qiu, J., Jiao, X., Baseler, M. W., Lane, H. C., Imamichi, T., & Chang, W. (2022). DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research*. https://doi.org/10.1093/NAR/GKAC194

16. Weissmueller, S., Manchado, E., Saborowski, M., Morris IV, J. P., Wagenblast, E., Davis, C. A., Moon, S. H., Pfister, N. T., Tschaharganeh, D. F., Kitzing, T., Aust, D., Markert, E. K., Wu, J., Grimmond, S. M., Pilarsky, C., Prives, C., Biankin, A. V., & Lowe, S. W. (2014). Mutant p53 drives pancreatic cancer metastasis through cell-autonomous PDGF receptor beta signaling. *Cell*, *157*(2), 382. https://doi.org/10.1016/J.CELL.2014.01.066

17. Janiszewska, M., Primi, M. C., & Izard, T. (2020). Cell adhesion in cancer: Beyond the migration of single cells. *The Journal of Biological Chemistry*, *295*(8), 2495. https://doi.org/10.1074/JBC.REV119.007759

18. Lukey, M. J., Katt, W. P., & Cerione, R. A. (2017). Targeting amino acid metabolism for cancer therapy. *Drug Discovery Today*, *22*(5), 796–804. https://doi.org/10.1016/J.DRUDIS.2016.12.003

19. Stock, C., & Schwab, A. (2015). Ion channels and transporters in metastasis. *Biochimica et Biophysica Acta - Biomembranes*, *1848*(10), 2638–2646. https://doi.org/10.1016/J.BBAMEM.2014.11.012

20. Hammond, E., Khurana, A., Shridhar, V., & Dredge, K. (2014). The Role of Heparanase and Sulfatases in the Modification of Heparan Sulfate Proteoglycans within the Tumor Microenvironment and Opportunities for Novel Cancer Therapeutics. *Frontiers in Oncology*, *4*. https://doi.org/10.3389/FONC.2014.00195

21. Pei, H., Li, L., Fridley, B. L., Jenkins, G. D., Kalari, K. R., Lingle, W., Petersen, G., Lou, Z., & Wang, L. (2009). FKBP51 affects cancer cell response to chemotherapy by negatively regulating Akt. *Cancer Cell*, *16*(3), 259–266. https://doi.org/10.1016/J.CCR.2009.07.016

1. [↑](#footnote-ref-2)