Computational Method for Tumor Cell Detection in Single-Cell DNA Sequencing Data

Toluwanimi Ariyo, Department of Research, Biomedical Sciences Magnet, Ridge View High School

Department of Computer Science, University of Illinois-Urbana Champaign

National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health

Key Words: Single-Cell DNA sequencing, Tumor, Cancer, Machine Learning

ABSTRACT

Single-cell DNA sequencing (scDNA-seq) helps researchers study the evolutionary process of cancer. It is a process used to examine individual cells, describe intra-tumor heterogeneity, and reconstruct the evolutionary history of a tumor. Coverage is the number of reads at a given position in the genome. The depth of high-coverage scDNA-seq allows for analysis of point mutations while it is difficult to make these inferences within low-coverage scDNA-seq. However, due to the uniformity of coverage, ultra-low coverage scDNA-seq is ideal for copy number calling [6].

This study aims to develop a computational method, utilizing features computed from low-coverage scDNA-seq, to detect tumor cells and assist in future efforts of identifying technical errors. Data was pre-processed using Principal Component Analysis (PCA). A machine learning algorithm was implemented to detect tumor cells in this latent, dimensionally reduced space for two patients (patients S0 and S1) with breast cancer sequenced using 10x genomics. The training set (patient S0) had an accuracy of 98% for tumor cell detection. The testing set (patient S1) had an accuracy of 99% for tumor cell detection. This demonstrates that these features are useful for accurately detecting tumor cells in ultra-low coverage scDNA-seq data.

Spatial heterogeneity of tumor clones was observed, revealing correlations with cell type and sections. Doublet analysis revealed doublets concentrated between clusters, providing evidence that this feature may be useful for future detection of technical errors. Future studies will focus on improving the computational method for doublet detection and optimization of the tumor cell detection algorithm.

INTRODUCTION

Through its earliest documentations, cancer has been prevalent in several species, ranging from humans to birds, serving as a malignant growth occurring from uncontrolled cellular division. Utilizing natural selection and mutations, animals have evolved potent tumor suppressive mechanisms to prevent cancer development, which ultimately altered the development and architecture of several tissues [1]. Several researchers have attempted to understand cancer development at its various stages of growth in order to provide a better explanation regarding this widespread phenomenon. Cancer is classified as an evolutionary process: as species evolve by mutation and selection acting on individuals in a population, tumors evolve by mutation and natural selection acting on cells in a tissue [1]. This process of mutation and natural selection is integral to the evolution of cancer at every stage. Through extensive data, evolutionary theorists have been able to construct an evolutionary timeline of cancer, ranging from a normal cell to intra-tumor heterogeneity (Figure I). Recently, researchers have made efforts to study and reconstruct this evolutionary history of a tumor for a better understanding of cancer through single-cell DNA sequencing.

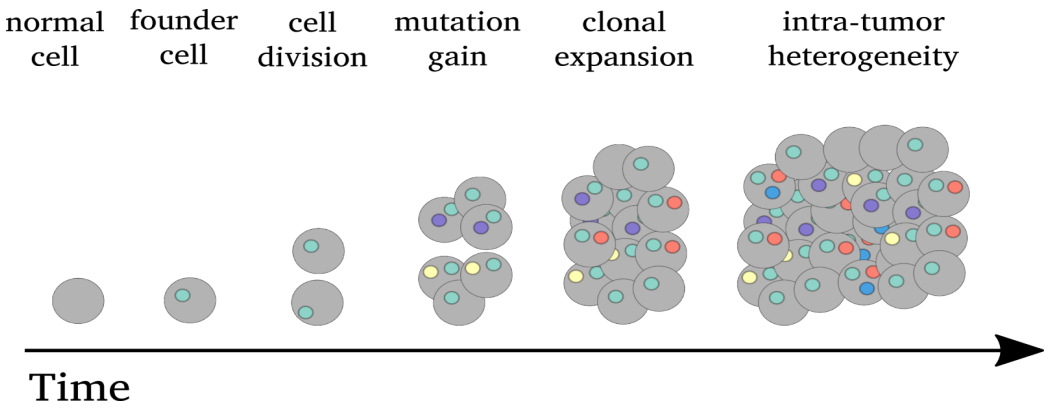

Figure I. Evolutionary Process of Cancer

Single-cell DNA sequencing (scDNA-seq) is a process used to detect the genome, transcriptome, and several segments of information from single cells [4]. This vital information may reveal cell population differences and cellular evolutionary history relationships. The increased use of single-cell sequencing for cancer research is providing a wealth of new insights regarding intra-tumor heterogeneity, metastasis, and the landscape of the tumor microenvironment [5]. This scDNA-seq process includes isolation, where each single-cell is isolated for further analysis, cycling to combine free DNA from the nucleus, barcoding of each droplet for later interpretation, amplification of DNA, and sequencing to produce reads that are aligned to a reference genome for further analysis. Coverage is the number of reads at a given position in the genome. The depth of high-coverage scDNA-seq allows for analysis of point mutations while it is difficult to make these inferences within low-coverage scDNA-seq. However, due to the uniformity of coverage, ultra-low coverage scDNA-seq is ideal for copy number calling and may reveal allele-specific mutations [6].

However, scDNA-seq generates high-dimensional data, which may require probing the expression level of all 20,000 different genes that can be expressed in cells for analysis. As a result, scDNA-seq generates low throughput and is more expensive compared to alternatives

such as mass cytometry [3]. Recently, researchers have resolved this issue by utilizing dimensionality reduction [6]. Specifically, principal component analysis (PCA) has been found to be effective in performing the task of dimensionality reduction. PCA is a mathematical algorithm that standardizes the range of continuous initial variables, computes the covariance matrix to identify correlations, computes the eigenvectors and eigenvalues of the covariance matrix to identify the principal components, creates a feature vector to identify necessary principal components, and recasts the data along the principal components axes (Figure II). This intricate method allows for dimensionality reduction, which can assist in mathematically modeling and further analysis with machine learning.

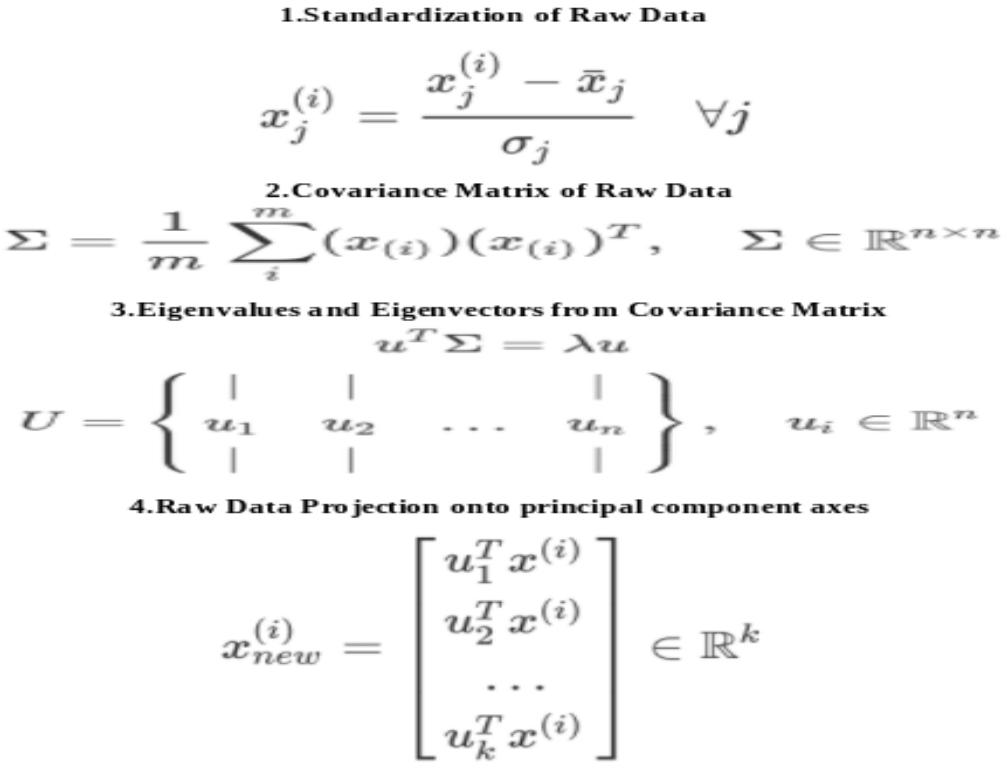

Figure II. Mathematical Algorithm of Principal Component Analysis (PCA)

Machine learning is a subset of artificial intelligence that utilizes data to construct algorithms that can automate analytical decisions, imitating an evolutionary process [8]. Several

algorithms have been constructed to perform different tasks within machine learning. Specifically, K-Nearest Neighbors (KNN) is a standard, supervised machine learning algorithm that is used for classification and regression. This algorithm is given a value of k to find the nearest neighbors to an indicated point, applying the euclidean distance for all points within the dataset, and the indicated point is classified based on the classes of the nearest neighbors.

Recent scientific advances have revealed tumor cells and other mutations within high-coverage scDNA-seq. In particular, one study has identified segments affected by copy-number aberrations and other mutations within high-coverage scDNA-seq [7]. However, due to copy-neutral loss of heterozygosity and other errors found within high-coverage scDNA-seq [6], tumor cell identification and further analysis of low-coverage scDNA-seq may prove integral to understanding the evolutionary history of tumors. This served as a reason for the development of a computational method to detect tumor cells within scDNA-seq. The present study aims to develop a computational method, utilizing features computed from low-coverage scDNA-seq and machine learning algorithms, to detect tumor cells and assist in future efforts of identifying technical errors.

MATERIALS AND METHODS

Materials. Patient S0 (Section A, B, C, D, E) scDNA-seq data, Patient S1 (Section A1, A2, A3, B1, B2) scDNA-seq data, Windows 10, Oracle VM Virtualbox 6.1.28, Ubuntu 20.04.1, packages and libraries: pandas, numpy, seaborn, matplotlib, sklearn, sklearn.decomposition (PCA), sklearn.neighbors (KNeighborsClassifier), linear_model, and preprocessing were utilized throughout the duration of this study.

Methods. Necessary packages and libraries were initially imported, namely, pandas, seaborn, matplotlib, and sklearn.decomposition (PCA) for the section analysis of Patient S0. Functions were initialized to process the Patient S0 scDNA-seq data. Specifically, create_rdr_baf(self) was a function initialized to select columns of interest for the Read-Depth Ratio (RDR) and B-Allele Frequency (BAF) by concatenating the chromosome, start, and end data into a single column and pivoting the data from vertical format to horizontal format. After the RDR and BAF were pivoted individually, the data was inner-merged to create a new dataframe, which allowed for accurate analysis as the cell sequence became identified within the index with RDR and BAF inputs for the columns. Another function, namely, pca_df(df) was created to perform the task of dimensionality reduction, specifically into 2 dimensions, creating a new dataframe with 'pca1' and 'pca2' columns. Subsequently, data files from Patient S0 were initialized into variables and read into dictionaries. This data was preprocessed with the use of create_rdr_baf(self) function, identifying "SECTION-BARCODE" as the index. Then, the dimensionality of the data was reduced into 2 dimensions with the use of the pca_df(df) function. The data was now visualized with the use of a seaborn relational plot, utilizing 'pca1' and 'pca2' section analysis data. This process was repeated for the section analysis of Patient S1 scDNA-seq data.

For the tumor cell analysis of Patient S0 scDNA-seq data, processed data was further analyzed with the pre-existing knowledge of normal cells. These normal cells were read into a dataframe and left-joined with Patient S0 scDNA-seq data. The resulting "NA" values in the new "TYPE" column were replaced with "TUMOR". This revealed the number of tumor cells and normal cells within the scDNA-seq data. This data was now visualized with the use of a seaborn relational plot, utilizing 'pca1' and 'pca2' tumor cell analysis data. This process was repeated for the tumor cell analysis of Patient S1 scDNA-seq data.

For the doublet analysis of Patient S1 scDNA-seq data, processed data was further analyzed with the pre-existing knowledge of doublets. These doublets were read into a dataframe, further processed for homogenous formatting, and left-joined with the processed Patient S1 scDNA-seq data. The resulting “NA” values in the new “DOUBLET” column were replaced with “NORMAL”. Then, the data was visualized with the use of a seaborn relational plot, utilizing ‘pca1’ and ‘pca2’ for doublet analysis and assisting in further efforts of detecting additional doublets.

Based on visualization of the patients’ scDNA-seq data, the KNN algorithm was considered an appropriate machine-learning algorithm for further analysis and tumor cell detection within the scDNA-seq data. Necessary packages and libraries were imported, namely, pandas, seaborn, matplotlib, sklearn.decomposition (PCA), numpy, sklearn, sklearn.neighbors (KNeighborsClassifier), linear_model, and preprocessing. Clonal data from Patient S1 was initialized into variables, concatenated into a single dataframe, and processed for further utilization. Normal cells were identified within the clonal data and faults were removed to reduce errors. The KNN algorithm was implemented with the initialization of x_train and x_test variables for the model, utilizing pca_df(df) and fit.transform(df) methods. A dataframe was created, containing the type of cells: normal cells and tumor cells. Preprocessing for the type of cells was implemented, creating an array with “1” corresponding to a tumor cell and “0” corresponding to a normal cell. The number of neighbors within the model was identified and an appropriate fitting was applied. The accuracy of the training set (Patient S0) and testing set (Patient S1) for tumor cell detection within the scDNA-seq data was found and subsequently printed.

RESULTS

Patient S0 scDNA-seq data and Patient S1 scDNA-seq data were accurately mathematically modeled within the PCA space, utilizing the origin of the section for classification. In Patient S1 scDNA-seq data, there was a significant correlation between the origin of section and corresponding coordinate positioning within the visualization (Figure I). However, within Patient S0 scDNA-seq data, there was not a significant correlation between the origin of section and corresponding coordinate positioning within the visualization (Figure I).

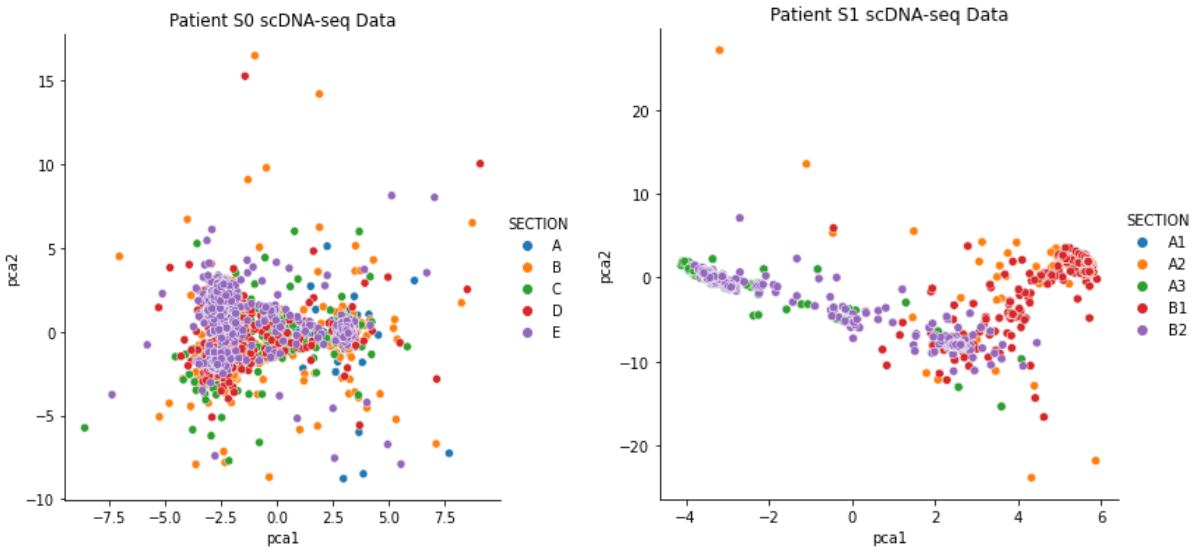

Figure I. Seaborn relational plot generated in the latent, dimensionally reduced space, visualizing the section analysis of Patient S0 scDNA-seq data and Patient S1 scDNA-seq data.

Patient S0 scDNA-seq data and Patient S1 scDNA-seq data were accurately mathematically modeled within the PCA space, utilizing doublets, an example of a technical error produced within scDNA-seq data, for classification. Identified doublets were concentrated within a specific location of the mathematical model space, residing between other clusters within the scDNA-seq data (Figure II).

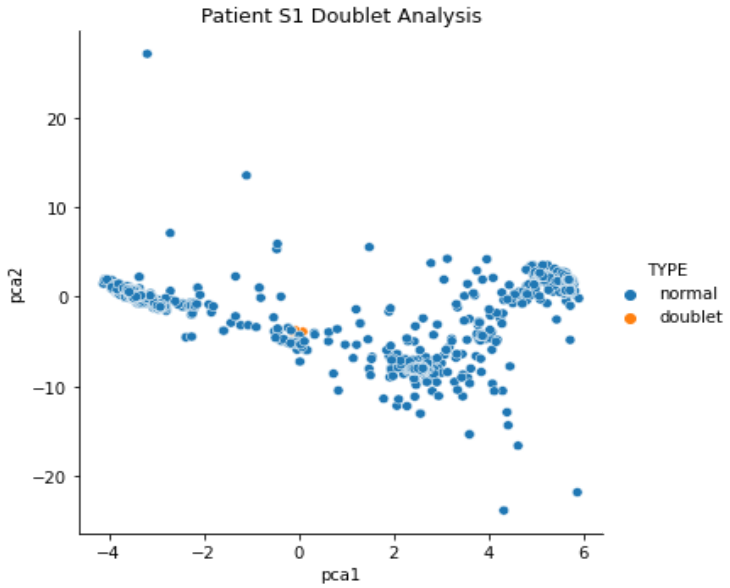

Figure II. Seaborn relational plot generated in the latent, dimensionally reduced space, visualizing the doublet analysis of Patient S1 scDNA-seq data.

Patient S0 scDNA-seq data and Patient S1 scDNA-seq data were accurately mathematically modeled within the PCA space, utilizing identified tumor cells for classification. In Patient S1 scDNA-seq data, there was a significant correlation between the identified tumor cells and corresponding coordinate positioning within the visualization (Figure III). Furthermore, within Patient S0 scDNA-seq data, there was a significant correlation between the tumor cells and corresponding coordinate positioning within the visualization (Figure III). In addition, spatial heterogeneity of tumor clones was observed, revealing correlations with cell type and sections (Figure III).

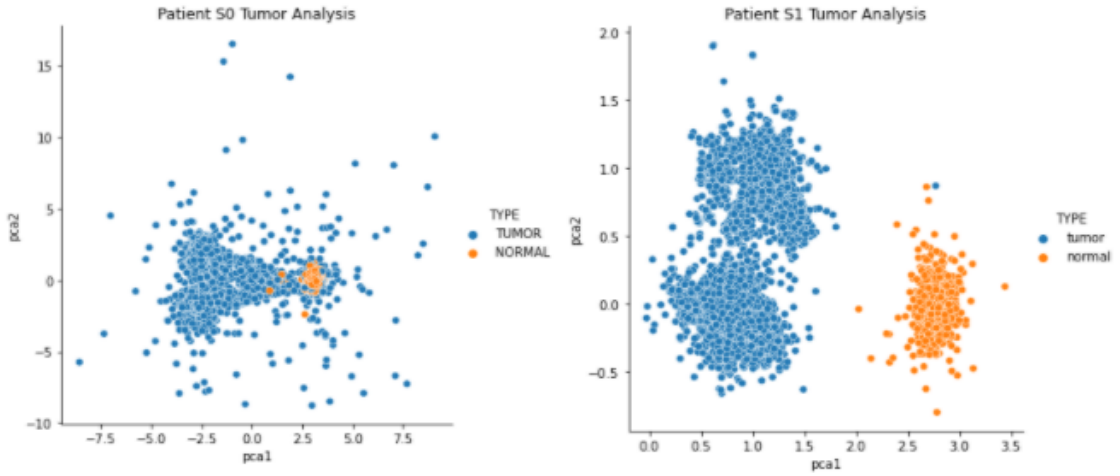

Figure III. Seaborn relational plot generated in the latent, dimensionally reduced space, visualizing the tumor analysis of Patient S0 scDNA-seq data and Patient S1 scDNA-seq data.

The machine learning algorithm, KNN, was implemented to detect tumor cells in this latent, dimensionally reduced space for the two patients (patients S0 and S1) with breast cancer sequenced using 10x genomics. The training set (patient S0) had an accuracy of 98% for tumor cell detection (Plate I). The testing set (patient S1) had an accuracy of 99% for tumor cell detection (Plate I).

```
vn = preprocessing.LabelEncoder() #Allows for preprocessing
cls = vn.fit_transform(list(pca_result["TYPE"])) #1 is Tumor, 0 is Normal
ycls = vn.fit_transform(list(x_testfinal['TYPE']))
x_testa = x_testfinal[['pca1', 'pca2']].values

model = KNeighborsClassifier(n_neighbors=5)
model.fit(x_train, cls)

acc_test = model.score(x_testa, ycls)
acc_train = model.score(x_train, cls)
print("Test Accuracy: ", acc_test)
print("Train Accuracy:", acc_train)
```

```
Test Accuracy:  0.9965928449744463
Train Accuracy: 0.9832385806704568
```

Plate I. Image of the Implementation of K-Nearest Neighbor Algorithm for tumor cell detection in the scDNA-seq data. The "Test Accuracy" corresponds to the accuracy of tumor cell detection within the Patient S1 scDNA-seq data. The "Train Accuracy" corresponds to the accuracy of tumor cell detection within the Patient S0 scDNA-seq data.

DISCUSSION

There was a significant correlation between the origin of section and corresponding coordinate positioning within the section analysis of Patient S1 scDNA-seq data (Figure I). This finding is supported by one study in the literature that has provided comparable analysis of sections within high-coverage scDNA-seq data [7]. However, there was no significant correlation between the origin of section and corresponding coordinate positioning within the section analysis of Patient S0 scDNA-seq data, providing knowledge that the section analysis method may only prove useful for specific patients (Figure I). Further analysis will be done to examine this occurrence and improve the reliability of the section analysis method.

Identified doublets were concentrated within a specific location of the mathematical model space, residing between other clusters within the doublet analysis of Patient S1

scDNA-seq data (Figure II). This finding provides evidence that the doublet analysis feature may prove useful for further analysis. The coordinates of the known doublets may allow for further interpretation of unknown doublets within proximity, utilizing an artificial intelligence algorithm for classification. Future studies will focus on improving the doublet analysis method to detect more technical errors within our dataset

In the tumor cell analysis of Patient S1 scDNA-seq data, there was a significant correlation between the identified tumor cells and corresponding coordinate positioning within the visualization (Figure III). Furthermore, within the tumor cell analysis of Patient S0 scDNA-seq data, there was a significant correlation between the tumor cells and corresponding coordinate positioning within the visualization (Figure III). Spatial heterogeneity of tumor clones was also observed, revealing correlations with cell type and sections (Figure III). These findings provide knowledge that the tumor cell analysis method may prove useful and reliable among various patients. In addition, this visualization allowed for further analysis with the application of a machine learning algorithm to detect additional tumor cells within the scDNA-seq data.

The training set (patient S0) of the machine learning algorithm resulted in an accuracy of approximately 98% for tumor cell detection (Plate I). The testing set (patient S1) resulted in an accuracy of approximately 99% for tumor cell detection (Plate I). These findings affirm that the KNN algorithm accurately detects tumor cells within low-coverage scDNA-seq data over numerous patients, providing evidence of its reliability. One study in the literature has observed similar trends to this finding with high-coverage scDNA-seq data [7].

The computational method performed extensive analysis on low-coverage scDNA-seq data, revealing origins of sections, doublets, and tumor cells. Utilization of this computational method may reveal several other significant factors that may aid in further analysis of

low-coverage scDNA-seq data. The accuracy of tumor cell detection from the machine learning algorithm allows for further implication that this computational method may play a crucial role in advanced tumor cell detection across numerous patient scDNA-seq data. Future studies will focus on improving the computational method for doublet detection and optimization of the tumor cell detection algorithm.

ACKNOWLEDGEMENTS

This research is supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health, Research Grant Number: R25DK078382. Mentoring throughout the research was provided by Mohammed El-Kebir, PhD and Leah Weber, PhD Candidate from the Department of Computer Science, University of Illinois Urbana-Champaign. Patient single-cell DNA sequencing data was provided by Simone Zaccaria, PhD, from the Department of Computer Science, Princeton University. Additional support was provided by the Department of Research, Biomedical Sciences Magnet, Ridge View High School.

LITERATURE CITED

1. Casás-Selves, M., & Degregori, J. (2011). How cancer shapes evolution, and how evolution shapes cancer. *Evolution*, 4(4), 624–634.
2. Ghodsi, A. (2006). Dimensionality reduction a short tutorial. *Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada*, 37(38).
3. Spitzer, M. H., & Nolan, G. P. (2016). Mass Cytometry: Single Cells, Many Features. *Cell*, 165(4), 780–791.

4. Tang, X., Huang, Y., Lei, J. et al. (2019). The single-cell sequencing: new developments and medical applications. *Cell Biosci*, 9, 53.
5. Weber, L. L., Sashittal, P., & El-Kebir, M. (2021). doubletD: detecting doublets in single-cell DNA sequencing data. *Bioinformatics*, 37(Supplement_1), i214-i221.
6. Zaccaria, S., & Raphael, B. J. (2021). Characterizing allele-and haplotype-specific copy numbers in single cells with CHISEL. *Nature biotechnology*, 39(2), 207-214.
7. Zaccaria, S., Raphael, B.J. (2020). Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *Nat Commun*, 11, 4301.
8. Zhang XD. (2020). Machine Learning. In: A Matrix Algebra Approach to Artificial Intelligence. *Springer, Singapore,* 8(6).