Mass Spectrometry Imaging in Detecting Tumor Heterogeneity

Ibrahim Elsayed#1, Donia Abd Elsalam#2, Renad Taher#3, Mariem Ahmed#4, and Mustafa Yehia#5

[1hemasayed600@gmail.com](mailto:1hemasayed600@gmail.com) [2donia.199887@gmail.com](mailto:2donia.199887@gmail.com)

[3renad.taher12@gmail.com](mailto:3renad.taher12@gmail.com)

[4mariem.ahmed.1608@gmail.com](mailto:4mariem.ahmed.1608@gmail.com)

5[mustafayehia4@gmail.com](mailto:mustafayehia4@gmail.com)

***Abstract*— Tumor 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 can 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.**

***Keywords*— Mass spectrometry imaging, t-SNE, intratumor heterogeneity, metastasis**

1. INTRODUCTION

**C**ancer is considered as one of the most important topics from the 20th century till now and needs a lot of attention, it is mainly an abnormal growth of cells. According to World Health Organization (WHO) 2020 cancer statistics (Fig. 1), it reflects our interest in both breast and gastric cancers. So, we are going to proceed with exploring causes of cancers, ways to discover/treat it and finally moving forward to a new approach called MSI to discover a lot of wonderful things about the field of cancer.

MSI 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. [1] 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.

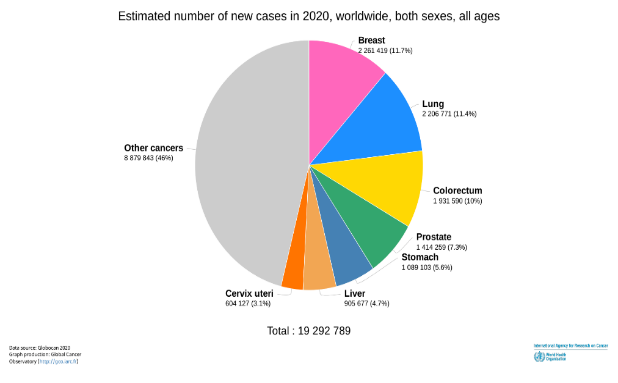


Fig. 1 WHO 2020 Cancer Statistics

1. MATERIALS AND METHODS

We start this section by identifying some important notations and show the problems we faced and how to solve them,

*A. Dataset*

We define some notations used here; gastric cancer data is ‘Gastric Data’, and breast cancer data is ‘Breast Data’. These two types could be divided from one category, which is ‘MSI Data’ as described previously. Our MSI data could be considered as pixel arrays that are high-dimensional data. There are many properties of this type of data that could be defined or explained by the fact ‘units that are partially redundant’. The redundancy in high-dimensional data means that there are parameters or features that can characterize different units are dependent on each other. From that, we must understand all units in data that require taking the redundancy into account. Gastric Data is a pixel array, or a group of gastric samples taken from *63* patients who had gastric cancer. and Breast Data is a group of breast samples taken from *32* patients, who are in differently risky stages. So, we deal with data in high dimension space and that would cause me a problem known as “Curse of Dimensionality” that can be solved using “Dimensionality Reduction” techniques.

*B. Dimensionality Reduction*

Curse of Dimensionality occurs when analyzing and organizing the data in high-dimensional space and does not occur in low-dimensional space. You can face this problem in different domains such as numerical analysis, machine learning like in our work, data mining and databases. Commonly these problems occur when the dimension of data increases, the volume of the data increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance. As a goal for that is to obtain the reliable and statistically sound result; Make the amount of data grows exponentially with increasing of the dimensionality. But it is not enough, you need to organize and search data that relies on detecting areas with similar properties. In high dimensional data, all objects appear to be sparse and dissimilar in many ways, which prevents common data organization strategies from being efficient. To solve this problem, we can use linear (e.g., PCA) or non-linear (e.g., t-SNE) dimensionality reduction technique.

*1) Principal Component Analysis (PCA)*

It is a technique in modern data analysis like in diverse the fields from neuroscience to computer graphics. That is a simple, non-parametric method, which is used to extract information from confusing datasets. We can use it to find a linear projection of our dimensional data, in such a way that the variance of the projected data is maximized. It provides a roadmap to reduce complex datasets and mainly concerned with preserving large pairwise distances in the map. [2]

*2) t-distributed Stochastic Neighborhood Embedding (t-SNE)*

Fig. 2 Microscopic Image of 63 Gastric Cancer Patient

PCA wants to make sure that stuff that is dis-similar, that ends up for apart like the zeros and ones. But there is one important question; Is that really what we want in visualization? Particularly, are those large pairwise instances in the data, are those things that are reliable? The answer is no, because if you think about data in terms of nonlinear manifold, to switch role here, then you will see that in this case the Euclidean distance between two points on this manifold would not reflect very well their similarity. Because the distance between these two points is similar, whereas if you would consider the entire structure of the data. (Swiss-role Problem).

t-SNE is an algorithm in ML used to visualize high-dimensional data that depends on stochastic neighbor embedding. For example, we can use it to visualize 30 features, where the features are intensity vectors that represents images. It can be used to represent word-count vectors or documents have thousands of dimensions. It is like the locally linear technique, but it solves the problem of collapse all points onto a single point. And solves the problem of visualize the real high-dimensional data with preserving the local and global structure of data in a single map. [3]

*C. Clustering*

A motivation: Clustering is an important technique for “Unsupervised Data” and our data is a simple example for that, but that is not the only reason that makes us turns our mind to this technique. The answer is for the following; As the colors in the t-SNE image are continuous and close colors represents close datapoints in the t-SNE space we used clustering to color the datapoints according to the cluster it belongs to as each cluster represents a discrete color so, we can determine the clinical outcome of each cluster.

*1) K-means Clustering*

It is an unsupervised learning algorithm which splits unlabeled datasets into number of clusters based on their properties such that datapoints with similar properties will be in the same cluster. K random centroids were initiated first in the data, where k is the number of clusters chosen. let us choose k=2 then we will have 2 initial centroids. Then assigning every datapoint to its nearest centroid by calculating the Euclidean Distance (ED). Then moving every centroid to the average of the datapoints assigned to it by calculating the center of gravity and repeating these steps until the position of the centroids is not changing. [4] We apply clustering on t-SNE to have a better look on intratumor heterogeneity and detect the stages of tumor in Gastric Data and metastatic level in Breast Data.

Fig. 3 Microscopic Image of 32 Breast Cancer Patient

*D. Statistical Analysis*

We used statistical approaches to study the survival status in Gastric Data and the metastatic level in Breast Data to see the significantly in our Data using “Kaplan-Meier Curves” and “Fisher’s Exact Test”.

*E. Significance Analysis of Microarrays (SAM)*

It 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 (Table 4.4.1). It computes a statistic for each gene , 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. The cutoff for significance is determined by a tuning parameter delta, chosen by the user based on False Positive Rate (FPR) [5]. One can also choose a fold change parameter, to ensure that called genes change at least a pre-specified amount.

1. RESULTS

*A. 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 (Fig. 2 & Fig. 3).

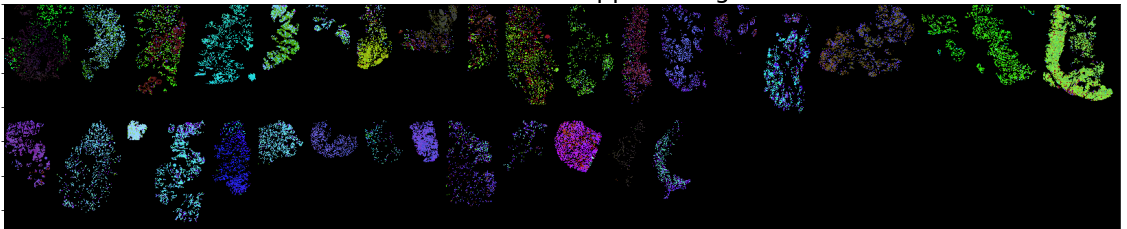
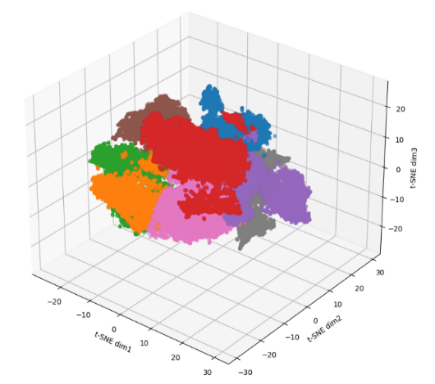


Fig. 5 Breast Data t-SNE Scatter Space (Left) & t-SNE Spatial Image (Right)

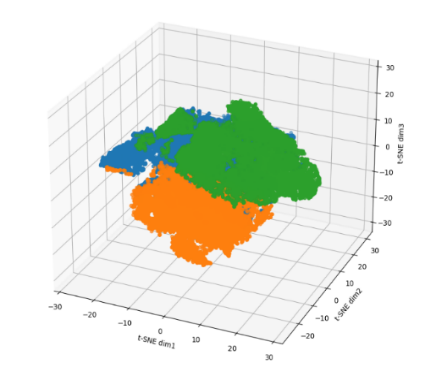


Fig. 4 Gastric Data t-SNE Scatter Space (Left) & t-SNE Spatial Image (Right)

*B. 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 (Fig. 4 & Fig. 5).

*C. 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 (Fig. 6). Breast Data; Cluster 3 is full of metastatic patients only (Fig. 7).

*D. Statistical Analysis*

The greatest significant difference in survival between the subpopulations in clusters 1 and 3 with P-value of 0.02 less than 0.05 (Fig. 8).

*E. Significance Analysis of Microarrays (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 (Fig. 9).

1. DISCUSSION

Identification of the tumor subpopulations that impact patient outcomes is essential for better characterizing the molecular changes that accompany tumor development and for optimizing patient management. MSI has several key characteristics that make it well suited to this task; it is an untargeted analysis that can simultaneously analyze hundreds of molecular ions, it can be directly applied to tissue sections, it is inexpensive, and it is fast. Several previous studies have reported MSI’s ability to uncover tumor subpopulations in histologically identical regions of tumor tissue. Here we have used dimensionality reduction based on t-SNE followed by bisecting k-means clustering to automatically segment the tumor-specific MSI data from a patient series into an optimum number of subpopulations. We used t-SNE because it is a nonlinear mapping technique that preserves the local and global similarity structure of the dataspace in a lower dimensionality representation. t-SNE has previously been shown to be a superior representation for the spatial organization of MSI and gene expression data. We reasoned that t-SNE’s nonlinear nature would also better equip it to distinguish the mass spectrometry differences between tumor subpopulations. Mapping of the tumor-specific MSI data from patients with gastric and breast cancer into the lower-dimensional t-SNE space revealed structured 3D data spaces.

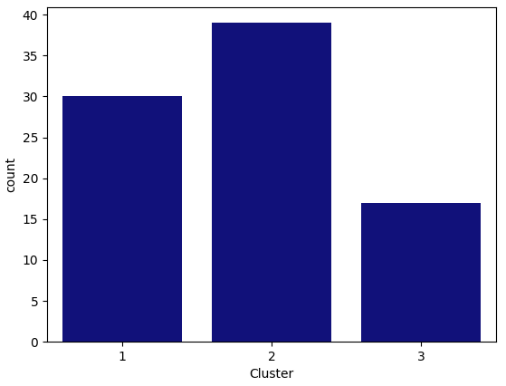
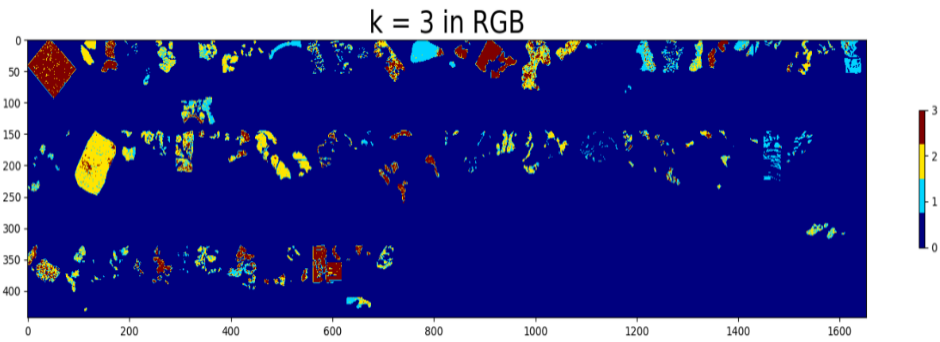


Fig. 6 Gastric Data K-means Spatial Image (Left) & Count Plot of Patients in each Cluster (Right)

1. CONCLUSION

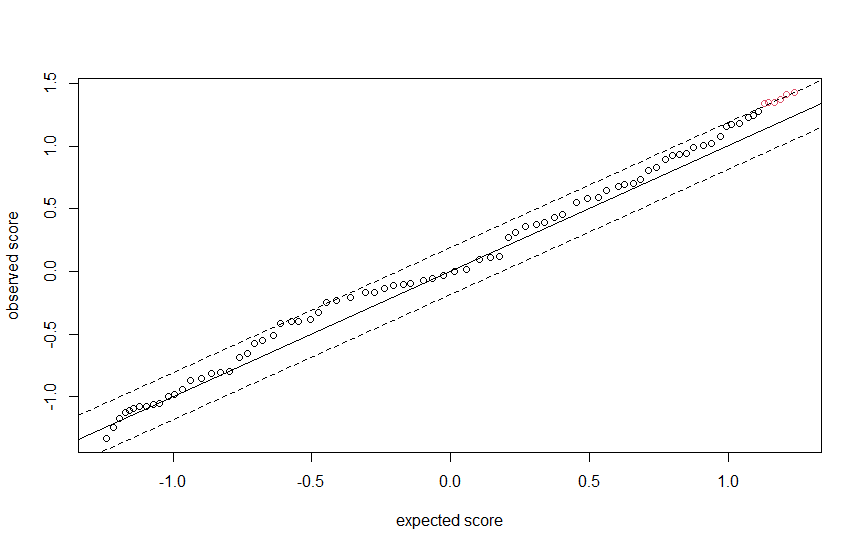
Intratumor heterogeneity is a key factor in tumor progression, affecting patient outcomes and treatment. Tumor subpopulations can be histologically indistinguishable but still have molecular phenotypes that drive tumor progression and determine disease outcome. The identification of these clinically 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.

Fig. 7 Breast Data K-means Count Plot with Metastasis Colored , &

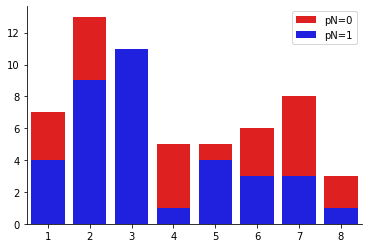
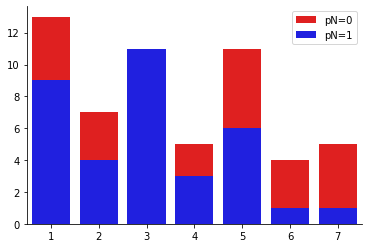
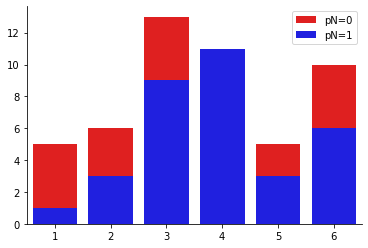


Fig. 9 SAM applied on Gastric Data

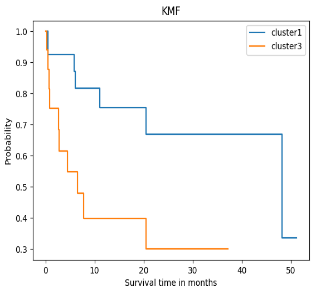
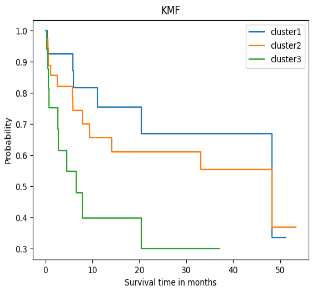


Fig. 8 Gastric Data K-M Curve (Left) & Significance between Cluster1 and 3 (Right)

1. FUTURE WORK

We shall now establish a technique to distinguish between biomarkers we got from SAM analysis, and to do that we have first to classify them using any classification technique (e.g., KNN) and then we will have to differ our data (As explained previously, our data is relatively small, but significant) so, we will use a technique known as Leave-One-Patient-Out (LOPO) and apply the same procedures on data extracted. That will give us a strong evidence for our result and solve the problem of small data.

ACKNOWLEDGMENTS

First, 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.

Second, 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.

REFERENCES

[1] Abdelmoula, Walid & Balluff, Benjamin & Englert, Sonja & Dijkstra, Jouke & Reinders, Marcel & Walch, Axel & Mcdonnell, Liam & Lelieveldt, Boudewijn. (2016). Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data. Proceedings of the National Academy of Sciences. 113. 10.1073/pnas.1510227113.

[2] J. Cadima, "Principal component analysis: a review and recent developments," THE ROYAL SOCIETY PUBLISHING, vol. 16, 2016.

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

[4] Sinaga, Kristina & Yang, Miin-Shen. (2020). Unsupervised K-Means Clustering Algorithm. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.2988796.

[5] Chu, G., Seo, M., Li, J., Narasimhan, B., Tibshirani, R., & Tusher, V. (n.d.). Users guide and technical document. Retrieved July 18, 2021, from Stanford.edu website: <http://statweb.stanford.edu/~tibs/SAM/sam.pdf>