2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS) | 978-1-6654-7180-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/MASS56207.2022.00123

Spatial Analysis of Tumor Heterogeneity Using Machine Learning Techniques

Chancharik Mitra
EECS
University of California, Berkeley
Berkeley, CA
cmitra@berkeley.edu

Jin Young Yoo
Food Science and Human Nutrition
University of Illinois Urbana-Champaign
Champaign, IL
jyoo19@illinois.edu

Zeynep Madak-Erdogan
Food Science and Human Nutrition
University of Illinois Urbana-Champaign
Champaign, IL
zmadake2@illinois.edu

Aiman Soliman
National Center for Supercomputing Applications
University of Illinois Urbana-Champaign
Champaign, IL
asoliman@illinois.edu

***Abstract*—The treatment and study of cancer are in part hindered by cells and tissue of the same cancer type exhibiting differences from one another. This *tumor heterogeneity* is thus an important characteristic worth better understanding and analyzing. In the past, this analysis has been mostly carried out manually by clinicians and researchers. However, with advances in algorithms and computational resources, we can analyze tumor samples using statistical methods and machine learning techniques. Our work features an automated pipeline for analyzing the spatial gene expression of tumor tissue samples. For the task of segmenting tissue regions into tumor, non-tumor, and hepatocyte regions, our models (logistic regression, support vector machine, and random forest classifier) achieve over 90 percent accuracy on all tests. We find these results to be encouraging for future research in spatial analysis of tumor heterogeneity using similar methods.**

***Index Terms*—machine learning, computational biology, cancer, tumor heterogeneity, spatial gene expression**

I. Introduction

In the study and development of treatments for cancer, many obstacles present themselves. Different cancer types and potential for metastasizing are some of the notable challenges. Another roadblock in the analysis and treatment development of cancer is tumor heterogeneity. Tumor heterogeneity describes the property of tumor cells of the same type to exhibit differences between samples, tissues, patients, etc. [8], [14]. Therefore, it becomes imperative to have analysis methods for gaining insight into tumor heterogeneity and the cellular microenvironment of a tumor sample. The method we choose for analyzing heterogeneity is spatial transcriptomics.

The Central Dogma of biology gives us a structure for understanding how the genetic information stored in DNA is ultimately expressed through the creation of proteins [3]. Gene expression profiling is a method for better understanding this information flow by revealing which genes are expressed in a sample. This profiling is done by measuring the mRNA levels associated with each gene. Several methods for gene expression profiling exist. For this work, we choose the method of RNA sequencing which provides a profile of the entire transcriptome as opposed to methods like DNA microarrays which only provides a profile of predetermined genes [5]. DNA microarray analysis may provide valuable information as a follow up to RNA sequencing, but it is currently out of the scope of this paper.

While RNA sequencing can be done at the cellular level (scRNA-seq), we propose analyzing spatial gene expression profiles of cancer tumor tissue samples to gain greater insight into the tumor's cellular microenvironment. Spatial transcriptomics methods will allow for mapping the gene activity of an entire tissue sample, with the benefit being the preservation of positional context of cells in the tissue [6], [15]. In particular, we look to apply computational methods including machine learning and deep learning methods to analyze spatial gene expression profiles for gaining further insight into the tumor heterogeneity of *metastatic estrogen receptor positive breast cancer* samples.

II. Related Work

In a review done by Nawaz et. al., researchers establish many important ecological relationships between cancerous, noncancerous, and immune cells [10]. Interestingly, the method for these findings involves applying spatial analysis commonly performed at the macroenvironmental scale to a tumor's cellular microenvironment. The study notes that computer vision and data analysis techniques are key to analyzing predatory, mutualistic, commensal, and parasitic relationships.

One relevant example is a study the group performed investigating "the spatial distribution of cancer and immune cells in breast tumors" [10]. The group measured the co-localization of cancer and immune cells by using the Morista-Horn index, which is an index typically used to quantify the level of co-localization of species in a macroenvironmental context. The study ultimately found that "a high degree of co-localization between cancer and immune cells measured

NSF funded, award number 2050195

2155-6814/22/$31.00 ©2022 IEEE
DOI 10.1109/MASS56207.2022.00123

by this index was found to be significantly associated with increased probability of ten-year disease-specific survival in human epidermal growth factor receptor" [10]. This result is promising in that our work looks to similarly utilize spatially rich features in order to derive insight into tumor therapy resistance as well as other tumor characteristics. The review also referenced studies that uncovered two mutualistic relationships between different types of cancer cells, which were found to be beneficial for tumor growth and created an immunosuppressive microenvironment respectively. Similarly, in our research, we hope to leverage gene expression data with spatial context for a different purpose, to gain a better understanding of the tumor's microenvironment and potential resistance to therapy.

In another study, spatial transcriptomics methods were used to create a library of gene expression profile data for interpreting intra-tumor heterogeneity in prostate cancer [1]. The study manually identifies cancer foci regions using those gene expression gradients, rather than the typical method of using histological factors. We look to delve deeper and analyze similar data for estrogen receptor positive breast cancer using machine learning methods rather than manual methodologies.

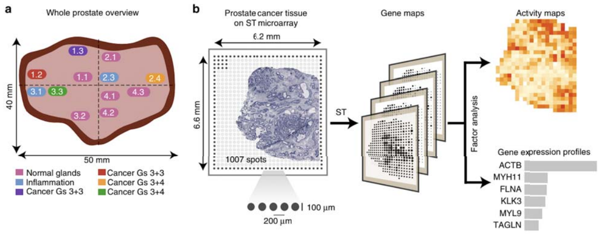


Fig. 1. This image comes from the prostate cancer study, and it provides a visualization of the gene expression maps that were utilized to delineate cancer foci. Image Citation: [1]

III. Method

(1) We use the process laid out in Zuo et. al. to prepare the fresh frozen tissue samples [15]. Our focus will be on the computational preparation of the raw gene expression data that results after sample preparation. The pipeline laid out in the following steps was applied specifically to one tissue sample, labeled F8_37. However, this automated pipeline can now easily be applied to new samples or batches of samples.

(2) Following its preparation, the sample is passed through the Visium 10X Genomics Spaceranger pipeline for spatial RNA sequencing of its gene expression profile. The spaceranger pipeline establishes a coordinate reference system (CRS) of circular Cartesian coordinates in order to perform its spatial analysis. A row-column as well as pixel representation of this CRS can be found in the tissue_positions_list.csv and scalefactors_json.json files in the spatial folder of the spaceranger outs (output folder). Each coordinate can also be identified by a unique barcode.

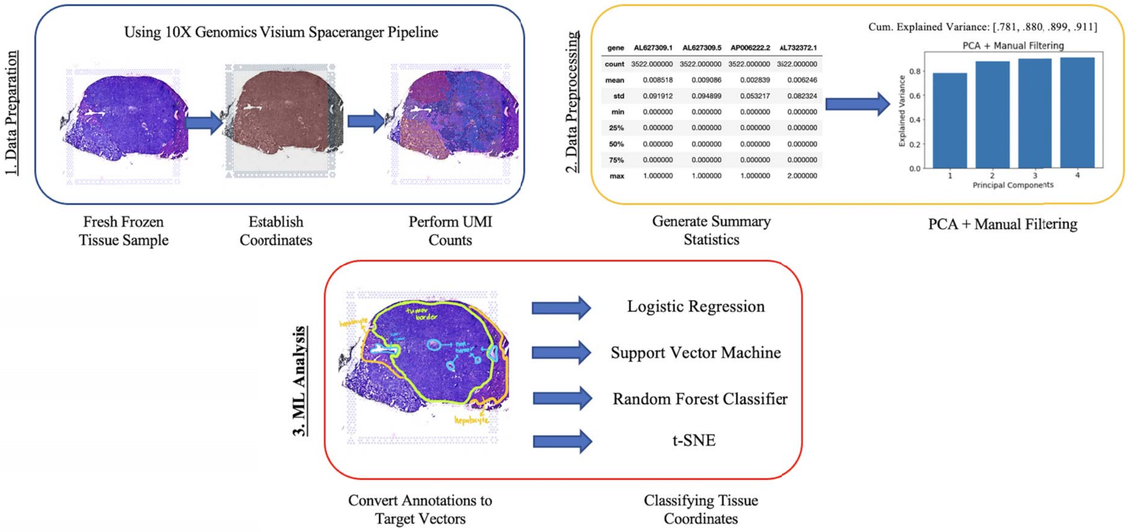
(3) The next step of the data preparation pipeline is to create the gene expression matrix upon which further analysis will be conducted. This was done by transforming the data in the filtered_feature_bc_matrix folder, which ultimately resulted in a (36601 x 3525) table consisting of coordinates (identified by their unique barcodes) as rows and gene expression for each coordinate measured by unique molecular identifier (UMI) count as columns (each column is represented by a gene). UMI count is a measure that represents the absolute gene transcript count, a metric which is calculated based on the absolute count of either RNA or DNA molecules that correspond to a gene [2], [5].

(4) Expression across the entire genome is often very sparse as most genes are either very minimally or not expressed in a small sample. To alleviate the issue of very sparse, high-dimensional data, we remove unexpressed genes from the dataset (i.e. genes that have zero UMI counts for all coordinates). For instance, 14060 of the 36601 reference genes were filtered out of the F8_37 sample. Then, to get a better sense of the remaining data, we calculate summary statistics (count, mean, standard deviation, min. val., 25% 50%, 75%, and max. val.) for each gene, revealing that much sparsity still existed.

(5) We address the dimensionality issue using PCA. We find that just 4 principal components can explain over 90 percent of the variation. Thus, we project the data onto these 4 principal components. Now, we have processed our data such that it is fit to be inputted into our machine learning models.

(6) We utilize the gene expression data to perform a segmentation/classification task. More specifically, each spatial coordinate acts as a data point, with the genes (in this case projected into four dimensions from PCA) acting as the features. At a high-level, we would like to learn useful representations of the genes that inform which coordinates are tumor, non-tumor, and hepatocyte. In order to create labeled training data, we annotate a tissue sample image, specifying regions as "tumor", "non-tumor", and "hepatocyte". Then, using 10X Genomics's Loupe Browser, this annotation can be converted into a label for every coordinate (i.e., data point). Thus, we now have target labels for training. We also prepare another set of labels for binary classification with just two classes "tumor" and "non-tumor" where the hepatocyte labels are subsumed into the non-tumor labels. (7) Equipped both with data and target labels, we perform the following analyses: t-SNE, logistic regression, support vector machines, and random forest classifiers. We perform t-SNE in two dimensions with no augmentation to the data. For the remaining supervised machine learning methods, we perform an 80%-20% train-test split respectively. We follow common convention and standardize the data with zero mean and unit variance for logistic regression and support vector machines, while leaving the data unstandardized for random forest classifiers.

After training the models on both labeled tasks (binary and ternary classification), we collect accuracy scores for predictions on the test set. In addition, we generate confusion matrices for all experiments to get a more granular sense of performance on the tasks.



gene	AL627309.1	AL627309.5	AP006222.2	AL732372.1
count	3522.000000	3522.000000	3522.000000	3522.000000
mean	0.008518	0.009086	0.002839	0.006246
std	0.091912	0.094899	0.053217	0.082324
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	2.000000

Fig. 2. An overview of the method of our work, from data preparation to machine learning analysis

IV. Results

We now discuss the results of our machine learning analysis on the spatial gene expression data.

A. t-SNE:

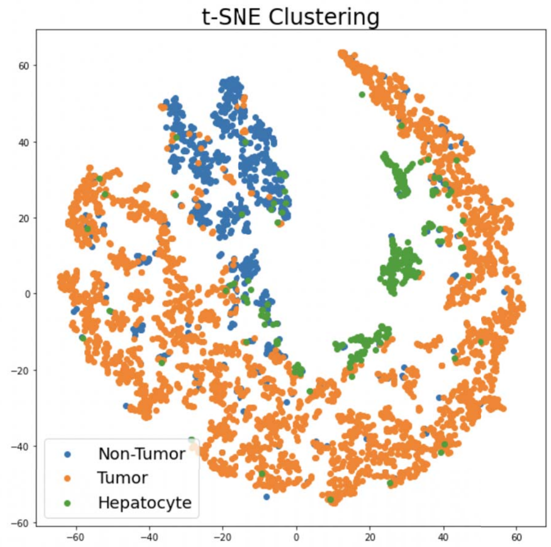
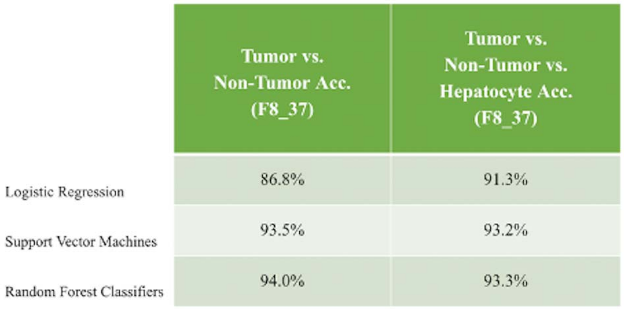


Fig. 3. 2-dimensional t-SNE plot of F8_37 gene expression data colored by target label

At a high level, t-SNE is a dimensionality reduction technique that tries to preserve high-dimensional distances between data points in the low-dimensional representation. Although this means the clustering is not semantic, t-SNE gives us a good idea of the separation between different labels. In this case (Fig. 3), we note there is good intra-class clustering as well as inter-class separation, indicating that the other machine learning methods will likely learn effective decision rules for classifying/segmenting these points/regions.

B. Accuracy of Classifiers:

Although the core task of the machine learning models is to classify coordinates as "tumor" vs "non-tumor" in the binary case and "tumor" vs "non-tumor" vs "hepatocyte" in the ternary case, this classification can also be considered a segmentation of the tissue sample into these regions as these coordinates retain spatial context of the tissue sample.



	Tumor vs. Non-Tumor Acc. (F8_37)	Tumor vs. Non-Tumor vs. Hepatocyte Acc. (F8_37)
Logistic Regression	86.8%	91.3%
Support Vector Machines	93.5%	93.2%
Random Forest Classifiers	94.0%	93.3%

Fig. 4. Accuracy scores for all machine learning classifier methods (F8_37 tissue sample)

Overall, we see very high accuracy scores on all tasks, with accuracy exceeding 90% in most cases. Comparing the accuracy between different types of models, we see that SVMs and random forest classifiers (RFCs) achieved similar accuracy scores, both outperforming the logistic regression model. This matches with our expectations given that logistic regression is a simpler model than SVMs or RFCs. High accuracy scores are encouraging for future work as well as clinical and research applications. However, it is also important to evaluate these models utilizing other metrics.

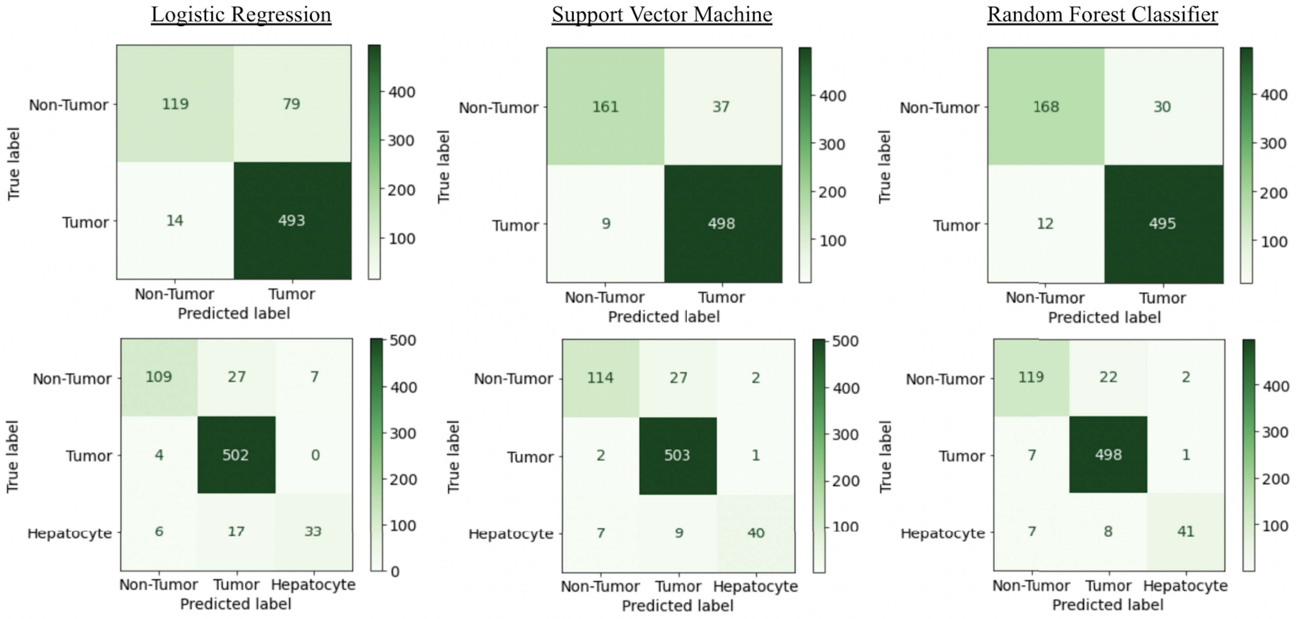


Fig. 5. Confusion matrices for all classification methods (F8_37 tissue sample)

C. *Confusion Matrices:*

The confusion matrices (Fig. 5) provide more granular insight that better contextualize the accuracy scores. There are a few important observations to make. In the binary case, we find that while logistic regression is effective in identifying tumor coordinates, it is less effective than the other methods at correctly classifying non-tumor coordinates. This holds true for the ternary case as well.

While RFCs obtained higher accuracy scores than SVMs, there are several nuances that are recovered from the confusion matrices. First, we note that the SVM method had a higher true positive rate for classifying tumors in the binary and ternary cases. Additionally, SVMs have a lower false-negative rate (i.e., classifying tumor coordinates incorrectly as either non-tumor or hepatocyte). Even though the difference is small, such a characteristic can be an potential reason for choosing SVMs over RFCs in clinical settings.

D. *Interpretability:*

One last, yet important, evaluation method is interpretability. In biological and medical-focused research, model interpretability is highly desirable. This is true largely because a diagnosis or treatment based on an ML model can be better trusted if the decision is interpretable. In this regard, RFCs are preferrable over logistic regression and SVMs because an RFC's optimized decision rule can be easily viewed as a tree structure of classification decisions. More about interpretability will be covered in the future work section.

V. Conclusion

In this research, we demonstrate that spatial gene expression data of tumor tissue samples can be used to analyze tumor heterogeneity. The following summarizes our findings and novel outcomes:

(i) We preprocess spatial gene expression data by filtering out unexpressed genes and performing PCA to compress the data.

(ii) We develop an automated pipeline for classifying spatial coordinates (which also translates to segmenting the tissue sample) as tumor, non-tumor, or hepatocyte using gene expression data. We achieve high performance on these tasks achieving over 90 percent accuracy for most cases. This type of classification can segment regions of a tissue sample, allowing for easier analysis of therapy resistance and other tumor characteristics.

(iii) We discuss and present benefits and drawbacks of each model, considering raw accuracy scores, confusion matrices, and interpretability.

These results provide an encouraging foundation for future research in analysis of tumor heterogeneity using machine learning methods.

VI. Future Work

We see three promising directions in which our future work on analyzing tumor heterogeneity can proceed. All of which will build off of the work done in this study.

A. *Further Interpretability Analysis:*

As we mention in our results section, the interpretability of RFCs is a desirable quality for a model to have in biological/medical research. In future work, we plan to leverage this property for greater insight by analyzing the resulting decision tree produced by our RFC model. Using this decision tree, we can gain further understanding of tumor properties like therapy resistance. One consideration for this work is the fact that we utilized PCA-compressed data as the input. Thus, the RFC decision rule will be in terms of our PCA vectors, so inverse PCA would need to be applied.

As a small demonstration of the value of interpretability analysis, we identified the highest contributing gene of the first principal component to be RPL41. There exists literature that suggests RPL41 acts as a tumor suppressor [13]. We expect interpretability analysis such as this will aid future research and clinical applications.

B. Spatial Feature Engineering:

In this work, we based our analysis solely on spatial gene expression. However, in future work we hope to leverage other spatial features which we will engineer based on biological principles. We provide a simple example:

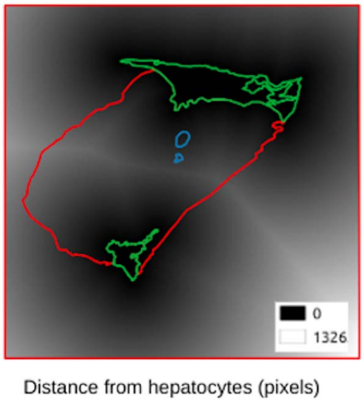


Fig. 6. Sample spatial feature of F8_38 sample

This proximity map of the F8_38 sample measures the distance of all points from hepatocyte regions and is created using QGIS geospatial tools. Thus, we can develop these spatial features and incorporate them into our input data. We expect that these features will improve the performance of ML models as well as providing greater insight into a sample's tumor heterogeneity and cellular microenvironment.

C. Deep Learning Methods:

Finally, we expect deep learning models to be a promising avenue for future work. Deep learning methods requires less manual feature engineering and fewer assumptions about the feature space in order to learn informative representations. For this reason, we expect deep learning models to provide unique insights about tumor heterogeneity compared to those from our ML methods.

Some architectures we plan to utilize include the UNET, DINO, and transformer-based architectures for segmenting the tissue sample [7], [9].

Another part of the pipeline that can be improved using deep learning is data compression. Using autoencoder methods, we can compress very high-dimensional data into semantically meaningful low dimensional encodings. In the example visualization below, the input and reconstructed output would be our gene expression in addition to any other engineered spatial features. We would like to take advantage of the dense semantic representation in the middle, using it as input for our models.

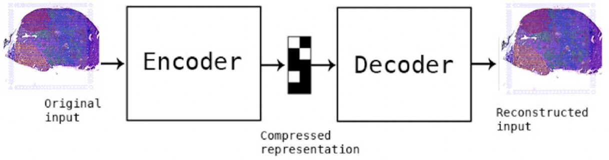


Fig. 7. A sample visualization of an autencoder compressing gene expression data. Image adapted from following source: [12]

In practice, we would apply this transformation on tabular data, but there is potential for applying a convolutional version of this method to preserve relational information of the pixels/coordinates.

Acknowledgement

We greatly appreciate the insight and advice of Priyam Mazumdar. We would also like to thank Olena Kindratenko and Dr. Volodymyr Kindratenko, the coordinators of the REU-FoDOMMaT program. The REU-FoDOMMaT program is supported by NSF funding, award number 2050195.

References

[1] Berglund, E., Maaskola, J., Schultz, N. et al. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. Nat Commun 9, 2419 (2018). https://doi.org/10.1038/s41467-018-04724-5

[2] Chen, W., Li, Y., Easton, J. et al. UMI-count modeling and differential expression analysis for single-cell RNA sequencing. Genome Biol 19, 70 (2018). https://doi.org/10.1186/s13059-018-1438-9

[3] Crick C (1970) Central dogma of molecular biology. Nature 227:561–563

[4] Fu GK, Hu J, Wang PH, Fodor SP. Counting individual DNA molecules by the stochastic attachment of diverse labels. Proc Natl Acad Sci USA. 2011 May 31;108(22):9026-31. doi: 10.1073/pnas.1017621108. Epub 2011 May 11. PMID: 21562209; PMCID: PMC3107322.

[5] Hurd PJ, Nelson CJ (2009) Advantages of next-generation sequencing versus the microarray in epigenetic research. Brief Funct Genomic and Proteomic 8(3):174–183. doi: 10.1093/bfgp/elp013.

[6] Marx, V. Method of the Year: spatially resolved transcriptomics. Nat Methods 18, 9–14 (2021). https://doi.org/10.1038/s41592-020-01033-y

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve J ´ egou, Julien ´ Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In ICCV, 2021.

[8] NCI Dictionary of Cancer terms. National Cancer Institute. (n.d.). Retrieved July 14, 2022, from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/tumor-heterogeneity

[9] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

[10] Sidra Nawaz, Yinyin Yuan, Computational pathology: Exploring the spatial dimension of tumor ecology, Cancer Letters, Volume 380, Issue 1, 2016, Pages 296-303, ISSN 0304-3835, https://doi.org/10.1016/j.canlet.2015.11.018.

[11] Spatial transcriptomics. 10x Genomics. (n.d.). Retrieved July 26, 2022, from https://www.10xgenomics.com/spatial-transcriptomics

[12] Trencseni, M. (2021, March 17). Building a pytorch Autoencoder for mnist digits. Bytepawn. Retrieved July 26, 2022, from https://bytepawn.com/building-a-pytorch-autoencoder-for-mnist-digits.html

[13] Wang S, Huang J, He J, Wang A, Xu S, Huang SF, Xiao S. RPL41, a small ribosomal peptide deregulated in tumors, is essential for mitosis and centrosome integrity. Neoplasia. 2010 Mar;12(3):284-93. doi: 10.1593/neo.91610. PMID: 20234822; PMCID: PMC2838445.

[14] Welch DR. Tumor Heterogeneity–A 'Contemporary Concept' Founded on Historical Insights and Predictions. Cancer Res. 2016 Jan 1;76(1):4-6. doi: 10.1158/0008-5472.CAN-15-3024. Epub 2016 Jan 3. PMID: 26729788; PMCID: PMC4773023.

[15] Zuo Q, Mogol AN, Liu YJ, Santaliz Casiano A, Chien C, Drnevich J, Imir OB, Kulkoyluoglu-Cotul E, Park NH, Shapiro DJ, Park BH, Ziegler Y, Katzenellenbogen BS, Aranda E, O'Neill JD, Raghavendra AS, Tripathy D, Madak Erdogan Z. Targeting Metabolic Adaptations in the Breast Cancer-Liver Metastatic Niche Using Dietary Approaches to Improve Endocrine Therapy Efficacy. Mol Cancer Res. 2022 Jun 3;20(6):923-937. doi: 10.1158/1541-7786.MCR-21-0781. PMID: 35259269; PMCID: PMC9177734.