2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

Pseudotime Based Discovery of Breast Cancer Heterogeneity

Tasmia Aqila
Computing and Information Sciences
Florida International University
Miami, FL, USA
taqil001@fiu.edu

Abdullah Al Mamun
Computing and Information Sciences
Florida International University
Miami, FL, USA
mmamu009@fiu.edu

Ananda Mohan Mondal
Computing and Information Sciences
Florida International University
Miami, FL, USA
amondal@fiu.edu*

***Abstract*—Breast cancer is highly sporadic and heterogeneous in nature. Even the patients with same clinical stage do not cluster together in terms of genomic profiles such as mRNA expression. In order to prevent and cure breast cancer completely, it is essential to decipher the detailed heterogeneity of breast cancer at genomic level. Putting the cancer patients on a time scale, which represents the trajectory of cancer development, may help discover the detailed heterogeneity. This in turn would help establish the mechanisms for prevention and complete cure of breast cancer. The goal of this study is to discover the heterogeneity of breast cancer by ordering the cancer patients using pseudotime. This is achieved through two objectives: *First*, a computational framework is developed to place the cancer patients on a time scale, meaning construct a trajectory of cancer development, by inferring pseudotime from static mRNA expression data; *Second*, discovering breast cancer heterogeneity at different time periods of the trajectory using statistical and machine learning techniques.**

In this study, the trajectory of breast cancer progression was constructed using static mRNA expression profiles of 1072 breast cancer patients by inferring pseudotime. Three sets of key genes discovered using supervised machine learning techniques are used to develop the trajectories. The first set of genes are PAM50 genes which is available in literature. The second and third sets of genes were discovered in the present study using the clinical stages of breast cancer (Stage-I, Stage-II, Stage-III, and Stage-IV). The proposed computational framework has the capability of deciphering heterogeneity in breast cancer at a granular level. The results also show the existence of multiple parallel trajectories at different time periods of cancer development or progression.

Keywords—Breast cancer heterogeneity, mRNA expression, pseudotime, t-SNE, Trajectory of cancer development.

I. INTRODUCTION

The Breast Cancer Landscape published by the Department of Defense reported that 2.1 million women were diagnosed with the disease worldwide in 2018 which accounts for nearly a quarter of all cancers in women [1]. It is estimated that 268,600 new cases will be diagnosed and 41,760 breast cancer deaths would occur in 2019 in the United States alone [2]. Consequently, there is an overarching need to unearth the factors that will help identify the heterogeneity in breast cancer. According to DoD Breast Cancer Landscape, it is well established that there are several different major molecular subtypes of breast cancer including luminal A, luminal B, HER2-overexpressing, and basal-like. Expression of estrogen receptor (ER), progesterone receptor (PR), and HER2 can be used to approximate these four major subgroups (luminal A: ER+ and/or PR+/HER2-; luminal B: ER+ and/or PR+/HER2+; HER2 overexpressing: ER-/HER2+; and basal-like: ER-/PR-/HER2-). The latter group is commonly called the triple-negative subtype of which basal-like tumors are one of its primary components. In the United States, 71% of tumors are Luminal A, 12% are Luminal B, 12% are triple-negative, and 5% are HER2 [3].

To decipher the breast cancer heterogeneity in terms of gene expression profiles, longitudinal or time-series data for the same cohort (reasonably large size) of patients are necessary. However, no such temporal data are available for patients with breast cancer. Recent studies show that single-cell gene expression with no temporal information can be analyzed to discover the mechanism of cell development by inferring pseudotime [4] [5][6][7][8]. These approaches allow us to observe the continuous changes at gene expression levels of cells and provide far more insights into the transcriptional kinetics of cell differentiation. These studies motivated us to hypothesize that the static mRNA expression data for breast cancers can be explored to decipher the trajectories of breast cancer development as well as heterogeneity at different points of the trajectory leading to metastasis by inferring pseudotime. The proposed study assumes that a cancer sample or patient represents the average behavior of a cell population, meaning that different patients represent different states of cell dynamics or different states of cancer development on a continuous trajectory.

In the present study, *first* -- feature selection approaches are used to reduce the dimension from 20K to few hundreds; *second* -- t-Distributed Stochastic Neighbor Embedding (t-SNE) is used to reduce the dimension to 3 t-SNE components; *third* – Principal curve analysis is done on samples with 3 t-SNE components to draw a smooth curve connecting all the clusters. The pseudotime for a patient is evaluated from the length of projection on the curve; *fourth* – k-means and gap statistic are applied to discover the heterogeneity from the inferred the pseudotime for cancer patients from static mRNA expression data.

Two major contributions of this study are: First, a computational framework is developed which is capable to construct the trajectory of breast cancer development by inferring pseudotime from static mRNA expression data. Second, it is also capable of deciphering heterogeneity in cancer along the trajectory of cancer at a granular level.

II. DATASET PREPARATION

The RNASeq mRNA expression profiles and clinical data of breast cancer patients were obtained from LinkedOmics [9]. The expression data contains profiles for 1093 patients whereas clinical data contains cancer stage information for

1072 patients. Finally, 1072 patients having both expression profiles and clinical stage information were used for analysis.

Usually, mRNA expression profiles come with expression values of about 20K genes for each patient. However, all these genes are not related to cancer and cancer heterogeneity. The list of genes related to cancer are discovered using supervised machine learning techniques, such as a set of 50 genes for breast cancer known as PAM50 (Prediction Analysis of Microarray), to classify intrinsic subtypes [10]. For the present study, we used three different sets of genes to infer the trajectory of cancer development – the first set is PAM50 genes, the second set (214 genes) and third set (233 genes) were obtained using feature selection approaches SVM-RFE (Support Vector Machine Recursive Feature Elimination) [11] and RF (Random Forest) [12] with respect to four clinical stages of cancer – Stage-I, Stage-II, Stage-III, and Stage-IV. It is surprising that gene sets produced by SVM-RFE and RF have only 16 genes in common.

III. METHODOLOGY

Four different statistical and machine learning techniques, namely, t-Distributed Stochastic Neighbor Embedding (t-SNE) [13], Principal Curve Analysis [14], gap statistic [15], and k-means [16] are used to develop the computational framework for discovering heterogeneity in breast cancer. The different components of methodology are: a) Constructing the trajectory of breast cancer development, b) Method to check quality of constructed trajectory, and c) Deciphering heterogeneity in breast cancer

A. Constructing the Trajectory of Breast Cancer Development

The dimension reduction technique, t-SNE [13] is used to reduce the dimension of mRNA expression profiles of cancer patients from the number of key genes (50 from PAM50, 214 from SVM-RFE, and 233 from RF) to three t-SNE components. Then the principal curve analysis [14] on cancer patients with three t-SNE components is conducted to construct a smooth curve connecting the clusters of cancer patients. The value of projection on the principal curve from each point (patient) represents the pseudotime for that patient on the trajectory of cancer development. A normalized pseudotime scale, ranging from 0 to 1, is developed from the projection values. This pseudotime based ordering of breast cancer patients represents the trajectory of breast cancer development.

B. Method to Check Quality of Constructed Trajectory

The cancer stage information has been used to evaluate the pair-wise order of cancer samples on the trajectory of cancer development thus determining the quality of a trajectory. Let *t* represents the trajectory of cancer development, an ordered path of *n* samples, produced by the pseudotime-based reconstruction method. Let $s(t,i,j)$ be a score to represent how well the order of the ith and jth samples in the ordered path *t* matches their expected order based on the stage information. A pseudo-temporal ordering score (POS) [6] for trajectory *t* are evaluated as the sum of $s(t,i,j)$ for all pairs of samples.

$$POS_t = \sum_{i=1}^{n-1} \sum_{j:j>i} s(t,i,j)$$

Where $s(t,i,j)$ is 0, if two samples are in the same stage. When the two samples are in different stages i.e., the ith and jth samples are collected at stages x and y respectively, then the value of $s(t,i,j)$ is either 1, if x is an earlier stage, or -1 if x is a later stage than y.

C. Deciphering Heterogeneity in Breast Cancer

The trajectory developed above can be analyzed to decipher the heterogeneity of breast cancer at different time periods along the trajectory in terms of - a) clustering to see how many clusters the patients form in a specific time period and b) mRNA expression profiles of key genes in each cluster of patients in a specific time period.

To discover the heterogeneity in terms of cluster at different time periods, patients are divided into groups of evenly spaced time periods between 0 and 1. For example, the four equal time periods will be 0.00 – 0.25, 0.25 – 0.50, 0.50 – 0.75, and 0.75 – 1.00. For each time period, the samples will be clustered based on three t-SNE components using k-means and gap statistic [15]. Gap statistic provides the optimum number of clusters that can be produced from a given set of samples.

To decipher the heterogeneity in terms of mRNA expression profiles at different time periods, normalized expression values for key genes are plotted along the trajectory for patients in different clusters in a time period.

IV. RESULTS AND DISCUSSION

At the present state of the developed computational framework, there are three challenges. *Challenge-1:* In different runs, t-SNE produces different sets of component values for a particular patient due to the stochastic nature of the algorithm, which results in different trajectories with different ordering scores; *Chanllenge-2*: Principal curve fitting may not start from a patient at Stage-I; *Chanllenge-3*: In finding optimum clusters using gap statistic, which depends on the generation of a reference dataset from the given dataset, may (most of the time it produces the same number of clusters) result in different number of optimum clusters in different runs. These challenges will be addressed in our future work.

Table 1 shows the pseudo-temporal ordering scores (POSs) for the trajectories developed in 10 runs with each of the three sets of key genes - PAM50, SVM-RFE, and RF. Due to the randomness of t-SNE technique, it produces trajectories with both positive and negative scores. Positive score represents that the patients are in correct order whereas negative score means the patients are out of order. Higher the positive score the better is the ordering, which means that the constructed trajectory is good. The perfect score, which is the maximum score for an ideal trajectory, depends on the number of patients in each stage of cancer. The highest score produced by individual sets of genes are 25,500, 46,016, and 17,292 for PAM50, SVM-RFE, and RF respectively. The trajectory with the highest score 46,016 is used for further analysis in deciphering heterogeneity in breast cancer.

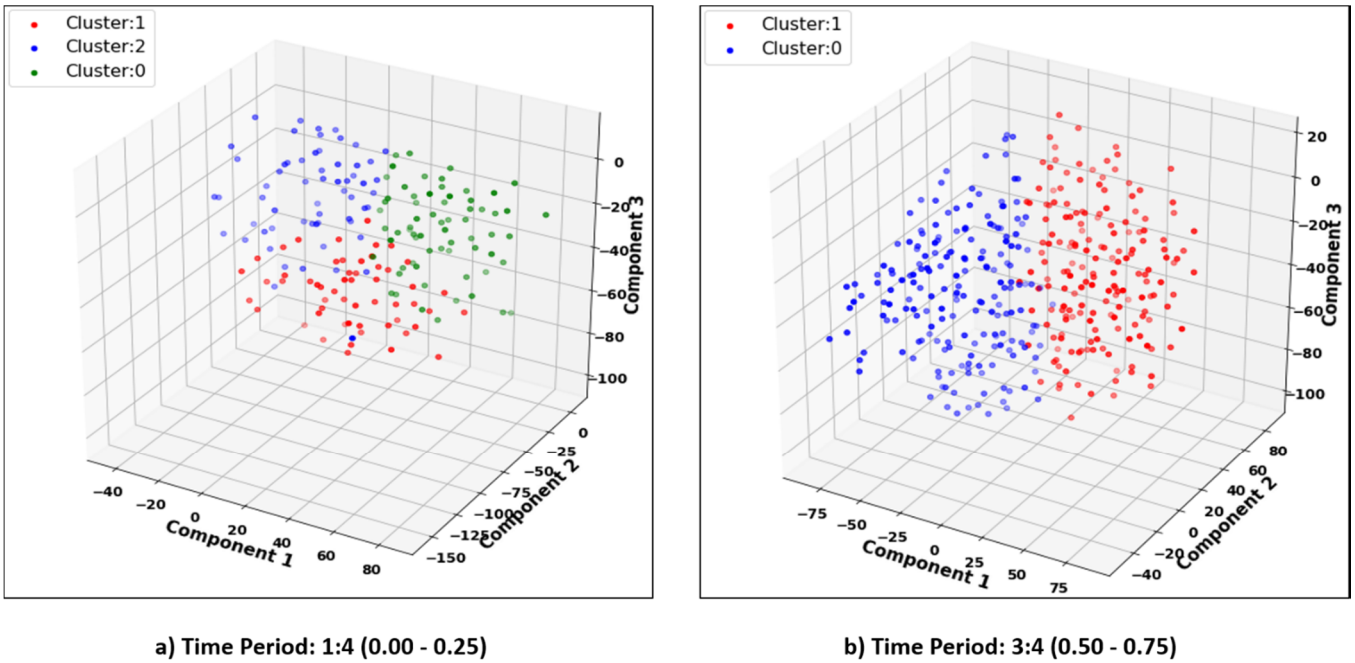
TABLE I. PSEUDO-TEMPORAL ORDERING SCORES FOR THE TRAJECTOIRES DEVELOPED IN DIFFERENT RUNS WITH EACH SETS OF GENES OBTAINED FROM PAM50, SVM-RFE, AND RF.

t-SNE Run	PAM50	SVM-RFE	RF
1	-5348	**46016**	**17292**
2	10364	-15504	-27134
3	**25500**	-23572	6760
4	14466	3586	9426
5	14776	19454	11944
6	5530	-17606	-8294
7	8664	-9134	-19762
8	7196	-13998	-5532
9	4426	16720	8208
10	3646	-10566	-16992

A. Heterogeneity in Terms of Cluster of Patients

To discover the heterogeneity of breast cancer patients in terms of cluster: *First*, the patients are divided into groups by dividing the trajectory into four (Fig. 1) and eight (Fig. 2) equal time periods; *Second*, the optimum number of clusters among the patients at each time period are determined using k-means [16] and gap statistic [15].

Fig. 1 shows the clusters of breast cancer patients at 1st (0.00 to 0.25) and 3rd (0.50 to 0.75) time periods from 4 equal divisions of whole trajectory on a scale of normalized pseudotime between 0 and 1. Similarly, Fig. 2 shows the clusters at 5th (0.500 to 0.625) and 6th (0.625 to 0.750) time periods from 8 equal divisions of the whole trajectory. Table II presents the number of clusters at each time period for 4 and 8 divisions of the whole trajectory. For four equal divisions, patients in each time period are grouped into 2 to 3 clusters producing a total of 10 clusters along the trajectory, while in case of eight divisions, patients are grouped into 2 to 4 clusters producing a total of 23 clusters. Existence of multiple clusters in each time period indicates the heterogeneity of breast cancer patients in that time period. Usually, for a trajectory of a normal event, all samples at a time period should form a single cluster. But in case of breast cancer trajectory, which is highly heterogeneous, one would expect multiple clusters at each time period as evidenced from Fig.1 and Fig.2. Fig. 2 shows the heterogeneity at a finer level by dividing the trajectory into smaller time periods. In Fig.2, patients of Fig.1(b) are divided into two equal time periods (0.500 to 0.625 and 0.625 to 0.750). Now, one can see, there are 6 clusters in Fig. 2 compare to only 2 clusters in Fig. 1(b). It is clear from two figures that the proposed computational framework is capable of deciphering heterogeneity in cancer at a granular level.



a) Time Period: 1:4 (0.00 - 0.25) **b) Time Period: 3:4 (0.50 - 0.75)**

Fig. 1. Clusters of breast cancer patients from four equal divisions of trajectory on a scale of normalized pseudotime between 0 and 1. a) 1st time period from 4 divisions represented by 1:4 and range of time for 1st time period is 0.00 to 0.25; b) 3rd time period from 4 divisions represented by 3:4 and range of time for 3rd time period is 0.50 to 0.75. These clusters are generated using k-means [16] and gap statistic [15] on cancer patients with three t-SNE components.

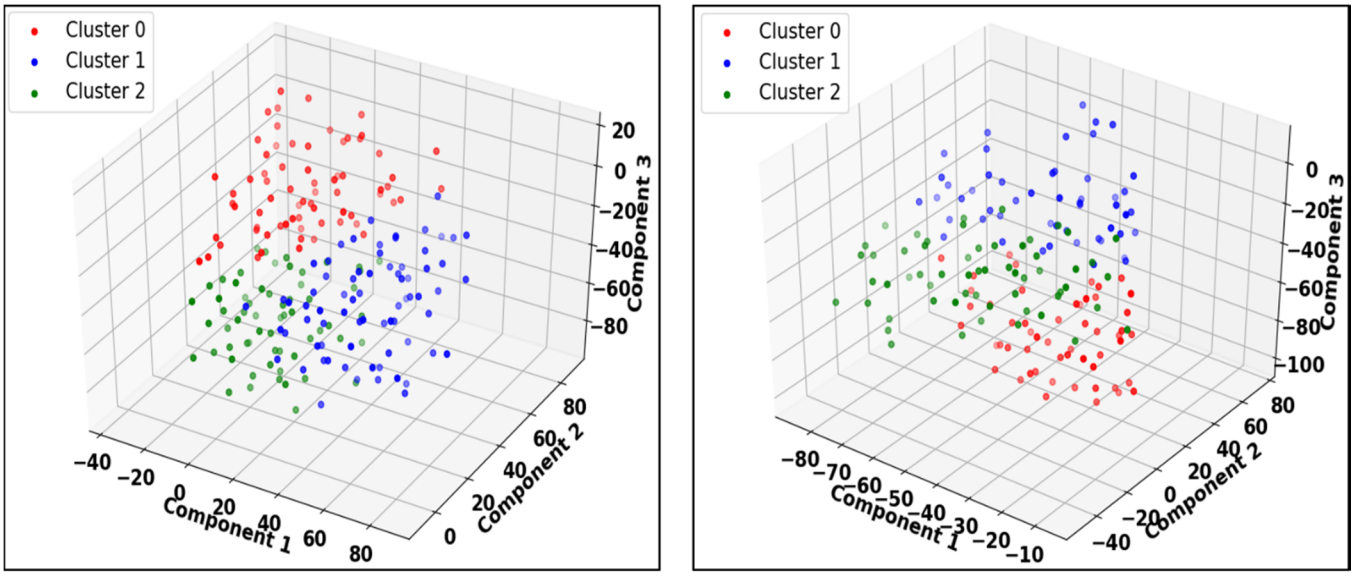


Fig. 2. Clusters of breast cancer patients from eight equal divisions of trajectory on a scale of normalized pseudotime between 0 and 1. a) 5th time period from 8 divisions represented by 5:8 and range of time for 5th time period is 0.500 to 0.625; b) 6th time period from 8 divisions represented by 6:8 and range of time for 6th time period is 0.625 to 0.750. These clusters are generated using k-means [16] and gap statistic [15] on cancer patients with three t-SNE components.

TABLE II. NUMBER OF CLUSTERS AT EACH TIME PERIOD FOR 4 AND 8 DIVISIONS OF THE WHOLE TRAJECTORY. P-1 REPRESENTS TIME PERIOD 1 AND SO ON.

Div.	P-1	P-2	P-3	P-4	P-5	P-6	P-7	P-8
4	3	2	2	3				
8	2	2	3	3	3	3	3	4

B. *Heterogeneity in Terms of mRNA Expression*

Figs. 3 and 4 present the heterogeneity of breast cancer at different time period by showing the trajectories of cancer patients in different clusters with respect to mRNA expression of one of the key genes, DARC. Fig. 3 shows the expression profile of breast cancer patients at 1st (0.00 to 0.25) and 3rd (0.50 to 0.75) time periods from 4 equal divisions of whole trajectory on a scale of normalized pseudotime between 0 and 1. Similarly, Fig. 4 shows the expression profile at 5th (0.500 to 0.625) and 6th (0.625 to 0.750) time periods from 8 equal divisions of the whole trajectory. It is clear from Fig. 3(a) that patients in cluster 0 (blue) and cluster 2 (orange) follow two parallel trajectories between time period 0.10 and 0.25, which represents heterogeneity of cancer patients with respect to mRNA expression. Fig. 3(b) shows a little evidence of heterogeneity due to the overlap of two clusters by 8 to 10 patients from cluster 1 (orange) between time period 0.50 and 0.62. Smaller division of the trajectory, Fig. 4, which divides the patients of Fig. 3(b) into two equal time periods (0.500 to 0.625 and 0.625 to 0.750) shows clear evidence of heterogeneity by producing parallel trajectories. This experiment shows the existence of parallel trajectories for breast cancer development. So, the proposed method has the capability of discovering parallel trajectories for breast cancer development.

C. *Heterogeneity of Trajectory in Terms of Cancer Stages*

Table III shows the distribution of breast cancer patients along the trajectory of cancer development with respect to cancer stages. In ideal cases, patients with stage-i would lie at the beginning and with stage-iv would lie towards the end of the trajectory. But, in case of breast cancer, patients from each stage are distributed all over the trajectory as indicated by Table III. This clearly shows that breast cancer is highly heterogeneous. This means that for some patients, cancer may progress fast and for others it may be slow.

TABLE III. DISTRIBUTION OF BREAST CANCER PATIENTS WITH RESEPECT TO CANCER STAGES ALONG THE TRAJECTORY OF CANCER DEVELOPMENT.

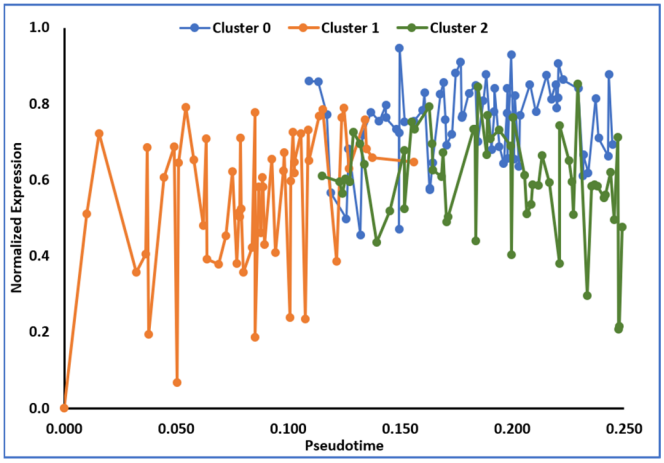
Stage	**Period 1 0.00 – 0.25**	**Period 2 0.25 – 0.50**	**Period 3 0.50 – 0.75**	**Period 4 0.75 – 1.00**
Stage-i	45	57	74	4
Stage-ii	96	230	201	92
Stage-iii	36	85	86	43
Stage-iv	4	5	4	7

D. *Heterogeneity of Trajectory in Terms of Intrisic Subtypes*

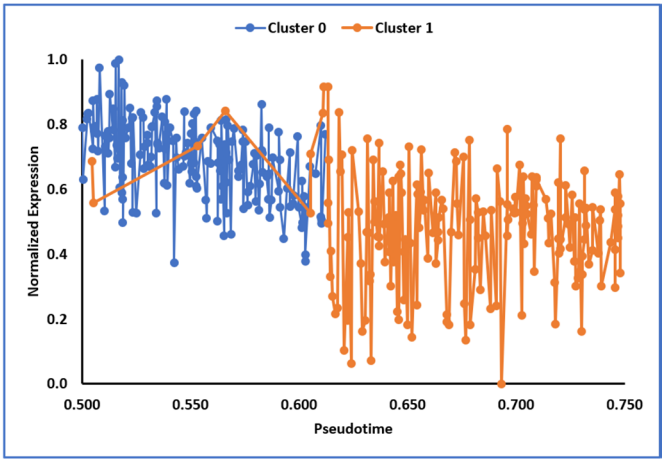
Table IV shows the distribution of breast cancer patients along the trajectory of cancer development with respect to intrinsic subtypes. The patients with subtypes LumA and LumB are distributed over the whole trajectory, while most of the Basal and HER2 are concentrated at time period between 0.25 and 0.50.

TABLE IV. DISTRIBUTION OF BREAST CANCER PATIENTS WITH RESEPECT TO INTRINSIC SUBTYPES ALONG THE TRAJECTORY OF CANCER DEVELOPMENT.

Intrinsic Subtypes	**Period 1 0.00 – 0.25**	**Period 2 0.25 – 0.50**	**Period 3 0.50 – 0.75**	**Period 4 0.75 – 1.00**
LumA	115	61	203	34
LumB	20	23	70	70
Basal		135	2	6
HER2	5	52		8

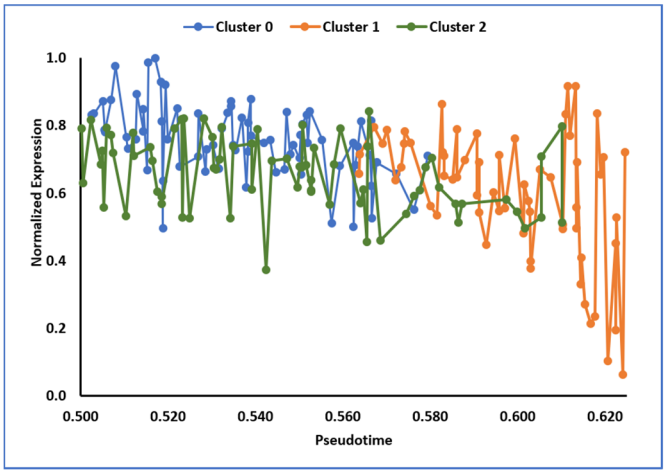


a) Time Period: 1:4 (0.00 - 0.25)

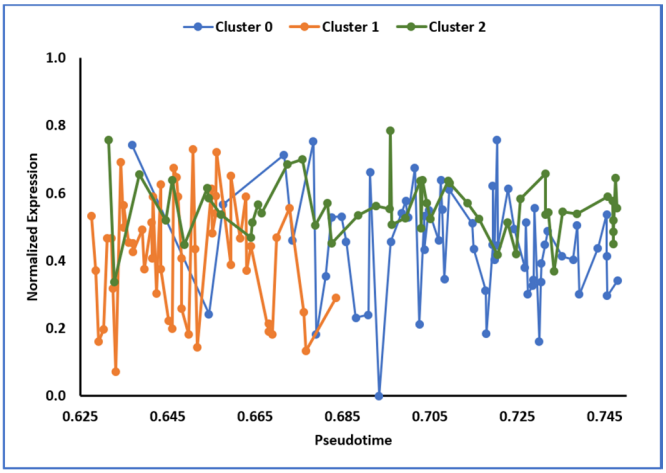


b) Time Period: 3:4 (0.50 - 0.75)

Fig. 3. Expression profile of breast cancer patients from four equal divisions of trajectory on a scale of normalized pseudotime between 0 and 1. a) 1st time period from 4 divisions represented by 1:4 and range of time for 1st time period is 0.00 to 0.25; b) 3rd time period from 4 divisions represented by 3:4 and range of time for 3rd time period is 0.50 to 0.75. The expression profile is based on one of the key genes, DARC.



a) Time Period: 5:8 (0.500 - 0.625)



b) Time Period: 6:8 (0.625 - 0.750)

Fig. 4. Expression profile of breast cancer patients from eight equal divisions of trajectory on a scale of normalized pseudotime between 0 and 1. a) 5th time period from 8 divisions represented by 5:8 and range of time for 5th time period is 0.500 to 0.625; b) 6th time period from 8 divisions represented by 6:8 and range of time for 6th time period is 0.625 to 0.750. The expression profile is based on one of the key genes, DARC.

V. Conclusion and Future Remarks

A computational framework is developed to construct a trajectory of cancer development by inferring pseudotime from static mRNA expression data. The developed trajectory is used to analyze the heterogeneity of breast cancer. The proposed method discovered that there exists multiple parallel trajectories for breast cancer at different time periods along the trajectory. Though the proposed computational framework is capable of deciphering the heterogeneity of breast cancer at a granular level, it comes with three challenges: *Challenge-1:* In different runs, t-SNE produces different sets of component values for a particular patient due to the stochastic nature of the algorithm, which results in different trajectories with different ordering scores; *Chanllenge-2:* Principal curve fitting may not start from a patient at Stage-I; *Chanllenge-3:* In finding optimum clusters using gap statistic, which depends on the generation of a reference dataset from the given dataset, may (most of the time it produces the same number of clusters) result in different number of optimum clusters in different runs. These challenges will be addressed in future work. Solving these challenges will provide a computaional framework for developing the representative trajectory of cancer development.

Acknowledgment

This research is partially funded by NSF CAREER award #1651917 (transferred to #1901628) to AMM.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA. Cancer J. Clin., vol. 68, no. 6, pp. 394–424, Nov. 2018.

[2] "Cancer facts and figures-2019." [Online]. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-

facts-and-figures/2019/cancer-facts-and-figures-2019.pdf. [Accessed: 19-Oct-2019].
[3] L. Tao, S. L. Gomez, T. H. M. Keegan, A. W. Kurian, and C. A. Clarke, "Breast cancer mortality in African-American and non-hispanic white women by molecular subtype and stage at diagnosis: A population-based study," Cancer Epidemiol. Biomarkers Prev., vol. 24, no. 7, pp. 1039–1045, Jul. 2015.
[4] C. Trapnell et al., "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.," Nat. Biotechnol., vol. 32, no. 4, pp. 381–386, Apr. 2014.
[5] E. Marco et al., "Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape.," Proc. Natl. Acad. Sci. U. S. A., vol. 111, no. 52, pp. E5643-50, Dec. 2014.
[6] Z. Ji and H. Ji, "TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis," Nucleic Acids Res., vol. 44, no. 13, pp. e117–e117, Jul. 2016.
[7] J. Shin et al., "Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis," Cell Stem Cell, vol. 17, no. 3, pp. 360–372, Sep. 2015.
[8] K. Campbell, C. P. Ponting, and C. Webber, "Laplacian eigenmaps and principal curves for high resolution pseudotemporal ordering of single-cell RNA-seq profiles," bioRxiv, p. 027219, Sep. 2015.
[9] S. V Vasaikar, P. Straub, J. Wang, and B. Zhang, "LinkedOmics: analyzing multi-omics data within and across 32 cancer types," Nucleic Acids Res., vol. 46, no. D1, pp. D956–D963, Nov. 2017.
[10] P. S. Bernard et al., "Supervised risk predictor of breast cancer based on intrinsic subtypes," J. Clin. Oncol., vol. 27, no. 8, pp. 1160–1167, Mar. 2009.
[11] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Mach. Learn., vol. 46, no. 1–3, pp. 389–422, 2002.
[12] L. Breiman, "Random Forests. transparencias," Mach. Learn., vol. 45, no. 1, pp. 1–33, 2001.
[13] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," J. Mach. Learn. Res., vol. 9, no. Nov, pp. 2579–2605, 2008.
[14] T. Hastie and W. Stuetzle, "Principal Curves," J. Am. Stat. Assoc., vol. 84, no. 406, pp. 502–516, Jun. 1989.
[15] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," J. R. Stat. Soc. Ser. B (Statistical Methodol., vol. 63, no. 2, pp. 411–423, May 2001.
[16] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1965, pp. 281–297.