# MASS SPECTROMETRY IMAGING IN DETECTING TUMOR HETEROGENEITY

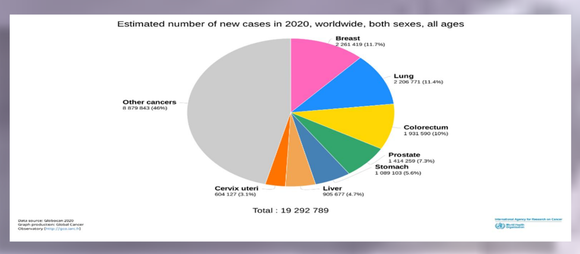

## Abstract

**T**umor subpopulations have molecular phenotypes that drive tumor progression and determine disease outcome which is essential for a more personalized therapy. Mass spectrometry imaging has proven its ability to identify diagnostic and prognostic biomarkers. In this research, we seek to determine tumor subpopulations that affect patient outcomes and the statistically associated subpopulations with poor survival and tumor metastasis. Here we introduce spatially mapped t-distributed stochastic neighbor embedding (t-SNE), a nonlinear visualization of the data that is able to better resolve the biomolecular intratumor heterogeneity. The outcomes will allow us to uncover subpopulations statistically associated with patient survival in primary tumors of gastric cancer and with metastasis in primary tumors of breast cancer.


## Introduction

**M**SI can uncover molecular intratumor heterogeneity. The challenge has been to identify those tumor subpopulations that drive patient outcomes within the highly complex datasets (hyperdimensional data, intratumor heterogeneity, and patient variation). Here we report an automatic, unbiased pipeline to nonlinearly map the hyperdimensional data into a 3D space, and identify molecularly distinct, clinically relevant tumor subpopulations. We demonstrate this pipeline's ability to uncover subpopulations statistically associated with patient survival in primary tumors of gastric cancer and with metastasis in primary tumors of breast cancer.

## Motivation




**Diagnosis:**

**D**iscovering new techniques as the cancer itself are developing with time and be more cruel than previous. MSI shows many things that the microscopic level could not determine and that helps us in different aspects in the cancer field.

**Prognosis:**

**P**rognosis is simply a follow-up approach, but in an expert way. For cancerous patients, it is very important to periodically check up the patient's status. One of the techniques used to assure this approach is "Survival Analysis".

## Dimensionality Reduction (t-SNE)

**A**s we were facing a problem in our data known as "short, fat data problem" which means the number of features "proteins" are larger than the number of samples so, the features are redundant or irrelevant and can thus be removed without incurring much loss of information. We used a non-linear dimensionality reduction technique called t-SNE to select some relevant features from the variables and predictors to construct our model.

## Clustering (K-means)

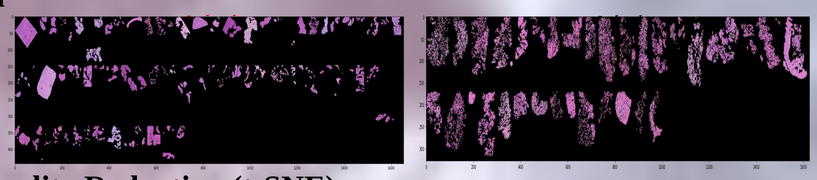
As the colors in the t-SNE image are continuous and close colors represents close datapoints in the t-SNE space we used k-means clustering to color the datapoints according to the cluster it belongs to as each cluster represents a discrete color so, we can determine the significance between clusters using statistical analysis methods.

## Statistical Analysis

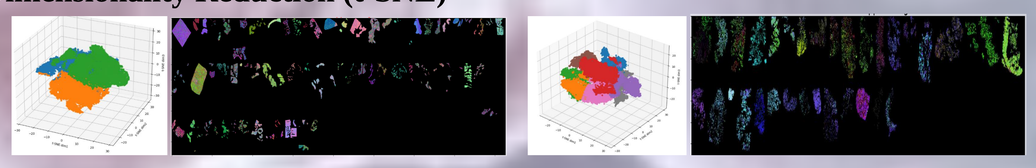
We used two techniques as a statistical analysis which are "Kaplan-Meier Curves" for survival rates and "Fisher's Exact Test" for building the contingency tables. It is known as (time to event analysis) that describes some methods for analyzing the length of time till a well-defined end point of interest happens (The point of interest in cancer is death indeed).

## Results

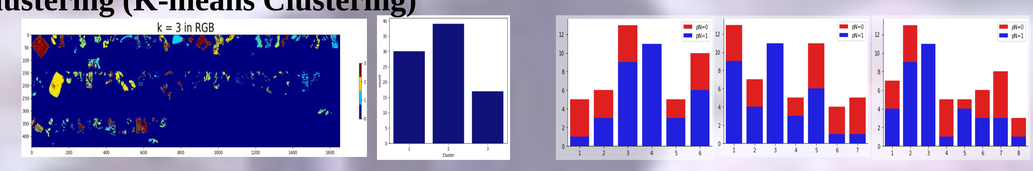
**MSI Data**



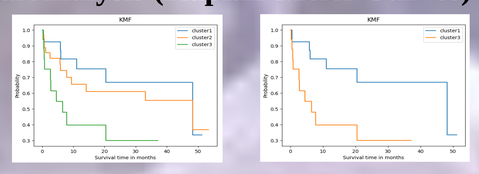
**Dimensionality Reduction (t-SNE)**



**Clustering (K-means Clustering)**




**Survival Analysis (Kaplan-Meier Curves)**

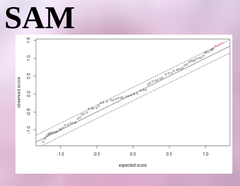






| K Values | P-value |
|---|---|
| 6 | 0.004824 |
| 7 | 0.005738 |
| 8 | 0.003732 |

**SAM**




## Discussion

**MSI Data**

HE Image is a remarkable tissue section. The recognition of the tumor subpopulations that affect the results of patients is important for better describing the changes in molecules. We used MSI because it has an ability to detect tumor subpopulations in histologically identical regions of tumor tissue.

**Dimensionality Reduction (t-SNE)**

It preserves the global and local similarity structure of the dataspace in the low dimensional representation. We can see samples representing two types of tumor heterogeneity; inter/intratumor heterogeneity of MSI Data. And contribution of the molecular heterogeneity in Data due to intratumor heterogeneity and patient variability.

**Clustering (K-means Clustering)**

Gastric Data; blue color: background, others: cluster. We have a total of 79 patients which is more than our number of patients and that shows us that some patients are assigned to more than one cluster. Breast Data; Cluster 3 is full of metastatic patients only.

**Survival Analysis (Kaplan-Meier Curves)**

The greatest significant difference in survival between the subpopulations in clusters 1 and 3 with P-value of 0.02 less than 0.05.

**SAM**

The significance level of SAM output could be found at the high end of SAM plot revealing the expected score along the line of interest.

## SAM Analysis

**I**t is a statistical technique for finding significant genes in a set of microarray experiments. The input is gene expression measurements from a set of microarray experiments, as well as a response variable from each experiment. It computes a statistic $(d_i)$ for each gene $(i)$, measuring the strength of the relationship between gene expression and the response variable. It uses repeated permutations of the data to determine if the expression of any genes is significantly related to the response.

## Conclusion

**I**ntratumor heterogeneity is a key factor in tumor progression, affecting patient outcomes and treatment. Tumor subpopulations can be histologically indistinguishable but still have molecular phenotypes that drive tumor progression and determine disease outcome. The identification of these clinically relevant tumor subpopulations is of utmost importance for understanding cancer development and the management of cancer patients. Although localized genomic techniques have established branched evolution of tumors and single-cell transcriptional heterogeneity, the cost and throughput of these techniques are prohibitive for large scale multi-site sequencing of patient tissues. The automated identification of phenotypic tumor subpopulations reported here will allow better targeting of these powerful genomic methods to those subpopulations that are statistically associated with patient outcomes.

## References


[1] Buchberger, A. R., DeLaney, K., Johnson, J., & Li, L. (2018). Mass spectrometry imaging: A review of emerging advancements and future insights. Analytical Chemistry, 90(1), 240–265.

[2] L. v. d. Maaten, "Visualizing non-metric similarities in multiple maps," Springerlink, vol. 23, 2011.

[3] Klein, John P., and Melvin L. Moeschberger. Survival analysis: techniques for censored and truncated data. Vol. 1230. New York: Springer, 2003.


## Acknowledgments


**F**irst, we want to say "Thank You" for Dr/ Walid Abd Elmoula for his time, assistance, and supervision of us along the past 10 months and for offering a suitable materials, resources, and opportunities reached to him for us. We also want to say the same thing for Dr/ Ahmed Morsy; Both showed honest and appreciated work.

**S**econd, we are so proud reaching this moment and with our love and appreciation, we want to thank each doctor/TA/person in Cairo University - Faculty of Engineering (CUFE) who helped us achieve this. It would be difficult without you. So, thank you is not enough, but totally respected.



Ibrahim Elsayed
Donia Abd Elslam
Renad Taher
Mariem Ahmed
Mustafa Yehia