2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871358

Identifying Transient Cells During Reprogramming via Persistent Homology

Aydolun Petenkaya[1], Farid Manuchehrfar[1], Constantinos Chronis*, Jie Liang*

***Abstract*— Single-cell RNA sequencing is a powerful method that helps delineate the regulatory mechanisms shaping the diverse cellular populations. Heterogeneous cell populations consist of individual cells, and the expression of distinct sets of genes can differentiate one sub-population of cells from another, as they are responsible for the emergence of distinct cellular phenotypes. Of particular importance are cells at transition states that bridge these different cellular phenotypes. In this study, we develop a method to identify the cells at transition states bridging different cellular phenotypes. Our approach is based on persistent homology, which enabled us to identify the group of cells located on the boundaries between different sub-populations of cells. We applied this method to study the reprogramming of human fibroblasts toward induced pluripotent stem cells using single-cell time-course data. Even though only the data that is representative of the early stages of the reprogramming process are analyzed, we are able to uncover transient cells bridging different cell sub-populations. The most prominent group of transient cells are found to be enriched for NANOG, which is a known stem cell transcription factor that takes part in the maintenance of pluripotency and other stem cell marker genes. Overall, our method can identify cells in transient states bridging major cellular phenotypes, even though they are only a small fraction of the overall cell population. We also discuss how this approach can link the topology of the surface of cellular transcripts and bring order to the transition between cellular states and how it automatically uncovers the underlying time process.**

I. INTRODUCTION

Somatic cells can be reprogrammed with low efficiency into induced pluripotent stem cell (iPSC) state by the over-expression of the Yamanaka transcription factors Oct4, Sox2, Klf4 and cMyc (OSKM) [1], [2]. To this day the molecular mechanisms that drive the reprogramming process and the order of the key events that facilitate successful iPSC conversion remain largely unexplored. Our goal is to characterize the reprogramming trajectories and key events occurring within the successfully re-wired cells using novel computational approaches.

Single-cell RNA sequencing can be used to delineate the key events that drive cell fate specification by identifying gene expression signatures that can distinguish between somatic, transient, and pluripotent cell states. In order to identify the different cell identities among a heterogeneous population, single-cell gene expression can help to infer how similar/dissimilar cells are. The single-cell gene expression profiles can be thought as a point cloud in a high dimensional space, and analysis often benefits from dimension reduction (Fig 1). Several methods, including t-SNE [3] and UMAP [4], have been adapted for dimensional reduction for single cell analysis [5]. Although these methods help to visualize the heterogeneity at the single cell level, the global and local structures that permeate single-cell expression data are often not fully preserved and the sub-populations of cells are often inaccurately defined, perhaps due to distortions in both global and local structures of the data [6].

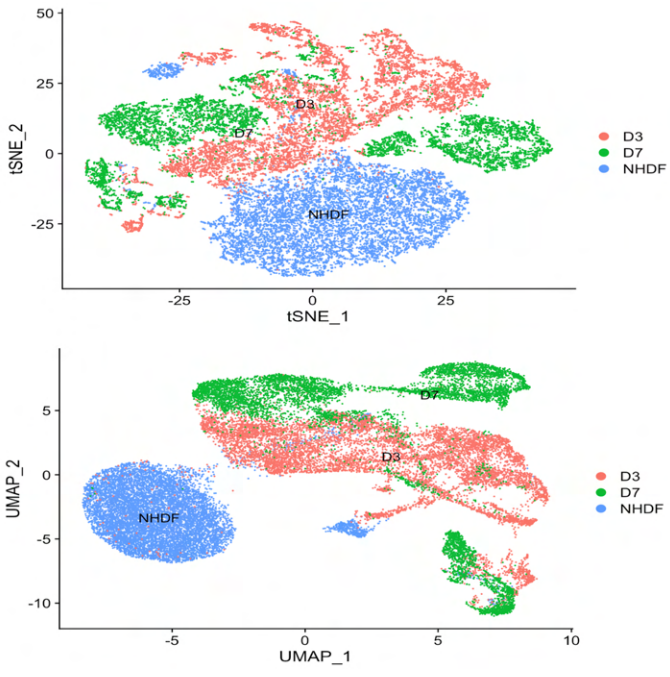

Fig. 1. Visualization of the reprogramming single-cell data with t-SNE (top) and UMAP (bottom) revealing structural differences in associations between individual cells. In these plots, each cell is represented by a dot. Cells are colored-coded, with Human fibroblasts (NHDF) in blue, fibroblasts over-expressing OSKM for three (D3) or seven (D7) days in red and green, respectively.

[1]Aydolun Petenkaya is with the Richard and Loan Hill Department of Biomedical Engineering, University of Illinois at Chicago, Chicago, IL, USA apeten2@uic.edu

[1]Farid Manuchehrfar is with the Richard and Loan Hill Department of Biomedical Engineering, University of Illinois at Chicago, Chicago, IL, USA fmanuc2@uic.edu

*Constantinos Chronis is with the Department of Biochemistry and Molecular Genetics & Department of Biomedical Engineering , University of Illinois at Chicago, Chicago, IL, USA chronis@uic.edu

*Jie Liang is with the Richard and Loan Hill Department of Biomedical Engineering & Center of Bioinformatics and Quantitative Biology, University of Illinois at Chicago, Chicago, IL, USA jliang@uic.edu2@uic.edu

We thank Chuansheng Hu for useful comments. This work is supported by NIH R35 GM127084.

All experimental cell lines and procedures described were performed in accordance to University of Illinois at Chicago, Institutional Bio-safety Committee approved protocols.

Topological data analysis provides potential resolutions to such issues [7]. Here we develop a novel approach and examine the topology of probability peaks located at coordinates determined by transcript levels. Through analysis of the topological structures of the high dimensional data cloud of single-cell RNA transcripts, our goal is to identify the cells successfully transitioning from one cellular state to another based on time-coursed measurements reflecting the beginning process of human fibroblast cells undergoing reprogramming.

II. METHODS

A. Single-cell RNA-seq Reprogramming Data

Human fibroblasts (NHDF) were infected with Sendai viruses encoding the transcription factors OSKM and single cell RNA-seq data were collected prior to infection (NHDF) and at Day3 (D3) and Day7 (D7) post infection, to investigate the earliest transcriptional changes induced in the context of iPSC reprogramming. All data used in this research were generated in house.

B. Persistent Homology to Identify Transient Cell States

Persistent homology can quantify topological features in the landscape of transcripts, whose abundancy change as cells enter different transcriptional states. Persistent homology has broad applications, including brain image analysis to investigate neurological disorders [8], unsupervised learning on network neuroscience [9] and microscopy data to study repair loci [10]. It has also been applied to analyze gene expression data [11]. Here we apply a novel persistent homology method to study single-cell gene expression of cells during reprogramming [12].

We analyze the topological space of high dimensional transcriptomic data by studying the states of probability peaks. We focus on their appearance and disappearance, which correspond to the birth and death of 0-th homology groups, namely, the appearance and disappearance of independents components above the level sets of a specified density or probability level [13]. Thus, 0-th homology groups indicates the number of connected components in the data at a particular density or probability levels [12]. We take the probability *p(x)* as the height function. At different thresholds $\{r_i\}$ of the height function:

$$1 = r_0 > r_1 > r_2 > \cdots > r_{i_{n-1}} > r_{i_n} = 0, \qquad (1)$$

we examine the sets $\{X_i\}$, $X_i = \{x \in X \,|\, p(x) \geq r_i\}$, which form a sequence, namely, a *filtration*:

$$\emptyset \equiv X_{i_0} \subset X_{i_1} \subset X_{i_2} \subset \cdots \subset X_{i_{n-1}} \subset X_{i_n} \equiv \Omega, \qquad (2)$$

As the threshold changes, different peaks start to appear from below the sea level of the threshold. This height is referred to as the birth of the component corresponding to the new peak, which is identified by the 0-th homology group. At a different threshold, some components may disappear. This process is referred to as the death of those components. This process continues until the last components merge at the ground level $p(x) = 0$ [14]. We study the 0-th homology groups and count the connected components at different heights via the he aforementioned approach [12].

We discretize the transcript levels using the first three principal components to analyze our single-cell data and regard the resulting discrete bins as different cellular states. The ideal bin size is determined by the density of the cells undergoing reprogramming and the dimensions of the bins. After determining the states to which every cell belongs, a frequency count is calculated for each state, which is proportional to the probability of how often a state appears throughout the single-cell space. Using these probabilities, we can examine the "peak-space" [12] and locate the peaks that represent a group or a sub-population of cells.

For each connected component, we place a dot according to its death probability $p_d(i)$ (*x*-axis) and its birth probability $p_b(i)$ (*y*-axis), respectively. Each dot on the persistence diagram corresponds to a probability peak. After all peaks are identified, we then examine the death probability for each peak in order to detect the component that the peak in question has merged into. Hence, with persistent homology, we can identify the bridge states connecting two components. This allows us to order the cellular states in a manner reflecting the differentiation time.

C. Identifying Stem cell Genes Using Bulk RNA-seq Data

In order to gain insight into the cells that we have identified as "transient cells", we use bulk RNA-sequencing of pluripotent stem cells to identify a group of signature genes that are highly expressed in stem cells compared to fibroblasts (NHDF). Through differential gene expression analysis, we have identified a set of 830 genes. We utilized this panel of signature genes to calculate a stem cell score, which is taken as the mean centered gene expression values per cell.

We average the expression of these genes for each cell within our time-course and generate an embryonic stem cell (ESC) "average" score. This is used to identify individual cells which are more similar to pluripotent stem cells due to expression of these signature genes.

III. RESULTS

The pre-processing of single-cell data involves the following steps: UMI (unique molecular identifier) correction, removal of doublets and cells with high

mitochondrial content, selection of cells with at least 200 detected genes and finally accepting genes that must be present in a minimum of 100 cells. To discover the bridge states amongst single cells, we have created a matrix where rows represent 25,417 cells and columns represent the first three principal components. These first three principal components captures sufficient variance at the transcriptome level while allowing us to avoid sparse representation of the space of cellular states.

After the identification of the connected components in the single-cell RNA-seq data, we selected the connected components harboring death states to pinpoint the transitory cells and explore whether they can provide topological order to the cellular reprogramming process. Specifically, a component with a death state indicates that it merges with another component. In other words, these components represent the group of cells that are following a certain trajectory in the course of reprogramming. Overall, we have uncovered fourteen components or sub-populations of cells that have moved along the trajectory (Fig 2).

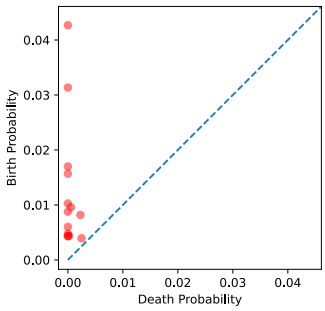

Fig. 2. Persistence diagram for the first three principal components where each dot represents a probability peak. The *x*-axis and *y*-axis represent the death and the birth probability, respectively.

Amongst these fourteen components, four significant components eventually merge with other components. We have visualized the transitory cells bridging these four components to other components on a two-dimensional PCA plot (Fig 3). Cells highlighted with solid colors in Fig 3 are the transient groups of cells on a successful path to complete the reprogramming process towards pluripotency. Altogether, there are 141 cells identified as "transient" with our novel approach.

The majority of these cells (121/141) are found in the D7 samples colored in green. As D7 samples are collected seven days post-infection, these cells have progressed further along the reprogramming process compared to cells collected at a very early time point of Day3 post-OSKM expression and parental cells (NHDF), which harbor fourteen and six transient cells, respectively.

As cells lose their fibroblast characteristics during reprogramming, a pluripotency-associated gene regulatory network (GRN) of transcripts must be established. We identified a set of signature genes that define the pluripotency GRN and are not expressed in the starting fibroblast through bulk RNA-seq analysis. We then calculated the sum of their average expression (stem cell score) for every cell in our time course.

A high stem cell score indicates high expression of these signature genes, and hence the more similar the profiled cells become to a pluripotent stem cell. Importantly, cells identified as transient using our persistent homology method exhibit the highest stem cell score out of all profiled cells, indicating that we can capture the individual cells that are successfully starting transitioning towards a pluripotent gene regulatory network (Fig 4).

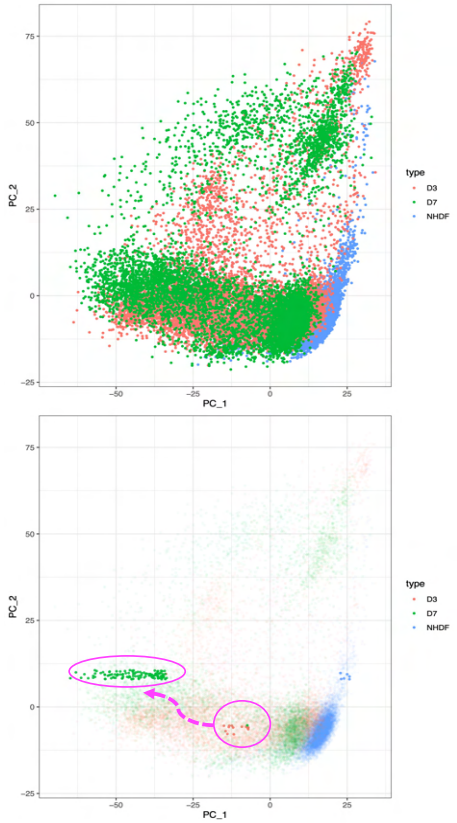

Fig. 3. Plots of principle component analysis of single cells. (Top) all the data points in the samples of single cells plotted by the first two PCAs. (Bottom) Cells from the four transition states that bridge different components of cell sub-populations. The arrow represents the transition direction. Colors blue, red, and green encode the human fibroblasts (NHDF), Day3 (D3), and Day7 (D7) samples, respectively.

To further validate that our captured "transient" cells are indeed expressing pluripotency-associated genes, we plotted the expression of the stem cell specific transcription factor NANOG (Fig 5). We find that only our "transient" population

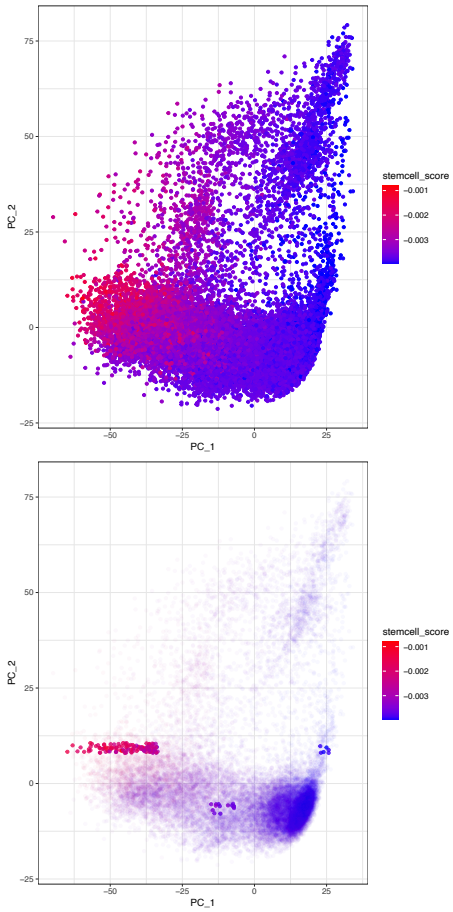

Fig. 4. PCA plots colored by the stem cell score (top) for all the data points in the samples and (bottom) for cells that belong to the four transition states.

of Day7 (D7) cells have up-regulated NANOG, a key event in the reprogramming process. These results indicate that our persistent homology method is capable of identifying the small number of cells that are successfully transitioning towards pluripotency withing a large and heterogeneous population of reprogramming cells.

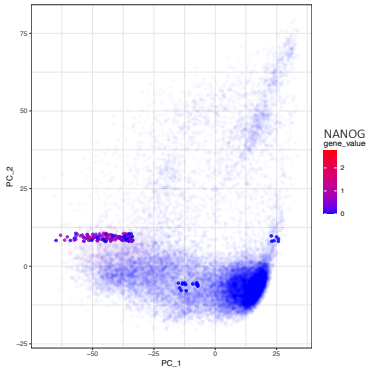

Fig. 5. The expression of NANOG gene in the transient cells. Transient cells of Day 7 exhibit heightened NANOG expression compared to other transient cells and the rest of non-transient cells.

IV. CONCLUSIONS

In this study, we have developed a new approach based on persistent homology to identify transient cells bridging different major cellular states. We applied this method to analyze single-cell RNA sequencing data from reprogrammed human fibroblasts. Even though we only included data representing the early stages of the reprogramming process, we successfully identified multiple density peaks that merge into other peaks, indicating the presence of transient cells in the overall cell population. Our current results indicate that major cellular phenotypes can be discovered from single-cell transcriptome and, more importantly, we can identify transient cell states bridging these major cellular phenotypes, a significant advantage of our approach. Upon successful additional analyses in understanding the nature of these transient cells, we plan to apply this method to study an extended range of reprogrammed cells to uncover the topological order of the entire reprogramming process.

REFERENCES

[1] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126(4):663–676, 2006.

[2] Constantinos Chronis, Petko Fiziev, Bernadett Papp, Stefan Butz, Giancarlo Bonora, Shan Sabri, Jason Ernst, and Kathrin Plath. Cooperative binding of transcription factors orchestrates reprogramming. *Cell*, 168(3):442–459, 2017.

[3] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[4] L McInnes, J Healy, N Saul, and L Großberger. Umap: uniform manifold approximation and projection. j. open source softw. 3, 861, 2018.

[5] Stefan Canzar Van Hoan Do. A generalization of t-sne and umap to single-cell multimodal omics. *Genome Biology*, 22, 2021.

[6] Tara Chari, Joeyta Banerjee, and Lior Pachter. The specious art of single-cell genomics. *bioRxiv*, 2021.

[7] Abbas H Rizvi, Pablo G Camara, Elena K Kandror, Thomas J Roberts, Ira Schieren, Tom Maniatis, and Raul Rabadan. Single-cell topological rna-seq analysis reveals insights into cellular differentiation and development. *Nature biotechnology*, 35(6):551–560, 2017.

[8] Hyekyoung Lee, Moo K Chung, Hyejin Kang, Bung-Nyun Kim, and Dong Soo Lee. Discriminative persistent homology of brain networks. In *2011 IEEE international symposium on biomedical imaging: from nano to macro*, pages 841–844. IEEE, 2011.

[9] Danielle S Bassett and Olaf Sporns. Network neuroscience. *Nature neuroscience*, 20(3):353–364, 2017.

[10] Andreas Hofmann, Matthias Krufczik, Dieter W Heermann, and Michael Hausmann. Using persistent homology as a new approach for super-resolution localization microscopy data analysis and classification of γh2ax foci/clusters. *International journal of molecular sciences*, 19(8):2263, 2018.

[11] Daniel Shnier, Mircea A Voineagu, and Irina Voineagu. Persistent homology analysis of brain transcriptome data in autism. *Journal of the Royal Society Interface*, 16(158):20190531, 2019.

[12] Anna Terebus, Farid Manuchehrfar, Youfang Cao, and Jie Liang. Exact probability landscapes of stochastic phenotype switching in feed-forward loops: Phase diagrams of multimodality. *Frontiers in Genetics*, 12, 2021.

[13] Herbert Edelsbrunner, John Harer, et al. Persistent homology-a survey. *Contemporary mathematics*, 453:257–282, 2008.

[14] Farid Manuchehrfar, Huiyu Li, Wei Tian, Ao Ma, and Jie Liang. Exact topology of the dynamic probability surface of an activated process by persistent homology. *The Journal of Physical Chemistry B*, 125(18):4667–4680, 2021.