CANCER HOMOGENEITY IN SINGLE CELL REVEALED BY BI-STATE MODEL AND BINARY MATRIX FACTORIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single cell RNA sequencing (scRNAseq) technology enables quantifying gene expression profiles by individual cells within cancer. Dimension reduction methods have been commonly used for cell clustering analysis and visualization of the data. Current dimension reduction methods tend overly eliminate the expression variations correspond to less dominating characteristics, such we fail to find the homogenious properties of cancer development. In this paper, we proposed a new and clustering analysis method for scRNAseq data, namely BBSC, via implementing a binarization of the gene expression profile into on/off frequency changes with a Boolean matrix factorization. The low rank representation of expression matrix recovered by BBSC increase the resolution in identifying distinct cell types or functions. Application of BBSC on two cancer scRNAseq data successfully discovered both homogeneous and heterogeneous cancer cell clusters. Further finding showed potential in preventing cancer progression.

1 INTRODUCTION

Cancer the biggest deadly threat to human has been a huge puzzle since its determination in 1775. From once considered as contagious to nowadays cancer immunotherapy, the modern medication continues to evolve in tackling this problem (Dougan et al., 2019). And yet, not enough to make a huge difference, 1,762,450 people have been diagnosed with cancer and 606,880 has died in 2018 (Siegel et al., 2019). The development of single cell RNA sequencing (scRNA-seq), which measures each single cell in cancer tissue with over 20,000 dimension of genes (features), picturized the hologram of cancer and its micro-environment with high resolution (Picelli et al., 2014; Puram et al., 2017; Tirosh et al., 2016). As illustrated in Figure 1A, classic analysis pipeline takes a linear (PCA) or non-linear (t-SNE) dimension reduction of the high dimensional input data, by which loadings of the top bases are further used for cell clustering and visualization (Tirosh et al., 2016).

Cancer cell heterogeneity hampers theraputic development. We use the melanoma dataset as an example. Cells in a scRNA-seq data are always with multiple crossed conditions, such as types of cancer, origin of patients and different cell types. By analyzing melanoma scRNA-seq data with classic pipeline, we differentiated the cell type of each cell in its cancer microenvironment (CME) (figure 1B). All cell types other than cancer cell are constituted by multiple patients (figure 1C), validated the accuracy of classic pipeline in cell type identification. While on cancer cell, each patient forms a distinct cluster (highlighted in shadow), suggesting confounding patient-wise heterogeneity. Similar phenomenon also exists in breast cancer and head and neck cancer. On the other hand, being an investment-heavy industry like medical industry, the uniqueness of each cancer patient contradicts its general principle as to

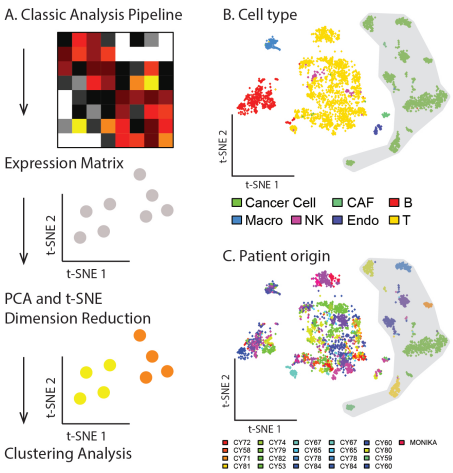

Figure 1: Classic analysis pipeline for scRNA-seq data and Melanoma example

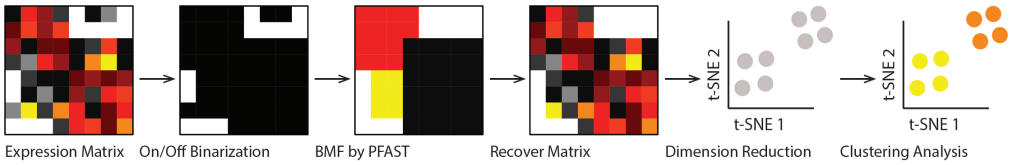

Figure 2: BBSC Pipeline for scRNA-seq data

seek a rather universal treatment for a broad range of patients. To solve this dilemma, major modifications are needed for the analysis pipeline of cancer scRNA-seq data.

Approximate gene expression with Bi-state model. The expression of one gene in a single cell is characterized as the following two-state bursting model determined by two factors, transcriptional frequency (f) and size (k_{size}) (Larsson et al., 2019)

$$f|k_{on}, k_{off} \sim Beta(k_{on}, k_{off})$$
$$y|k_{size}, f \sim Poisson(k_{size} \cdot f)$$
$$x \sim y + \epsilon,\ \epsilon \sim N(\lambda_0, \sigma_0)$$

In addition, f follows a beta distribution accounting for the collective effect of the probability to shift the expression from off to on (k_{on}) and from on to off (k_{off}). y denotes the true expression of gene i inside cell j and x is the observation of y with Gaussian error. Recent study revealed that, regulated by enhancers, burst frequency f is the major facilitator of cell type specific gene expression landscape (Larsson et al., 2019). Though f and k_{size} cannot be precisely fitted from our observed data, since y follows the Poisson distribution of the pure product of k_{size} and f, we could still capture the most significant frequency changes across different cells. That is, we could infer whether f is above or equal to zero, corresponding to expression/no-expression of the gene, from our observed data. Counting this property, we thus propose the following approximate gene expression bi-state models.

$$F^{n\times m} = A^{n\times k} \otimes B^{k\times m} + E, \tag{1}$$

$$Y_{ij} \sim \begin{cases} Poisson(\lambda_i),\ if\ F_{ij} = 1, \\ 0,\ if\ F_{ij} = 0, \end{cases} \tag{2}$$

$$X_{ij} \sim Y_{ij} + \epsilon_{ij},\ \epsilon_{ij} \sim N(\lambda_0, \sigma_0), \tag{3}$$

where F denotes a latent binary matrix of f, which is considered as a low rank representation of k different cell types, generated by the Boolean product of two binary matrix A and B plus a Boolean flipping error E. Y denotes the true quantitative expression level generated from F, and X is considered as a measure of Y with i.i.d. Gaussian error ϵ. Here our approach takes the approximating Y by Hadamard product between X and $\hat{A}^{n\times k} \otimes \hat{B}^{k\times m}$, i.e.

$$\hat{Y} = X \circ (\hat{A}^{n\times k} \otimes \hat{B}^{k\times m}),$$

where $\hat{A}^{n\times k}$ and $\hat{B}^{k\times m}$ are the estimation of $A^{n\times k}$ and $B^{k\times m}$.

Bi-state and Boolean matrix factorization for scRNA-seq data (BBSC). In sight of this, we developed a novel scRNA-seq pattern mining and analysis pipeline namely BBSC (Figure 2), by implementing a data binarization process for the inference of ON/OFF bi-state expression patterns. In addition, we proposed a fast binary matrix factorization (BMF) method, namely PFAST, adapting to the large scale of scRNA-seq data. BBSC can be easily implemented with classic dimension reduction based analysis procedure. Application of BBSC on scRNA-seq of the head and neck cancer and melanoma data successfully revealed the cancer homogeneity hence increased the sensitivity in identifying sub types of cells. In addition, cancer cell clusters expressing the epithelial mesenchymal transition (EMT) markers were specifically identified by BBSC in head and neck cancer study, which consist cancer cells from different patient samples, suggesting heterogeneous cancer cells may adopt a similar strategy in cancer metastasis process.

We summarize our contributions as follows:

- We constructed a scRNA-seq analysis pipeline, BBSC, for retrieving cancer homogeneity properties. BBSC is by far the first analysis pipeline accounting the fundamental interplay between cell type and gene expression in the analysis of scRNA-seq data.
- As a major component in BBSC pipeline, we proposed a fast and efficient BMF algorithm, PFAST, in adapting to the large scale of scRNA-seq data.
- In the analysis of head and neck cancer data, BBSC identified that cancer cell may adapt similar strategies in metastasis. This finding could be applied to prevent cancer progression.

2 Related Work

So far, two strategies have been used to optimize the classic pipeline for scRNA-seq data analysis: (1) using extra information to supervise the dimension reduction, such as CITE-seq and REAP-seq data combining scRNA-seq with additional protein information (Stoeckius et al., 2017; Peterson et al., 2017) or a recent work by Peng et al. (2019), by maximizing the similarity with bulk RNA seq data for scRNAseq imputation; and (2) limiting analysis to the genes known to be related with desired biological features Tirosh et al. (2016). Both strategies require substantial prior information that is either expensive or unsuitable for studying biological characterization. In this paper, we developed a new strategy rooted from a perspective that differences in cell types and physiological states correspond to different bi-state frequency patterns, which could retrieve effectively by Boolean matrix factorization.

Following the Boolean algebra, BMF decomposes a binary matrix as the Boolean product of two lower rank binary matrices and has revealed its strength in retrieving information from binary data. Due to the NP completeness of the BMF problem, several heuristic solutions have been developed, among which two series of works are most frequently utilized (Miettinen et al., 2008; Lucchese et al., 2010). One is ASSO algorithm developed by Miettinen et al. (2008). ASSO first generates potential column basis from row-wise correlation. Then adopts a greedy searching from generated basis for the BMF fitting. The second series of work is the PANDA algorithm developed by Lucchese et al. (2010). PANDA aims to identify the top 1-enriched submatrices in a binary matrix from background noise. In each iteration, PANDA excludes the current fitting from the input matrix and retains a residual matrix for further fitting. More recently, Bayesian inference has involved in this field. Ravanbakhsh et al. (2016) retrieve patterns from factor-graph model by deriving MAP using message passing (denoted MP). Rukat et al. (2017) proposed OrMachine, provide full probabilistic inference for binary matrices. While ASSO and PANDA being regarded as the baseline in BMF, MP and OrMachine represent state-of-the-art performance.

3 BBSC Analysis Pipeline

As shown in Figure 2, we implemented a data binarization and PFAST algorithm to constrain scRNA-seq data before a regular dimension reduction based analysis, which forms a new analysis pipeline namely BBSC. BBSC first binarizes the input data via the on/off expression states of each gene. The approximated matrix, namely recover matrix, is further constructed by the Hadamard product of the original expression matrix and the BMF fitted binary matrix. Regular dimension reduction and cell clustering analysis is then conducted on the recovered matrix.

3.1 Characterization of ON/OFF expression state

To determine a gene is truly expressed or not is to examine X_{ij} on ϵ. Empirically, we assume the lowest none zero expression value of each gene approximates the distribution of ϵ. Since type I error is far damaging than type II error in biological experiments, we utilized the 95% quantile of ϵ distribution as the threshold

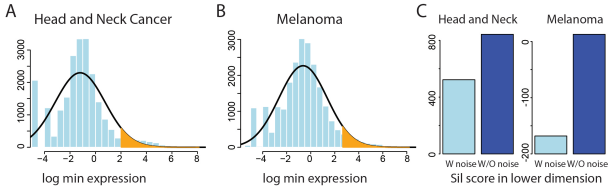

Figure 3: Infer F from scRNA-seq data

of ON expression state, i.e., gene expression above the threshold is considered as $f > 0$ while expression below the threshold is considered as with an OFF state, i.e. $f = 0$. We applied this binarization procedure on two high quality scRNAseq cancer datasets of head and neck cancer (Figure 3A) and melanoma (Figure 3B). To justify the threshold of ON/OFF state computed in this way, we compared the representation of data in the lower dimension by the overall silhouette score, which measures the similarity of each data point to own cluster compare to others. The overall silhouette score represents the goodness of the clustering. Note that cell cluster information is retrieved directly from original paper. In both datasets, the binarization approach significantly increased the performance of cluster representation, suggesting our binarization can remove true noise and still maintains the biological information.

3.2 PFAST ALGORITHM

We developed a fast and efficient BMF algorithm, namely PFAST to cope with the large scale of modern data of interest. PFAST follows the general framework of PANDA algorithm. In each iteration, PANDA has two main sub functions, core pattern discovery (Core) and extensions of a core pattern (Core_ext). Core finds the most enriched square of 1s under current residual matrix. Core_ext expands the generated core patterns with not included area. To find most precise patterns amid noise, PANDA calculates global loss at each step. Though PANDA only works on the residual matrix in each iteration, it still involves already generated patterns for calculating loss. This look back property and global loss calculation may play a major role in decomposing noisy binary data. However, the associated computational pressure makes PANDA inapplicable for large-scale scRNA-seq data. Fortunately, during our binarization process, 95% of noise has been eliminated, which compensates an extensive binary pattern mining as PFAST. Unlike PANDA, PFAST only focus on the loss in a local scale. Moreover, PFAST abolished the look back property, only focus the loss decrease for current pattern. Taken together, PFAST is an extensive BMF algorithm. Each iteration of PFAST has a computational complexity of O(mn). Like PANDA, PFAST will only work iteratively on residual matrix that has not been covered by any identified patterns before hitting the convergence criteria. The choice of convergence criteria can be modified for different needs. The popular convergence criteria are set by identifying top k patterns or covering certain proportion of the non-zero values in the matrices. Detailed algorithms of PFAST is illustrated below:

Algorithm 1: PFAST

Inputs: Binary matrix F, Threshold t, and τ
Outputs: $A \in \{0,1\}^{n\times k}$, $B \in \{0,1\}^{k\times m}$
$PFAST(F, t, \tau)$:
$A \leftarrow \emptyset$ $B \leftarrow \emptyset$ $Fr \leftarrow F$
while $!\tau$ **do**
  $(\mathbf{a}, \mathbf{b}) \leftarrow PFAST\_core(Fr)$
  $(\mathbf{a}, \mathbf{b}) \leftarrow PFAST\_ext\_core(Fr, \mathbf{a}, \mathbf{b}, t)$
  $A \leftarrow A \cup \mathbf{a}$ $B \leftarrow B \cup \mathbf{b}$
  $Er_{ij} \leftarrow 0$ where $(\mathbf{a} \otimes \mathbf{b})_{ij} = 1$
end

3.3 EVALUATION OF PFAST ALGORITHM ON SYNTHETIC DATA

Since OrMachine has been deprecated, we compared the performance of PFAST with ASSO, PANDA ,and MP on simulated datasets. We simulated binary matrices$X^{n\times m} = U^{n\times k} \otimes V^{k\times m}$ where each element of U and V follows an identical Bernoulli random variable. In the simulation, we set $n = m = 1000, k = 5$, and two signal level $p = 0.2/0.4$, corresponding to sparse and dense matrix. We compared the performance with three criterion: reconstructed error, sparsity, and time cost. Specifically, reconstructed error measures the overall fitting of each method, and sparsity measures the parsimonious level of the pattern matrices. Detailed definition of reconstructed error and sparsity are given below. Intuitively, a good binary matrix factorization should have small reconstructed error and proper sparsity level. To the best of our knowledge, the conditions to guarantee

Algorithm 2: PFAST_core

Inputs: Residual matrix Fr
Outputs: $\mathbf{a} \in \{0,1\}^n$, $\mathbf{b} \in \{0,1\}^m$
$PFAST\_core(Fr)$:
$\mathbf{s} = \{\mathbf{s}_1, ..., \mathbf{s}_n\} \leftarrow$ sorting based on row-wise sum
$\mathbf{a} \leftarrow 0^n$; $\mathbf{b} \leftarrow 0^m$; $\mathbf{a}_{\mathbf{s}_1} \leftarrow 1$; $\mathbf{b}_{\mathbf{s}_1} \leftarrow 1 \; \forall i \; s.t. Fr_{\mathbf{s}_1,i} = 1$
for $l \leftarrow 2, ..., n$ **do**
  $\mathbf{a}^* \leftarrow \mathbf{a}$; $\mathbf{a}^*_{\mathbf{s}_l} \leftarrow 1$; $\mathbf{b}^* \leftarrow \mathbf{b}$; $\mathbf{b}^*_i \leftarrow 0 \; \forall i \; s.t. \; Fr_{\mathbf{s}_l,i} = 0$
  if $sum(Fr_{\boldsymbol{a}^*,\boldsymbol{b}^*}) > sum(Fr_{\boldsymbol{a},\boldsymbol{b}})$ **then**
    $\mathbf{a} \leftarrow \mathbf{a}^*$; $\mathbf{b} \leftarrow \mathbf{b}^*$
end

Algorithm 3: PFAST_ext_core

Inputs: Fr **a** **b** t
Outputs: $\mathbf{a} \in \{0,1\}^n$, $\mathbf{b} \in \{0,1\}^m$
$PFAST\_ext\_core(Fr, \mathbf{a}, \mathbf{b}, t)$:
$Fext \leftarrow Fr_{-,\mathbf{b}}$
for $i \; in \; 1, ...n$ **do**
  $\mathbf{a}_i \leftarrow 1 \; \forall i \; |Eext_{i,-}| > |\mathbf{b}| * t$
end

a unique solution of the BMF problem have not been theoretically derived, thus we do not directly compare the factorized and true pattern matrices directly, i.e., U vs A^*, and V vs B^*, where A^* and B^* denote the pattern matrices decomposed by the three different algorithms. Note that ASSO and PFAST require one additional parameter as a standard input. To achieve a fair comparison, we tested different parameters for each method and used the parameter with the best performance for the comparison. The convergence criteria for all the methods were set as when (1) 5 patterns were identified, corresponds to the true rank of simulated matrices; (2) identified patterns already covers 95% of the non-zero values. All the experiments ran on the same laptop with i7-7600U CPU and 16 GB memory. We conducted the evaluation for 10 times, detailed results are shown in Figure 4. The definitions of reconstructed error and sparsity are

$$\text{reconstructed error} = \frac{|(U \otimes V) \ominus (A^* \otimes B^*)|}{|U \otimes V|} \quad \text{sparsity} = \frac{|A^{*n \times k}| + |B^{*k \times m}|}{(n+m) \times k}.$$

Comparing to ASSO, PANDA, and MP, our analysis showed that PFAST achieved superior performance in both sparse and dense matrices. The running time of PFAST is significant lower than all other methods. We also observed better convergence of PFAST. ASSO tended to find the most inclusive patterns so that they usually converged with very few dense patterns. PANDA was designed to identified significant patterns from background noise. Its low tolerance to noise caused a relative slow pace in convergence. MP revealed its robustness in fitting binary data. However, it has the highest computational cost compared to others. The performance of PFAST demonstrated its balanced computational cost and fitting accuracy. With the significant improvement of speed, PFAST still manages to

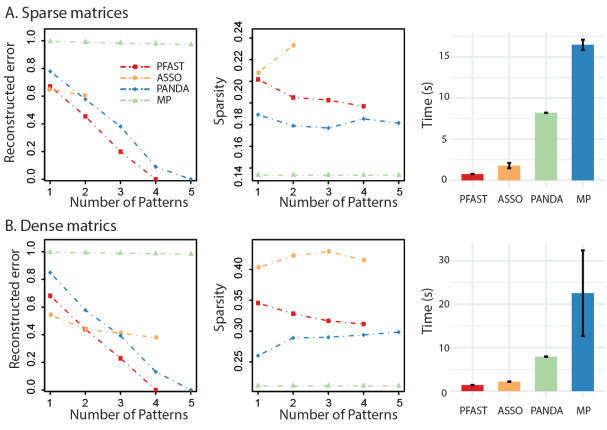

Figure 4: Performance comparison of PFAST with ASSO, PANDA and MP

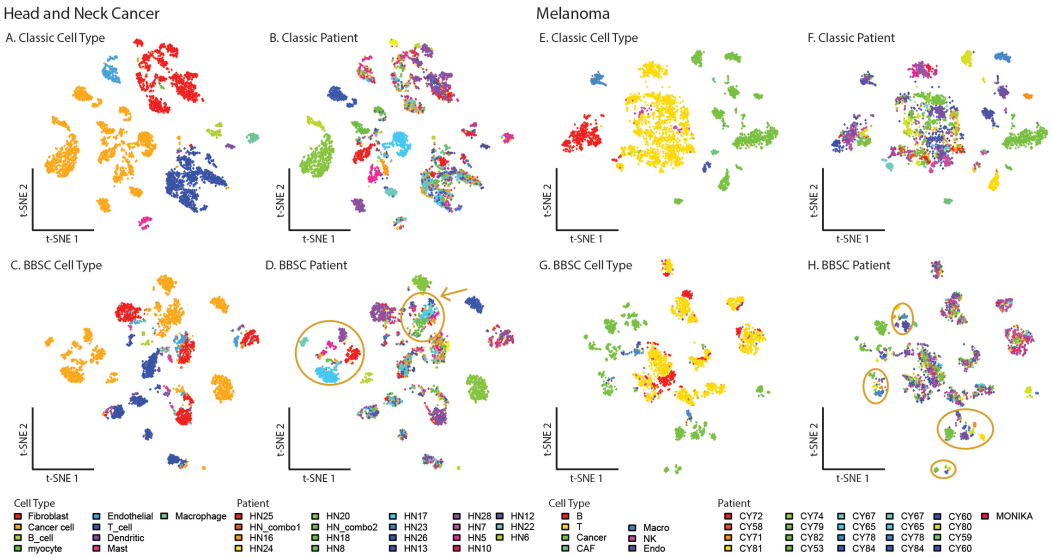

Figure 5: BBSC analysis of Head and neck and Melanoma scRNA-seq data

maintain information by decomposed patterns. Meanwhile, the PFAST has a pattern sparsity level very close to true density 0.2 and 0.4, also indicating the rationality of PFAST decomposition. Thus, PFAST is suitable in dealing with large scale data like scRNA-seq data.

4 Application of BBSC on real Cancer Datasets

We applied classic tSNE-based dimension reduction and BBSC analysis on the head and neck cancer and melanoma data sets, as detailed below. For both datasets, we recovered the bi-state model of data by binarizing the expression matrix into ON/OFF expression state with 95% Gaussian noise quantile. PFAST was applied on the binary matrices with threshold setting to 0.6. The choice of convergence criteria can vary according to different needs. Here, we set convergence as 1) top 10 patterns have been identified, 2) 40% of non-zero values has been recovered. The rationale here is that scRNA-seq data is overall sparse. It usually cost extensive patterns to achieve a small reconstructed error. However, the later discovered patterns introduced more bias, where the later patterns are more likely to be related to other factors rather than cell type. Empirically, top 10 patterns and 40% cutoff achieve better cell type identification ability. In analyzing the head and neck cancer and melanoma data sets, it resulted in 5 and 10 patterns respectively. In both analysis pipeline, we conducted dimension reduction using t-SNE with perplexity setting to 30 with 20000 max iterations. It is noteworthy that no cell clustering was made in this analysis. All the cell type annotation and patient information were directly retrieved from the original paper. As illustrated in Figure 5A,E, the 2D embedding achieved from the classic pipeline well separated cells by their phenotypic types. Fibroblast, T-, B-, myeloid and cancer cells et al forms distinct individual clusters. Further analysis of the association between cell groups and patient information confirmed same type of the immune and stromal cells from different patients form one cell group, while the cancer cells are grouped by specific patient over the 2D embedding (Figure 5B,F). These observations are consistent with original work.

On the other hand, on the 2D embedding of the BBSC pipeline, cell of different phenotypic types form into distinct groups. Comparing to the classic pipeline, BBSC retrieved data generated more groups of subtypes of Fibroblast, T cells and cancer cells (Figure 5C,G). The split cell groups identified by BBSC show higher association with intra-cancer heterogeneity. We further investigated the association between the patient origin and cell group over the 2D embedding of the BBSC data (Figure 5D,H). Interestingly, in both datasets, we observed several cell groups, marked with yellow circles, that are constituted by cancer cells for different patients. These cancer cell groups correspond to the common sub cell populations prevalently shared by cancer tissues with different patients, which may suggest hallmark functions developed in the disease progression.

To identify the functional characteristics of BBSC derived cell groups, we checked the differentially expressed genes associated with the cell groups of cancer cells. We first achieved five distinct clusters of the cancer cells over the 2D embedding of the BBSC retrieved data by using k-mean method (Figure 6A). Figure 6B illustrates the newly clustered cancer cell in the 2D embedding derived by the classic tSNE method. The cluster 1 and 2 are formed by cells from different patients while the cluster 3 to 5 were associated with specific patients. We identified significant differential expression of epithelial-mesenchymal transition (EMT) marker genes among the five clusters (Figure 6C). EMT is regarded as a hallmark event in cancer cells metastasis approach for carcinomas such as head and neck cancer [13]. Under this process, cancer cells lose their epithelial properties and become mesenchymal-like cells with higher migratory capabilities for escaping the cancer tissue into circulating system. We identified the cluster 1 and 2 behaved distinct difference compared with cluster 3 to 5 on EMT marker genes. Cells in the cluster 1 and 2 are with overly expressed mesenchymal markers such as CDH3, TGFB1, ITGB6 and VIM. While the cluster 3 to 5 overly express epithelial markers genes such as CDH1, CLDN4, CLDN7, KRT19 and EPCAM. Our analysis clearly demonstrated the BBSC substantially removed inter-cancer heterogeneity that enables the identification of cancer cells from different patients with common functional characteristics. More importantly, the observation also suggests though cancer cell are very different in each patient, they ought to take similar strategy in the metastasis process. Targeting the progression strategy revealed in this study may have huge therapeutic impact in preventing cancer progression.

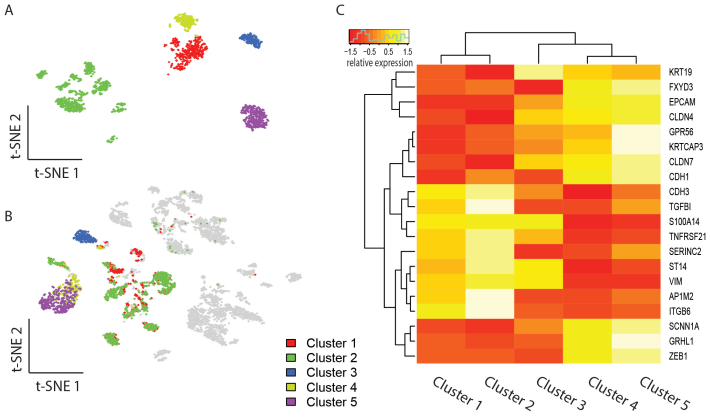

Figure 6: Detailed analysis of cancer cell clusters

5 Discussion

Enabled by the development of single cell technology, we now can observe the complicated biological process like cancer with unprecedented resolution. However, the classic analysis pipeline fails to deliver detailed information: 1) it does not reveal common characteristic of cancer cell in different cancer patients. 2) Even it separates functional cells; it fails to reveal intra-cluster heterogeneity. To solve above problems, we have developed BBSC analysis pipeline. Rooted from casting the frequency change in gene expression, we have applied BMF in the feature selection process, which avoids adding new expensive and potentially noisy information. We have applied tailored binarizing process for each dataset. Moreover, to deal with big scale tall matrix like scRNAseq data, we have developed a fast and efficient algorithm called PFAST. Letting alone its fast speed in handling large-scale data, it shows high accuracy compared with state-of-art BMF algorithms. We have applied BBSC on two high quality cancer studies, head and neck cancer and melanoma. In both datasets, BBSC shutters the big clusters into several sub clusters, and promotes a gateway to analysis intra-cluster heterogeneity. Moreover, BBSC manages to get common cancer sub cell clusters in both datasets, and decreases the patient-wise heterogeneity that hindered cancer therapeutic development. We next have justified the biological meanings of BBSC derived sub clusters by looking into the sub cancer clusters in head and neck cancer. By analyzing their detailed expression profile, We find out that the common clusters are in the EMT transition process indicating these cancer cells play an important part in cancer metastasis. While patient specific clusters are in the early EMT process indicating that these cells are still in the original cancer micro environment. These findings have first justified the biological importance of BBSC derived sub clusters. Secondly, it brings much insightful ideas in the clinical application. We now can hypothesize that when cancer cells

seek metastasis, they will transform into similar states that are common across different patients. The characteristic of the common clusters may serve as target in preventing cancer metastasis. Furthermore, we validate that the heterogeneity of cancer comes from the original cancer tissue. Also BBSC shows promising results in deciphering this kind of heterogeneity. Especially in head and neck cancer study, BBSC distinctly divides cancer cells from the same patient into two sub clusters. Due to our limited expertise in cancer biology, we did not look closely in this property. However, we believe this would bring insightful ideas in the cause of cancer origin heterogeneity. Overall BBSC is an efficient and valuable analysis platform for scRNAseq or other single cell data. It is capable to bring insightful knowledge for our detailed understanding of complicated biological process.

REFERENCES

Michael Dougan, Glenn Dranoff, and Stephanie K Dougan. Cancer immunotherapy: beyond checkpoint blockade. *Annual Review of Cancer Biology*, 3:55–75, 2019.

Anton JM Larsson, Per Johnsson, Michael Hagemann-Jensen, Leonard Hartmanis, Omid R Faridani, Björn Reinius, Åsa Segerstolpe, Chloe M Rivera, Bing Ren, and Rickard Sandberg. Genomic encoding of transcriptional burst kinetics. *Nature*, 565(7738):251, 2019.

Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. Mining top-k patterns from binary datasets in presence of noise. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pp. 165–176. SIAM, 2010.

Pauli Miettinen, Taneli Mielikäinen, Aristides Gionis, Gautam Das, and Heikki Mannila. The discrete basis problem. *IEEE transactions on knowledge and data engineering*, 20(10):1348–1362, 2008.

Tao Peng, Qin Zhu, Penghang Yin, and Kai Tan. Scrabble: single-cell rna-seq imputation constrained by bulk rna-seq data. *Genome biology*, 20(1):88, 2019.

Vanessa M Peterson, Kelvin Xi Zhang, Namit Kumar, Jerelyn Wong, Lixia Li, Douglas C Wilson, Renee Moore, Terrill K McClanahan, Svetlana Sadekova, and Joel A Klappenbach. Multiplexed quantification of proteins and transcripts in single cells. *Nature biotechnology*, 35(10):936, 2017.

Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1):171, 2014.

Sidharth V Puram, Itay Tirosh, Anuraag S Parikh, Anoop P Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina L Luo, Edmund A Mroz, Kevin S Emerick, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, 171(7):1611–1624, 2017.

Siamak Ravanbakhsh, Barnabás Póczos, and Russell Greiner. Boolean matrix factorization and noisy completion via message passing. In *ICML*, pp. 945–954, 2016.

Tammo Rukat, Chris C Holmes, Michalis K Titsias, and Christopher Yau. Bayesian boolean matrix factorisation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2969–2978. JMLR. org, 2017.

Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(1):7–34, 2019.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865, 2017.

Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.