A data denoising approach to optimize functional clustering of single cell RNA-sequencing data

Changlin Wan*
Electrical and Computer Engineering
Purdue University
wan82@purdue.edu

Dongya Jia*,†
Center for Theoretical Biological Physics
Rice University
dj9@rice.edu

Yue Zhao
H. John Heinz III College
Carnegie Mellon University
zhaoy@cmu.edu

Wennan Chang
Electrical and Computer Engineering
Purdue University
chang534@purdue.edu

Sha Cao
Biostatistics
Indiana University
shacao@iu.edu

Xiao Wang
Statistics
Purdue University
wangxiao@purdue.edu

Chi Zhang†
School of Medicine
Indiana University
czhang87@iu.edu

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313483

***Abstract*—Single cell RNA-sequencing (scRNA-seq) technology enables comprehensive transcriptomic profiling of thousands of cells with distinct phenotypic and physiological states in a complex tissue. Substantial efforts have been made to characterize single cells of distinct identities from scRNA-seq data, including various cell clustering techniques. While existing approaches can handle single cells in terms of different cell (sub)types at a high resolution, identification of the functional variability within the same cell type remains unsolved. In addition, there is a lack of robust method to handle the inter-subject variation that often brings severe confounding effects for the functional clustering of single cells. In this study, we developed a novel data denoising and cell clustering approach, namely CIBS, to provide biologically explainable functional classification for scRNA-seq data. CIBS is based on a systems biology model of transcriptional regulation that assumes a multi-modality distribution of the cells' activation status, and it utilizes a Boolean matrix factorization approach on the discretized expression status to robustly derive functional modules. CIBS is empowered by a novel fast Boolean Matrix Factorization method, namely PFAST, to increase the computational feasibility on large scale scRNA-seq data. Application of CIBS on two scRNA-seq datasets collected from cancer tumor micro-environment successfully identified subgroups of cancer cells with distinct expression patterns of epithelial-mesenchymal transition and extracellular matrix marker genes, which was not revealed by the existing cell clustering analysis tools. The identified cell groups were significantly associated with the clinically confirmed lymph-node invasion and metastasis events across different patients.**

***Index Terms*—Cell clustering analysis, Data denoising, Boolean matrix factorization, Cancer microenvirionment, Metastasis.**

I. Introduction

The rise of single-cell RNA sequencing (scRNA-seq) technology has revolutionized the biological and biomedical research fields in recent years [1], [2]. One important mission of scRNA-seq is to explore the inter/intra-subject heterogeneity of the tissue microenvironment by enumerating compositions of the cells, and their functional states. Essentially, most of the state-of-the-art approaches detect cell clusters based on cell-wise distances that is calculated based on a pre-selected set of gene features, or their projections. However, technical confounders, such as dropouts, are prevalent in single cell data that could introduce large variability in single cell expression data, and severely affect the clustering efficiency. On the other hand, recent studies revealed that many varied gene expressions may not necessarily contribute to different cell types or their functional activity states [1], [2], due to cell type unrelated expression and the confounding factor induced gene expression variation. Such that the challenge to robustly identify the functional variations within the same cell type while selecting and relying on only the informative genes remains unsolved.

In sight of these challenges, we here developed a novel data denoising framework, namely CIBS (**C**ell type **I**dentification by fast **B**oolean matrix factorization of **S**cRNA-seq data)[1], that can be seamlessly implemented with existing cell clustering analysis, to optimize the detection of cell groups with converged functional activities from scRNA-seq data. The key ingredient of CIBS is derived from a systems biology perspective, by considering the variation of observed gene expression as a result of the on-/off-switch of the gene's transcriptional regulators. A multi-modal model is applied to transform a scRNA-seq data into a binary expression matrix, where each matrix element represents a discretized gene expression status. A functional gene module is further modeled as a subset of genes showing consistently active expression states across a subset of cells, and such a module can be identified by a Boolean Matrix Factorization (BMF) approach on the binarized expression state matrix. A new BMF algorithm, namely PFAST, was developed to empower the analysis on large scale scRNA-seq data. The input scRNA-seq data will be further denoised by removing the expression signal that cannot be explained by principle binary matrix factors. Both CIBS framework and PFAST algorithm were benchmarked with state-of-the-art methods on real-world and synthetic data sets. Specifically, application of CIBS on two scRNA-seq data of cancer cells identified distinct cell group associated with

[1]*equal contribution, † correspondence.
CIBS can be accessed at https://github.com/clwan/CIBS

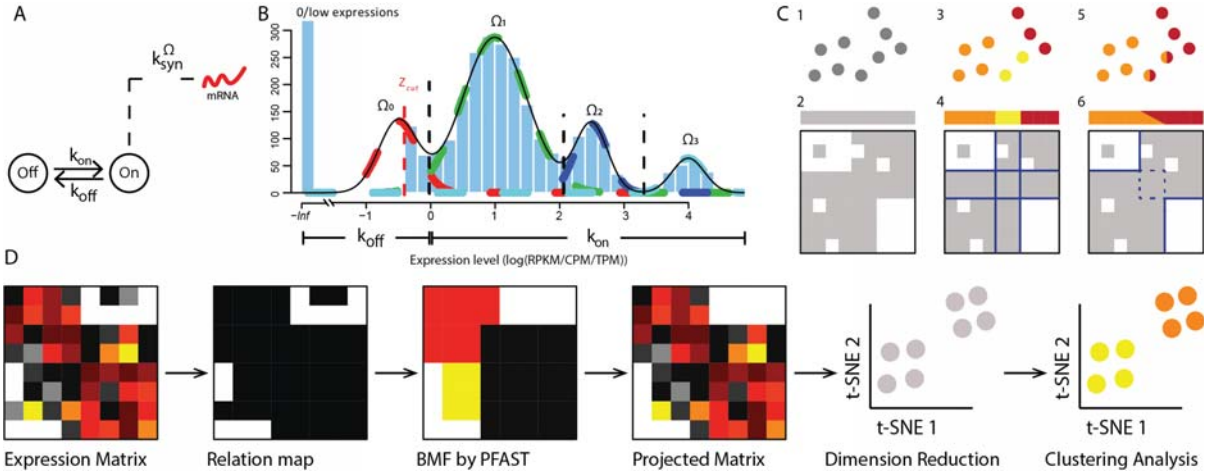

Fig. 1. Gene expression model and CIBS pipeline. A. two-state gene expression. B. LTMG models on/off and multimodality in gene expression. C. functional cell type with its gene expression modules. D. illustrations of CIBS pipeline.

epithelial-mesenchymal transition and metastasis, which was validated by matched clinical data.

The key contributions of this work include:

- We developed CIBS, a user-friendly data denoising framework that can be implemented with existing cell clustering methods to optimize the detection of functionally meaningful cell groups.
- We developed PFAST, a computationally efficient BMF algorithm that is customized to handle large scale scRNA-seq data with more than 20,000 gene features and over 5,000 cells.
- Application of CIBS on cancer scRNA-Seq datasets collected from different patients identified distinct cell groups and signature genes related to epithelial-mesenchymal transition (EMT) and metastasis.
- Downstream analysis revealed the epithelial-like cancer cells have high inter-tumoral heterogeneity while the mesenchymal-like cancer cells tend to be more homogeneous, suggesting that the cancer cells share similar trancriptional variations during the EMT process.

II. Background and Problem Formulation

Numerous models have been developed to infer the multi-modality in the expression profile of each gene in scRNA-seq data [4], [9]. Recently, Larson et al utilized allele specific scRNA-seq to trace the allele origin of each mRNA molecule, which for the first time validated that gene's expression was determined by an on-/off-state, where they also revealed the determination of different expression states are the major facilitators in determining the functional groups of cell (Fig 1A) [5]. We have recently developed a left truncated mixture Gaussian model to simultaneously discriminate the on- and off- expression state and estimate the multi-modality of the on expression state of single gene's expression profile through multiple cells. Denote $X \in \mathbb{R}^{m \times n}$, where X_{ij} is the expression level of gene j in the cell i, the LTMG model enables a direct discrimination of the expression profile of each gene into on- and off- expression states. Specifically, X_{ij} is inferred as with an on-state if $\exists l > 0, s.t., a_l^j p_l(X_{ij}|u_{on}^{\Omega_l,j}, \sigma_{on}^{\Omega_l,j}) > a_0^j p_0(X_{ij}|u_{off}^{\Omega_0,j}, \sigma_{off}^{\Omega_0,j})$, and X_{ij} is inferred as with an off-state if $a_0^j p_0(X_{ij}|u_{off}^{\Omega_0,j}, \sigma_{off}^{\Omega_0,j}) > a_l^j p_l(X_{ij}|u_{on}^{\Omega_l,j}, \sigma_{on}^{\Omega_l,j})$ for all $l > 0$ (Fig 1B, see details in [4]). We introduce a binary expression state matrix P with P_{ij} defined as

$$P_{ij} = 1\, X_{ij} \in \Omega_{1,2,\ldots,m}, \quad P_{ij} = 0\, X_{ij} \in \Omega_0$$

, i.e. the $P_{ij} = 0$ *or* 1 indicates X_{ij} is with an off- or on-expression state. As discussed above, we deem the variations in the expression state matrix P can better represent true functional variations comparing to the original expression profile X. Another advantage to consider P instead of X is that the expression variation led by different batch or other noise were eliminated in P. While we shift the focus to the binary expression state matrix P, the functional clustering problem is thus transferred as mining the relational information among modules of cells and genes, i.e., to conduct a disentangled representation learning of P [6], [7].

Low rank representation of binary matrix is a warehouse for disentangling the 1s enriched pattern matrices, i.e rank-1 matrices, from a binary data (Fig 1C1,2), which can be solved by a co-clustering or Boolean Matrix Factorizaton (BMF) approach[8]. Noted. existing co-clustering method tend to identify non-overlapped patterns (Fig 1C3,4). However, this assumption does not fit scRNA-seq data since each gene can serve multi-functionalities in different cells, i.e. the activated expression of one gene can be regulated for attending multiple functional modules [10]. Hence it is necessary to enable overlaps among the rank-1 patterns, i.e. the observed expressions state matrix is the Boolean aggregation of different functional modules, which forms a BMF problem [6], [7], [11] (1C5,6). However, the binarization and BMF process may over simplify the expression variations. For example, different activated expression state of one gene may correspond to different func-

tions, which are presented as same values (1) in P but different values (original expression level) in X. To better characterize functional related gene expression variations, instead of using P alone, we project the original expression profile X onto the BMF fitted expression state matrix $\hat{P}$, i.e., the Hadamard product, $\hat{X} = X \circ \hat{P}$, for cell clustering analysis.

III. CIBS PIPELINE

We introduce the CIBS analysis pipeline (Fig 1D). Based on the LTMG model, CIBS first derives the expression state matrix P from expression profile X as described above. The BMF identifies functional modules of genes that are consistently with on-expression state in a subset of cells. To encourage the within-group difference, the original expression values was then projected onto then BMF fitting expression state matrix P, which only leaves the gene expressions represent in at least one functional module. Dimension reduction and cell clustering methods are further applied on the denoised data to derive functional cell clusters (Fig 1D).

To cope with the large scale of scRNA-seq data, we propose a fast BMF algorithm, namely PFAST. PFAST follows the general framework of well-received PANDA algorithm [7] with some major improvements: (1) In searching optimal basis, PANDA needs to calculate global loss with $O(mn)$ complexity on every attempt. On the other hand, PFAST only needs to sum up the positive values within covered regions, which has approximate complexity of $O(m)$. (2) Both PFAST and PANDA update residual matrix after each iteration. However, PFAST removed the look back step, which was set in PANDA to consider already generated patterns in decomposing current residual matrix. We found this approach affects the efficiency along with the increase of pattern number but has limited improvement to the matrix decomposition. (3) While expanding patterns generated by PFAST_core, PFAST_ext_core simplifies the computational cost by introducing a similarity cutoff t that reduces its complexity to $O(n)$. In summary, each iteration of PFAST has an approximated complexity of $O(mn)$, and the total complexity of PFAST is $O(kmn)$, where k is number of identified patterns. The detailed illustration of PFAST as well as the auxiliary algorithm PFAST_core and PFAST_ext_core are listed in Appendix.

We benchmark PFAST with ASSO, PANDA and MP on simulated data (see experiment detail in Appendix). Comparing to ASSO, PANDA and MP, our analysis suggested that PFAST achieved superior performance in both sparse and dense matrices (2). The running time of PFAST is significantly lower than other methods. We also observed a better convergence rate of PFAST. Meanwhile, the number of patterns detected by PFAST are closer to the number ture patterns, for both sparse and dense matrices.

IV. APPLICATION OF CIBS ON REAL DATASETS

We applied CIBS on the head and neck cancer and melanoma data with the following analysis settings (see experiment detail in appendix). We compared the cell clusters identified by using Seurat on the original data and CIBS denosised data, as it represent the state-of-the-art performance in cell clustering analysis. All the cell type annotation and patient information were directly retrieved from the original works. Seurat identified cells clusters correspond to different cell types, including distinct clusters of Fibroblast, T-, B-myeloid, and cancer cells (Fig 3A, E,). By investigating the patient origin of each cell in both datasets, the stromal and immune cell types from different patients form consistent cell type specific cluster while clusters of cancer cells were largely separated by their patient origin (Fig 3B, F). These observations are consistent with the original work [1], [2]. On the other hand, applicaton of Seurat on CIBS denoised data also identified clusters of cells with distinct stromal and immune types (Fig 3C, G). But for cancer cells, we observed several sub clusters (marked with yellow circles in Fig 3D, H) that are constituted by cancer cells from multiple patients in both datasets.

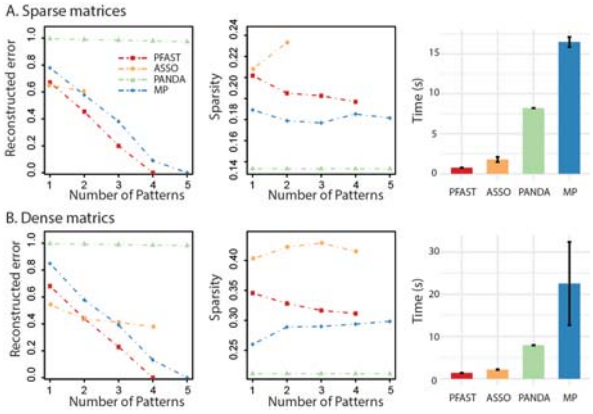

Fig. 2. Performance comparison of PFAST with ASSO, PANDA and MP. For sparse and dense matrices (A and B) respect to reconstructed error, pattern density and time cost to reach convergence.

We further focused on the analysis and biological interpretation of the functional variations of the identified clusters of cancer cells [1], [2]. We retrieved all cancer cells in the head and neck dataset and identified 5 clusters from the CIBS denoised data, as shown in (Fig 3I). The cluster 1 and 2 are formed by cells from multiple patients, while the cluster 3 to 5 are associated to specific patients. We identified significant differential expression of epithelial-mesenchymal transition (EMT) genes among the five clusters (Fig 3J).Cells in the cluster 1 and 2 are with overly expression of well confirmed marker genes of mesenchymal-like cell, such as CDH3, TGFB1, ITGB6 and VIM, while the cluster 3, 4 and 5 overly express epithelial marker genes such as CDH1, CLDN4, CLDN7 KRT19 and EPCAM, suggesting the cell cluster 1 and 2 are of mesenchymal-like cells and cluster 3-5 are of epithelial-like cells. In addition, the patient origins dominating the cell clusters 1 and 2 are with higher number of lymph-node invasion comparing to the patients of the cell cluster 3-5, suggesting the EMT-associated cell groups identified from CIBS denoised data are truly related to metastasis. On

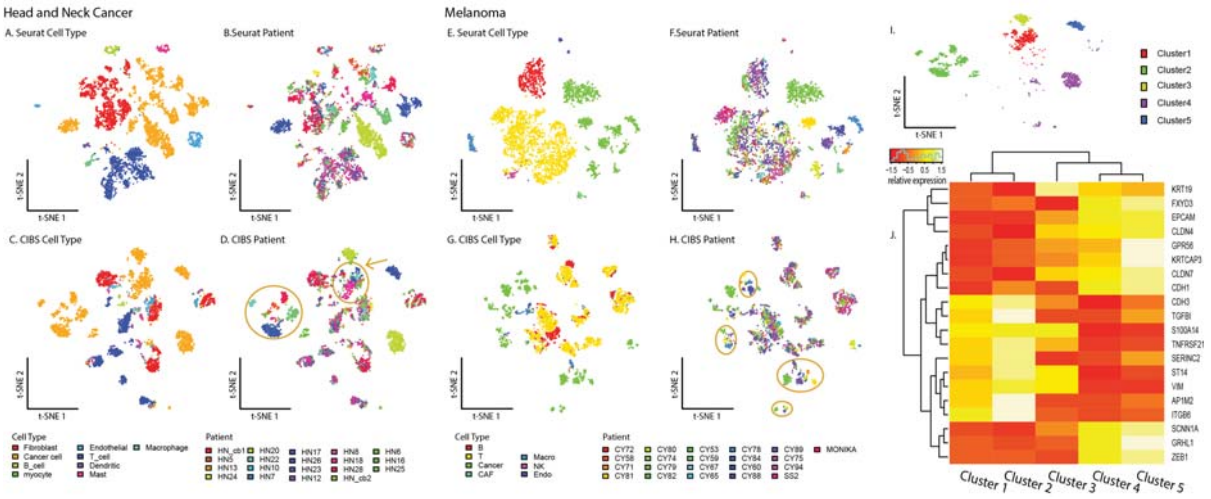

Fig. 3. Application of CIBS on real cancer dataset. A-H. CIBS clustered cancer cell from multiple head and neck, and melanoma patients. I-J. Detailed analysis revealed CIBS identified functional cell types are in different states in EMT process.

the other hand, the cell clusters inferred by Seurat on the original data are all patient origin specific. Our observation clearly suggested the clusters identified by CIBS correspond to sub groups of cancer cells with varied functional states of metastasis, demonstrating CIBS can effectively eliminate the gene expression variation irrelevant to cell functions, such as the sample-wise batch effect, and identify cell groups with truly varied functional states.

V. Discussion

Distinguishing cells of different phenotypic types or functional groups is major challenge in cell clustering analysis of scRNA-seq datasets. To tackle this challenge, we developed the CIBS data denoising approach, by detecting the gene expressions that are more likely caused by a functional variation, from a systems biology perspective. The computational model is inherited from our previously developed statistical distribution [4], which not only reduces observational noise, but also eliminates the bias led by biological factors such as different mRNA degradation rate and unfully degraded mRNAs. To deal the large feature and sample size of a scRNA-seq data, we also developed a fast Boolean matrix factorization algorithm, namely PFAST. PFAST is with a significantly decreased computational cost and detection accuracy compared with state-of-art approaches. Noted, CIBS can be easily implemented with existing cell clustering methods by providing a data denoising of the input data. We applied CIBS on two high quality cancer scRNA-seq data. Compared with classic cell clustering methods, CIBS identified clusters of cancer cells with distinct expression pattern of EMT and metastasis related genes. Interestingly, the identified epithelial-like cancer cells possess higher patient specificity but the mesenchymal-like cells from different patients are more homogeneous. CIBS enables an elimination of functional irrelevant inter-tumoral variation. Hence amplify the intra-tumorual gene expression variations of true biological functions in cell clustering analysis.

References

[1] Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., ... Fallahi-Sichani, M. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science, 352(6282), 189-196.

[2] Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., ... Deschler, D. G. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell, 171(7), 1611-1624.

[3] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., ... Satija, R. (2019). Comprehensive integration of single-cell data. Cell, 177(7), 1888-1902.

[4] Wan, C., Chang, W., Zhang, Y., Shah, F., Lu, X., Zang, Y., ... Zhang, C. (2019). LTMG: a novel statistical modeling of transcriptional expression states in single-cell RNA-Seq data. Nucleic acids research, 47(18), e111-e111.

[5] Larsson, A. J., Johnsson, P., Hagemann-Jensen, M., Hartmanis, L., Faridani, O. R., Reinius, B., ... Sandberg, R. (2019). Genomic encoding of transcriptional burst kinetics. Nature, 565(7738), 251-254.

[6] Miettinen, P., Mielikäinen, T., Gionis, A., Das, G., & Mannila, H. (2008). The discrete basis problem. IEEE transactions on knowledge and data engineering, 20(10), 1348-1362.

[7] Lucchese, C., Orlando, S., & Perego, R. (2010, April). 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)

[8] Xie, J., Ma, A., Zhang, Y., Liu, B., Cao, S., Wang, C., ... Ma, Q. (2020). QUBIC2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale RNA-Seq data. Bioinformatics, 36(4), 1143-1149.

[9] Zhang, Y., Wan, C., Wang, P., Chang, W., Huo, Y., Chen, J., ... Zhang, C. (2019). M3S: A comprehensive model selection for multi-modal single-cell RNA sequencing data. BMC bioinformatics, 20(24), 1-5.

[10] Brabletz, T., Kalluri, R., Nieto, M. A., & Weinberg, R. A. (2018). EMT in cancer. Nature Reviews Cancer, 18(2), 128.

[11] Ravanbakhsh, S., Póczos, B., & Greiner, R. (2016, June). Boolean Matrix Factorization and Noisy Completion via Message Passing. In ICML (pp. 945-954).

[12] Wan, C., Chang, W., Zhao, T., Li, M., Cao, S., & Zhang, C. (2019). Fast And Efficient Boolean Matrix Factorization By Geometric Segmentation. arXiv, arXiv-1909.

[13] Wan, C., Chang, W., Zhao, T., Zang, Y., Cao, S., Zhang, C. (2020). Denoising individual bias for a fairer binary submatrix detection. arXiv preprint arXiv:2007.15816.

APPENDIX

A. PFAST algorithm

We propose PFAST, tailored to the cost function of BMF and maintains a low computational complexity requirements. It follows the general framework of PANDA algorithm [7]. In each iteration, PFAST has two functions – core pattern discovery (PFAST_core) and extension of core pattern (PFAST_ext_core). Specifically, PFAST_core detects the most enriched square of 1s within current residual matrix, and PFAST_ext_core expands the generated core pattern with uncovered area. Starting with original binary matrix, after each iteration, PFAST iteratively retains the residual matrix with setting the 1s covered by existing derived patterns to 0s, and detect largest pattern from the residual matrix until the convergence criteria η is met. Different convergence criteria can be set based on the analysis demand. Common settings include identifying the top k patterns or covering certain ratio of 1s in the input data. The general framework of PFAST is illustrated below:

Algorithm 1: PFAST

Inputs: Binary matrix P, Threshold t, and convergence criteria η
Outputs: $A \in \{0,1\}^{n\times k}$, $B \in \{0,1\}^{k\times m}$
PFAST(P, t, η):
$A \leftarrow \emptyset$ $B \leftarrow \emptyset$ $Pr \leftarrow P$
while $!\eta$ **do**
  $(\mathbf{a}, \mathbf{b}) \leftarrow$ PFAST_core(Pr)
  $(\mathbf{a}, \mathbf{b}) \leftarrow$ PFAST_ext_core$(Pr, \mathbf{a}, \mathbf{b}, t)$
  $A \leftarrow A \cup \mathbf{a}$ $B \leftarrow B \cup \mathbf{b}$
  $Pr_{ij} \leftarrow 0$ where $(\mathbf{a} \otimes \mathbf{b})_{ij} = 1$
end

1) Core pattern discovery: Here we denote the to-be-fitted residual matrix as Pr. For each Pr, the solution space of the optimal decomposition can be as big as 2^{mn} and the optimal solution can not be guaranteed [6], [7]. PFAST inherited the searching strategy from PANDA[7]. In each iteration, PFAST first calculates and ranks row-wise $l1$ norm, $|Pr_{i,:}|$, corresponding to number of 1s for each row. And return a vector of row indices $\mathbf{s}$ of the descending order for $|Pr_{i,:}|$. Such that, the first element of $\mathbf{s}$ corresponds to the row indices with largest $l1$ norm. I.e., $\mathbf{s}_1 = argmax_i(|Pr_{i,:}|)$. Searching is initialized with $\mathbf{a} \in \{0,1\}^n$, $\mathbf{b} \in \{0,1\}^m$, where $\mathbf{b}$ is the same as $Pr_{\mathbf{s}_1,:}$. And $\mathbf{a}$ is a vector with $\mathbf{a}_{\mathbf{s}_1}$ setting to 1 and other elements as 0. PFAST_core detects 1s enriched sub matrices by following the order of $\mathbf{s}$. At term l, new patterns $\mathbf{a}^*$, $\mathbf{b}^*$ are generated, where $\mathbf{a}^*$ is the same as $\mathbf{a}$ except $\mathbf{a}_{\mathbf{s}_l}$ setting to 1, $\mathbf{b}^*$ becomes the intersection of $\mathbf{b}$ and $Pr_{\mathbf{s}_l,:}$, i.e., $\mathbf{b} \wedge Pr_{s_l,:}$. Here $\wedge$ stands for and operation under Boolean algebra, where $1 \wedge 1 = 1, 1 \wedge 0 = 0 and 0 \wedge 0 = 0$. The comparison of new patterns and current patterns is tailored to the cost function that is to find the maximum number of 1s covered. If new patterns covered more positive values, the new patterns will be promoted. Moreover, scRNAseq data are generally sparse so that the covered region is usually far less than n or m. Thus, counting the covered positive values has a approximate complexity of $O(m)$. By going this process for n-1 times, PFAST_core finds a rather dense square. And the overall complexity is $O(mn)$.

Algorithm 2: PFAST_core

Inputs: Residual matrix Pr
Outputs: $\mathbf{a} \in \{0,1\}^n$, $\mathbf{b} \in \{0,1\}^m$
PFAST_core(Pr):
$\mathbf{s} = \{\mathbf{s}_1, ..., \mathbf{s}_n\} \leftarrow$ sorting based on row-wise sum
$\mathbf{a} \leftarrow 0^n$; $\mathbf{b} \leftarrow 0^m$; $\mathbf{a}_{\mathbf{s}_1} \leftarrow 1$; $\mathbf{b} \leftarrow Pr_{s_1,:}$
for $l \leftarrow 2, ..., n$ **do**
  $\mathbf{a}^* \leftarrow \mathbf{a}$; $\mathbf{a}^*_{\mathbf{s}_i} \leftarrow 1$; $\mathbf{b}^* \leftarrow \mathbf{b} \wedge Pr_{s_l,:}$
  if $sum(Pr_{\boldsymbol{a}^*,\boldsymbol{b}^*}) > sum(Pr_{\boldsymbol{a},\boldsymbol{b}})$ **then**
    $\mathbf{a} \leftarrow \mathbf{a}^*$; $\mathbf{b} \leftarrow \mathbf{b}^*$
end

2) Core pattern extension: In the process of finding dense square, PFAST_core keeps updating the patterns. Such that the final patterns can be greatly different from its initialization. It is highly likely that some rows in Pr can fit the final patterns very well but were neglected in the searching process. PFAST_ext_core is designed to rescue the rows with good fittings to the final patterns but were omitted in the pattern searching. Based on the cost function, we are invariant to add zero values into the decomposition. This property granted a very fast process for PFAST_ext_core, that we can include $Pr_{i,:}$ as long as it mapped an acceptable ratio, t , of positive values with generated patterns. This mapping could be easily revealed by counting the row-wise sum of column pattern covered sub matrix $Pr_{:,\mathbf{b}}$. In general cases, $|\mathbf{b}| << m$, Counting row-wise sum of $Pr_{:,\mathbf{b}}$ has a complexity of $O(n)$, which makes PFAST_ext_core with an approximate complexity of $O(n)$.

Algorithm 3: PFAST_ext_core

Inputs: Pr **a** **b** t
Outputs: $\mathbf{a} \in \{0,1\}^n$, $\mathbf{b} \in \{0,1\}^m$
PFAST_ext_core$(Pr, \mathbf{a}, \mathbf{b}, t)$:
$P_{\text{ext}} \leftarrow Pr_{:,\mathbf{b}}$
for $i\ in\ 1, ...n$ **do**
  $\mathbf{a}_i \leftarrow 1\ \forall i\, |P_{\text{ext}_{i,:}}| > |\mathbf{b}| * t$
end

B. Experiments setting

1) On simulated data: We simulated binary matrices $P^{m\times n} = U^{m\times k} \otimes V^{k\times n}$, where each element of U, V follows an identical Bernoulli distribution with probability p. We set $n = m = 1000, k = 4$, and two signal levels $p = 0.2/0.4$ corresponding to sparse or dense matrix. We compared the methods follows the same metrics in [12]: (1) reconstruction error measures overall fitting, (2) density measures the parsimonious level of generated patterns, (3) time consumption. The solution with smaller reconstruction error

and density are regarded as more optimal. Detailed definitions of the metrics are listed below.

$$\text{reconstruction error} := \frac{|P \ominus (A \otimes B)|}{|P|}$$

$$\text{Density} := \frac{|A| + |B|}{(n+m) \times k}$$

The convergence criteria η for all the methods were set as: 1) 5 patterns were identified, i.e., one pattern more than the true rank, 2) the cost function stopped decreasing. For each scenario, we conducted the evaluation with 10 rounds of simulations.

2) On real cancer single cell data: The binary expression state matrix P was generated from the gene-wise LTMG fitting. PFAST was applied on the binary matrices with t equal to 0.6. Convergence criteria η for PFAST was set as: 1) top 10 patterns have been identified, 2) 30% of non-zero values has been recovered, which resulted 5 patterns in head and neck cancer and 10 patterns in melanoma datasets. ScRNA-seq data is overall sparse. A large number of patterns is usually needed to achieve a small-reconstruction error. However, some patterns identified in later iterations are likely to be less biologically explainable due their small size. Our analyses suggested an empirical setting of 10 patterns and 30% coverage rate can achieve a good derivation of functional clusters for cell clustering. We conducted the dimension reduction using t-SNE for visualization. Cell clustering analysis was conducted by using the default setting of Seurat – the most utilized cell clustering analysis for scRNA-seq data [3]. For the re-clustering in Fig3. I, we utilized k-mean clustering algorithm. Number of k is selected by using elbow method.

C. Acknowledgement

This work was supported by R01 award #1R01GM131399-01, NSF IIS (N0.1850360), Showalter Young Investigator Award from Indiana CTSI and Indiana University Grand Challenge Precision Health Initiative. This was also supported by a training fellowship from the Gulf Coast Consortia, on the Computational Cancer Biology Training Program (CPRIT grant NO.RP170593).