PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Unravelling cancer subtype-specific driver genes in single-cell transcriptomics data with CSDGI

Meng Huang[1,2☯], **Jiangtao Ma**[1,3☯], **Guangqi An**[4], **Xiucai Ye**[2*]

1 Department of Automation, Xiamen University, Xiamen, China, **2** Department of Computer Science, University of Tsukuba, Tsukuba, Japan, **3** School of Engineering, Dali University, Dali, Yunnan, China, **4** Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan

☯ These authors contributed equally to this work.
* yexiucai@cs.tsukuba.ac.jp

OPEN ACCESS

Citation: Huang M, Ma J, An G, Ye X (2023) Unravelling cancer subtype-specific driver genes in single-cell transcriptomics data with CSDGI. PLoS Comput Biol 19(12): e1011450. https://doi.org/10.1371/journal.pcbi.1011450

Editor: Simone Romeni, Ecole Polytechnique Federale de Lausanne, SWITZERLAND

Received: August 22, 2023

Accepted: December 5, 2023

Published: December 14, 2023

Data Availability Statement: The source code used to replicate all our analyses, including all real data, is available at the following link: https://github.com/linxi159/CSDGI.

Funding: This work was partly supported by JSTSPRING (JPMJSP2124) for MH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Cancer is known as a heterogeneous disease. Cancer driver genes (CDGs) need to be inferred for understanding tumor heterogeneity in cancer. However, the existing computational methods have identified many common CDGs. A key challenge exploring cancer progression is to infer cancer subtype-specific driver genes (CSDGs), which provides guidane for the diagnosis, treatment and prognosis of cancer. The significant advancements in single-cell RNA-sequencing (scRNA-seq) technologies have opened up new possibilities for studying human cancers at the individual cell level. In this study, we develop a novel unsupervised method, *CSDGI* (Cancer Subtype-specific Driver Gene Inference), which applies Encoder-Decoder-Framework consisting of low-rank residual neural networks to inferring driver genes corresponding to potential cancer subtypes at the single-cell level. To infer CSDGs, we apply *CSDGI* to the tumor single-cell transcriptomics data. To filter the redundant genes before driver gene inference, we perform the differential expression genes (DEGs). The experimental results demonstrate *CSDGI* is effective to infer driver genes that are cancer subtype-specific. Functional and disease enrichment analysis shows these inferred CSDGs indicate the key biological processes and disease pathways. *CSDGI* is the first method to explore cancer driver genes at the cancer subtype level. We believe that it can be a useful method to understand the mechanisms of cell transformation driving tumours.

Author summary

Cancer is recognized as a complex disease with diverse characteristics. In order to comprehend the diversity within tumors, it is essential to infer cancer subtype-specific driver genes (CSDGs), which offer valuable insights for investigating cancer progression and treatment. The remarkable progress made in single-cell RNA-sequencing (scRNA-seq) technologies has ushered in new prospects for studying human cancers at the cellular level. Cancer Subtype-specific Driver Gene Inference (CSDGI) is a novel unsupervised

method proposed. In our study, we use Encoder-Decoder-Framework to infer driver genes specific to cancer subtypes in the CSDGI. We apply CSDGI to three tumor single-cell transcriptomics data. The experimental results have shown the effectiveness of CSDGI. Furthermore, functional and disease enrichment analyses illustrate that these inferred CSDGs shed light on crucial biological processes and disease pathways. Our collection of driver genes will serve as a valuable resource in unraveling the mechanisms driving cell transformation in tumors.

Introduction

Cancer is a heterogeneous disease characterized by the unstable cellular growth [1,2], which is caused by a set of genes. These genes can drive tumorigenesis as drivers, and thereby are known as cancer driver genes (CDGs) [1]. They may affect the homeostatic development for a collection of critical cellular function containing cell proliferation, cell differentiation, cell death, etc. To advance the research of tumor emergence and evolution, it is crucial to infer CDGs [3,4]. Further, cancer molecular subtypes play a crucial role in deepening our insights of cancer within tumor cell subsets as representing a heterogeneous disease instead of a single disease [5], which may suggest therapeutic opportunities. Therefore, the discovery of these CDGs across cancer subtypes helps explore cancer progression. To guide the diagnosis, treatment and prognosis of cancer, it is also necessary to infer cancer subtype-specific driver genes (CSDGs).

Previous studies have shown that CDGs may be tissue-specific [6] and condition-specific [7,8]. For example, CDGs can be identified in a large number of tumor samples across cancer subtypes using bulk transcriptome data by RNA sequencing (RNA-seq) from The Cancer Genome Atlas (TCGA) consortium [6,9]. To infer the driver gene in large cohorts, many projects have been performed in the whole genome of thousands of tumour samples under the International Cancer Genome Consortium (ICGC) [10, 11]. Besides, transcriptomic studies have supported most cancer subtype discoveries rather than other omics data [12]. Nevertheless, previous computational methods [13–16] using bulking transcriptome data mostly identify CDGs involved in all of cells, rather than CDGs specific to cell subpopulations (cancer subtypes). This also shows that the previous studies fail to directly infer cancer subtype-specific driver genes in transcriptomic data.

This may conceal the heterogeneity of CDGs across different cancer subtypes. Fortunately, more transcriptome studies have become possible due to the advancements in single-cell RNA-sequencing (scRNA-seq) technologies [17–20]. In contrast to traditional bulk RNA-seq providing average gene expression of cell groups [12,21], scRNA-seq allows for the quantification of gene expression in individual cells, which facilitates the analysis of cellular differences [22–25]. Cancer is no longer considered as a uniform entity but is rather subdivided into various subtypes. This subtype classification is crucial for precision medicine. Understanding subtype-specific driver genes in cancer can aid in the development of more targeted treatment approaches. By targeting these subtype-specific driver genes, researchers can develop more effective drugs and treatment strategies, and provide personalized treatment and early diagnosis. This is particularly valuable for cancer-related research, as scRNA-seq allows for the identification and analysis of intratumoral heterogeneity [26–28]. As a result, researchers can obtain an in-depth understanding of cancer subtypes, cellular populations, and the functions of individual cells within tumor samples at single cell level [24,29,30]. Accordingly, the scRNA-seq can provide a way to infer CDGs at the single-cell or cancer-subtype level.

By using bulk transcriptome and other omic data, several methods have been developed to identify CDGs. For example, Akavia et al. developed CONEXIC to unveil potential CDGs in the deleted tumor region by combining gene expression data with copy number change [13]. Ng et al. proposed PARADIGM-SHIFT to infer gene activity utilizing pathway-level information, gene expression and copy number, which helps predict mutation drivers in cancer processes [14]. Paull et al. developed TieDIE to find small cancer driver pathway by using genomic and transcriptomic perturbations, which may predict transcription factor in cancer [15]. Chen et al. presented MAXDRIVER to infer CDGs by integrating genomic data and heterogeneous networks [16]. However, the above computational methods are limited to unveil CDGs specific to tissues or samples since they do not take into account the relationships between genes and cancer subtypes. To understand the heterogeneity of cancer across different cancer subtypes, it is necessary to infer cancer subtype-specific driver genes.

In ths work, we present a novel unsupervised computational method, *CSDGI* (Cancer Subtype-specific Driver Gene Inference) to infer driver genes specific to potential cancer subtypes at single-cell level. By only using single-cell transcriptomics data rather than a large cohort, the proposed *CSDGI* method applies the Encoder-Decoder-Framework that consists of low-rank residual neural network to selecting genes that are most associated with the identified cancer subtypes. *CSDGI* ranks potential driver genes based on the association between genes and cancer subtypes. In each cancer subtype identification module, genes with higher rankings are more likely to be drivers specific to the current cancer subtype. To filter the redundant genes, the differential expression genes (DEGs) are performed before inferring driver gene. We apply *CSDGI* to the real tumor single-cell transcriptomics datasets, which infers CSDGs relating to cancer subtypes. The experimental results demonstrate that *CSDGI* can be an effective method to infer driver genes for identified cancer subtypes. Especially, the associated driver genes can be applied to explore the interpretable biological meaning of each cancer subtype. Systematic gene ontology and disease enrichment analysis demonstrates the potential functions and key biological processes of identified CSDGs. *CSDGI* is the first method to explore CSDGs at the cancer subtype level. These CSDGs and the proposed methodology provide a new way to better understand the mechanisms of tumorigenesis, help classify cancer subtypes and explore the genotypic status of tumors. We believe that *CSDGI* can be a useful method to understand the mechanisms of cell transformation driving tumours, and cancer progression.

Materials

Single-cell transcriptomics data in the human cancer

In this study, we obtain three single-cell transcriptomics datasets from the data repository NCBI Gene Expression Omnibus. These data include melanoma, breast cancer, and chronic myeloid leukemia gene expression data from GSE72056 [28], GSE75688 [29] and GSE76312 [30]. The summary of these datasets is provided in Table 1. In the melanoma dataset (GSE72056), these cells comprise 1257 malignant melanoma tumor cells and 3388 benign tumor cells. Besides, the breast cancer dataset (GSE75688) consists of 317 tumor cells and 198 nontumor cells. Both GSE72056 and GSE75688 datasets were transformed using log*TPM* as

Table 1. The summary of single-cell transcriptomics data of human cancer used in this study.

The scRNA-seq data	# Number of cells	# Number of genes	Data source
Melanoma dataset	4,645	23,686	GEO access number: GSE72056[28]
Breast cancer dataset	515	57,915	GEO access number: GSE75688[29]
Chronic myeloid leukemia dataset	1,134	23,384	GEO access number: GSE76312[30]

https://doi.org/10.1371/journal.pcbi.1011450.t001

inputs for the model. Additionally, the chronic myeloid leukemia dataset (GSE76312) includes cells from the human chronic myeloid leukemia, comprising 902 tumor cells and 232 normal cells. The dataset from GSE76312 was transformed by log*RPKM* as inputs of model.

Data pre-processing

To investigate the intrinsic transcriptomic signatures of tumor cells, we apply a filtering process to the gene expression data. We filter out genes that are expressed in less than t% of cells, where t% is set to 6% based on the previous study [31]. Additionally, we also filter out genes that are expressed in more than (100−t)% of cells, as these ubiquitous genes are not helpful in inferring cancer subtypes. Following the gene filtering, we select the most r% variable genes by controlling the relationship between mean expression and variability. After gene filtering, there are 12693 genes in the melanoma dataset, 15451 genes in the breast cancer dataset, and 10862 genes in the chronic myeloid leukemia dataset. Subsequently, we utilize a R tool (EMDomics) [32] to obtain differential expression genes (DEGs) between tumor cells and benign (nontumor or normal) cells. As a result, we obtain 820 DEGs in the breast cancer dataset, 1048 DEGs in the chronic myeloid leukemia dataset, and 1170 DEGs in the melanoma dataset. These selected DEGs also help infer CSDGs in the breast tumor cells, the chronic myeloid leukemia tumor cells and the malignant melanoma tumor cells. More detailed information about DEGs can be found in S1 File.

Methods

Proposed method overview

To identify cancer subtype-specific driver genes, we propose a novel unsupervised method, cancer subtype-specific driver gene inference (CSDGI) by using Encoder-Decoder-Framework. This workflow is shown in Fig 1. Firstly, we download real tumor scRNA-seq datasets including nontumor cells and tumor cells. These data are preprocessed to obtain the gene expression data X. Then, we apply Encoder-Decoder-Framework to infer driver genes for each cancer subtype. We rank genes according to their low-rank weights in the Encoder-Decoder-Framework, i.e., the importance of a gene in various cancer subtypes will result in its higher ranking. The output of the rank display the overall impact of each gene for every cancer subtypes. Finally, we perform downstream analysis with cancer subtype-specific driver genes.

Encoder-Decoder-Framework

Encoder-Decoder-Framework consists of low-rank encoder module and decoder reconstruction module. As shown in Fig 1, the single-cell gene expression matrix is represented as $X = [X_1, X_2, X_3, \cdots, X_m]^T \in \mathbb{R}^{m\times n}$, where $X_i \in R^{n\times 1}$ indicate the gene expression profile of the i-th cell. Here, m and n denote the number of cells and genes, respectively. k is the number of estimated cancer subtypes In Encoder, motivated by the interpretable feature clustering method and the automatic association feature learning in scRNA data [33,34], we apply the low-rank residual network to encoding the low-rank representation of each subtype. In detail, the gene expression of each cell X_i, $i \in \{1,2,\ldots,m\}$ is input to the input layer. The deeper residual neural network (ResNet) [35] aims to redefine the layers by acquiring an understanding of residual functions associated with the input layer. The relationships between genes are often complex, which exhibit nonlinear characteristics. In the ResNet, the learnable parameters represent the complex relationships between genes, which is shared across each subtype. This shows that Encoder can represent nonlinear properties among genes in cells for the propsed Encoder-Decoder-Framework. The number of residual blocks is p. The objective function of

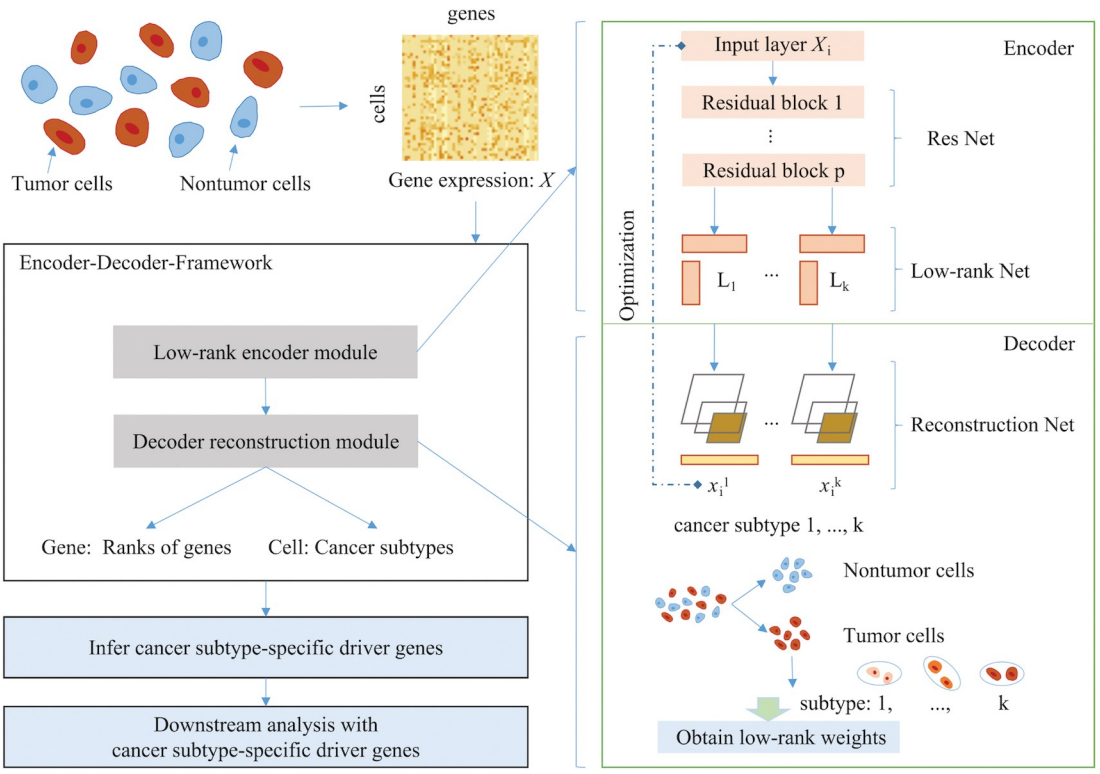

Fig 1. Workflow of CSDGI. After data pre-processing for real tumor scRNA-seq datasets, we infer different driver genes for different cancer subtypes using Encoder-Decoder-Framework and perform downstream analysis with cancer subtype-specific driver genes.

https://doi.org/10.1371/journal.pcbi.1011450.g001

ResNet is as follows:

$$X_i^{l+1} = X_i^l + H(X_i^l, W^l), \tag{1}$$

$$X_i^p = X_i^l + \sum_{j=l}^{p-1} H(X_i^j, W^j), \tag{2}$$

where H is the non-linear function representation, where the activate function is ReLU and the bias is b, $H(X_i^j, W^j) = ReLU(W^j * X_i^j + b^j)$. Let $l = 0, X_i^l = X_i^0 = X_i$.

To represent the gene association across different genes in each cancer subtype, we utilize low-rank network that is performed using low-rank matrix [36,37]. Here, we view L_k as the low-rank (r-rank) graph of cancer subtype k. The function of low-rank network is as follows:

$$D_i^k = tanh(X_i^p(V_k - diag(V_k))), \tag{3}$$

where $L_k = (B_k)^{-1} V_k (V_k)^T$, $V_k(V_k)^T$ is the r-rank matrix, $V_k \in R^{n \times r}$, the degree matrix of $V_k(V_k)^T$ is B_k, the output is $D_i^k \in R^{n \times 1}$. Empirically, we set r to 1.

In Decoder, we use the deep neural network to implement the Reconstruction Net. To obtain the reconstructed gene expressions x_i^k, we perform the following calculations:

$$x_i^k = M^k \circ F_i^k * + E_i^k, i = 1, 2, \ldots, m, \tag{4}$$

where $F_i^k = tanh(G^j * D_i^k + e^j)$, G and e are the weight and bias, $\circ$ is the dot product, M and E are the learnable weight matrix and bias for x_i^k.

To minimize the loss between source gene value X_i and reconstructed gene value x_i^k, we apply the mean squared error (MSE) to them. The residual between X_i and x_i^k is calculated as follows:

$$L = \frac{1}{m}\sum_{i=1}^{m} \min \parallel (X_i - x_i^k) \circ \lambda^k \parallel_2^2, k \in \{1, \ldots, s\}, \tag{5}$$

where m is the number of cells, s is the maximum number value for cancer subtype index k, the low-rank weight λ^k is the absolute value of μ_k, μ_k is the first nontrivial left eigenvector of $(I - L_k)$ of low-rank matrix in the Encoder. After obtaining optimal loss, the subtype index k with the minimum loss indicates a subtype label of each cell.

In Eq (5), it can be observed that the relevant genes of cancer subtype k have significantly higher absolute values compared to the irrelevant genes. These differences arises from their strong intra-associations with relevant genes and weak associations with irrelevant genes. As a result, we utilize the λ^k to ensure that X_i is assigned to the appropriate cancer subtype k. Its limitation is that we need to enter k in advance. Furthermore, k value shows that cancer subtypes are not identical in this proposed method.

Gene inference and downstream analysis

To infer CSDGs, we calculate the values of low-rank weight $\lambda^k \in R^{n\times 1}$ corresponding of each cancer subtype k after obtaining the optimal model of Encoder-Decoder-Framework. We calculate the differences between the generated cell samples and the original cell samples in each cancer subtype, respectively. Next, we obtain the subtype label of each cell that is assigned according to the minimum loss of model. In each cancer subtype, each value of λ^k is sorted in descending order according to the weight values, where λ^k indicates the importances of genes in the cancer subtype k. We define the top 5% of genes as cancer subtype-specific driver genes in the sorted value of weight λ^k. In order to explore the biological meaning of CSDGs, we perform downstream analysis including subtype classification, functional and disease enrichment analysis.

Measurement metrics

To quantitatively evaluate the performance of CSDGI, we adapt the accuracy metric for classification and adjusted rand index (ARI) for clustering. Theset metrics are defined as follows:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives}, \tag{6}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}, \tag{7}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}, \tag{8}$$

$$F_1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{9}$$

$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)}, \tag{10}$$

where True Positives represents the number of positive samples that are correctly classified as

positive, True Negatives represents the number of negative samples that are correctly classified as negative, False Positives represents the number of negative samples that are incorrectly classified as positive, False Negatives represents the number of positive samples that are incorrectly classified as negative, RI is the rand index and ARI is the adjusted rand index to assess the cluster performance.

Parameter settings

To identify unique driver genes distinguishing tumor cells from nontumor cells in three real tumor single-cell transcriptomics datasets, we set the value of k to 2. To explore different driver genes associated with inferred cancer subtypes within tumor cells, the factoextra (R package tool) is used to determine the cluster number of tumor cells. Based on the results of the cluster number analysis, k is set to 2, 2, and 3 in the proposed CSDGI method for the breast tumor cells, the chronic myeloid leukemia tumor cells, and the malignant melanoma tumor cells, respectivcely. All model optimization experiments of CSDGI are iteratively conducted for 400 epochs using a single NVIDIA Tesla Ampere A100-PCIe-40GB GPU and the Ubuntu 18.04 system. Furthermore, all evaluation experiments are run independently 100 times and the average results are shown in the study.

Results

Differential expression genes analysis

To filter redundant genes, we use a R tool (EMDomics) to obtain DEGs between tumor cells and nontumor cells. Fig 2 presents different density plots of genes illustrating the distribution of various genes in three real tumor scRNA-seq datasets. Besides, we obtain all emd scores and p-value of preprocessing genes by using the EMDomics tool. We focus on genes having higher emd scores with p-value < 0.05. For example, the emd score results with melanoma dataset (GSE72056) are shown in Fig 2A–2C. Fig 2A demonstrates that the EMD score is 0.01 for the gene RP11-567G24.1, which represents the lowest value, while Fig 2C exhibits the highest EMD score of 23.9 for the gene TMSB4X in the breast cancer dataset. Fig 2B displays the density plot for the gene C18orf32, which shows 7.01 EMD scores. Notably, higher EMD scores indicate a greater distinction between tumor cells (group A) and nontumor cells (group B). Consequently, we select 820 genes with EMD scores ranging from 7 to 23.9 as DEGs in the breast cancer dataset. Similarly, Fig 2D–2F show that we obtain 820 genes with emd scores varing from 2 to 13.57 as DEGs in the chronic myeloid leukemia dataset. As shown in Fig 2G–2I, 1170 genes with emd scores more than 7 are regarded as DEGs in the melanoma dataset. More comprehensive information of DEGs from the above three tumor datasets can be found in S1 File.

Distinguishable performance analysis between tumor and nontumor

To quantitatively evaluate the proposed CSDGI method, we run it between nontumor cells and tumor cells in the breast cancer, chronic myeloid leukemia and melanoma data. For all real tumor scRNA-seq data, we set the parameter about the cluster k to 2 in the CSDGI framework to select distinguishable genes between nontumor cells and tumor cells. As shown in S2 File, we list selected top-20 genes respectively for breast cancer, chronic myeloid leukemia and the melanoma data. Here, we perform classification and clustering for selected top-20 genes. To compare different methods, we use Recursive Feature Elimination (RFE) and Chi-square Test (Chi2) as gene identification methods to select genes. For the classification task, we use Support Vector Machine (SVM) and Random Forest (RF) as classification models. For the

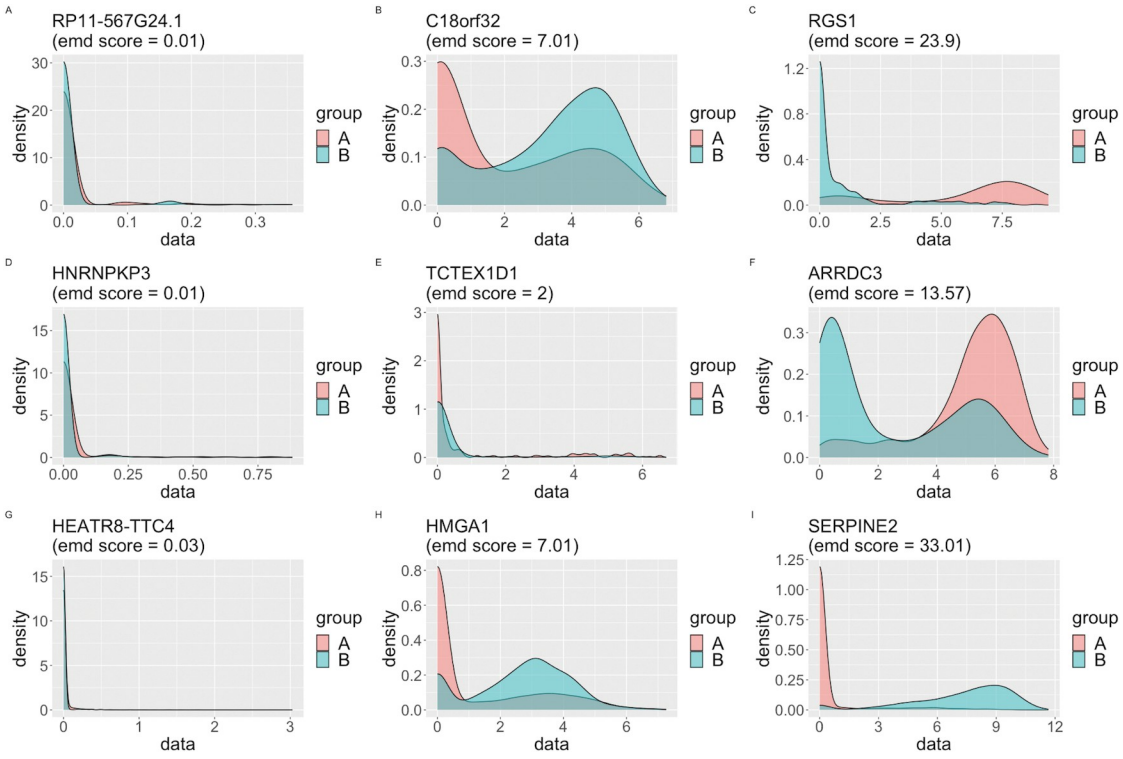

Fig 2. The density plot of all genes using the EMDomics tool between nontumor cells and tumor cells. (A, B, C) The density plots with RP11-567G24.1 (gene), C18orf32 (DEG) and RGS1 (DEG) in the breast cancer dataset. (D, E, F) The density plots with HNRNPKP3 (gene), TCTEX1D1 (DEG) and ARRDC3 (DEG) in the chronic myeloid leukemia dataset. (G, H, I) The density plots with HEATR8-TTC4 (gene), HMGA1 (DEG) and SERPINE2 (DEG) in the melanoma dataset. Group A represents the set of tumor cells. Group B represents the set of nontumor cells.

https://doi.org/10.1371/journal.pcbi.1011450.g002

clustering task, we use Gaussian Mixed Model (GMM) and K-means clustering (K-means) [38] as clustering models. As shown in Fig 3A, 3D and 3G, we find the classification accuracy increases progressively as the number of genes increases for different gene identification methods and classification models. However, the accuracy stabilizes when the number of selected genes fluctuates from 10 to 20. The proposed CSDGI method has outperformed other methods for SVM and RF classification tasks in all cases of selected genes from 2 to 20. Similarly, we also use F1-score to evaluate classification tasks. As shown in Fig 3B, 3E and 3H, our method has still the best classification results than other methods. The reasonable explanation is that the proposed method places its emphasis on identifying and selecting the most relevant genes with specific cell subpopulation. These selected genes may be considered to be the potential driver genes for influencing the characteristics of cancer subtypes. For the clustering tasks, we also compare different methods by using GMM and K-means in the different numbers of selected genes from 2 to 20. As shown in Fig 3C, 3F and 3I, we can find that the clustering ARI of each method has obvious fluctuations with the number of genes from 2 to 8 rather than continuously increasing. Besides, as the increasing number of genes from 8 to 20, the clustering ARI is sometimes reduced instead of rising steadily. However, our method has still outperformed other methods in most cases of selected gene numbers from 2 to 20. Therefore, the proposed CSDGI can effectively identify potential driver genes, which helps distinguish different cancer subtypes more correctly. These above classification and clustering results in the breast cancer dataset, chronic myeloid leukemia dataset and melanoma dataset also show that these selected genes have superior distinguishability for characterizing different cell

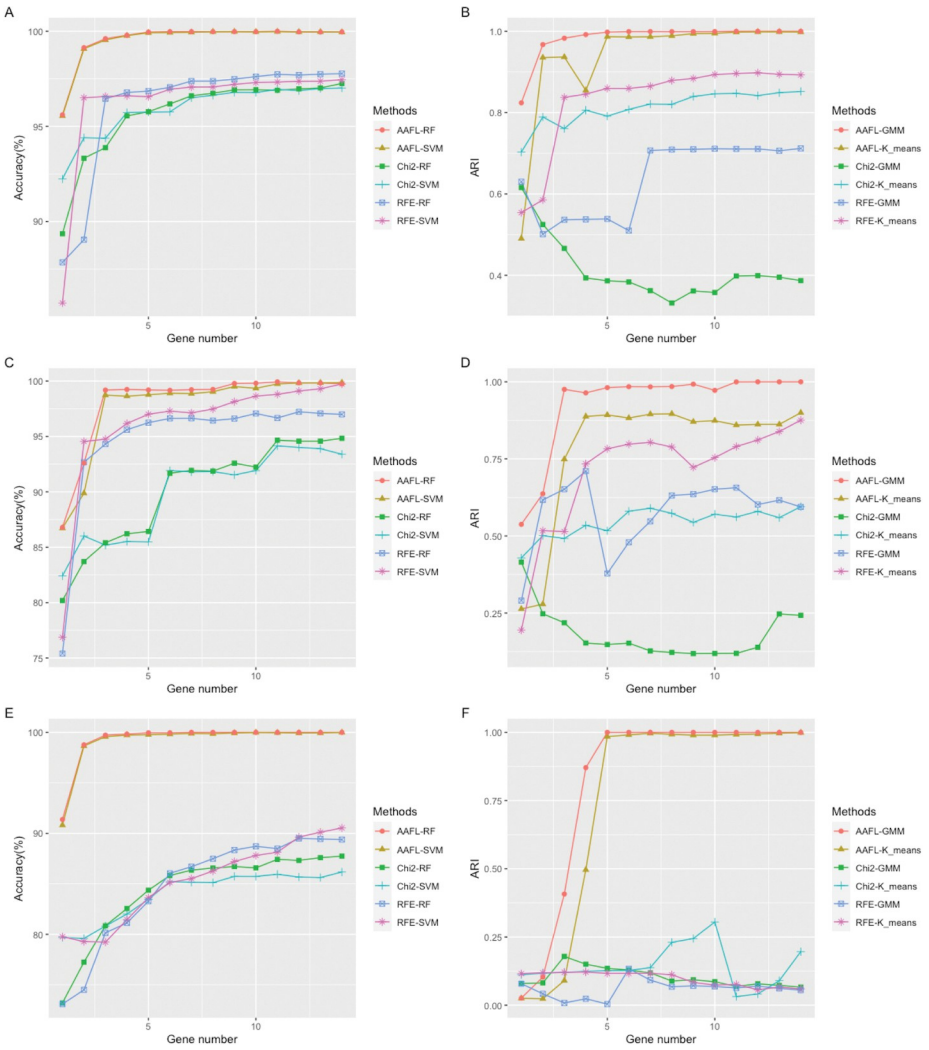

Fig 3. Classification and clustering evaluation results for selected top 20 genes in three real tumor scRNA-seq datasets. (A, B, C) The Accuracy, F1-score and ARI plots in the breast cancer dataset. (D, E, F) The accuracy, F1-score and ARI plots in the chronic myeloid leukemia dataset. (G, H, I) The accuracy, F1-score and ARI plots in the melanoma dataset. For the comparison of different methods, the number of selected genes is 2, 4, 6, 8, 10 12 14, 16, 18 and 20, respectively.

https://doi.org/10.1371/journal.pcbi.1011450.g003

subpopulations. More importantly, these identified driver genes between known cell types and potential cancer subtypes may unveil the important driver genes that drive the transformation from non-tumor cells to tumor cells (cancer subtype).

Estimation of cancer subtypes

To identify cancer subtypes, we apply the proposed CSDGI to the tumor cells in the breast cancer, chronic myeloid leukemia and melanoma data. After excluding nontumor, normal and benign tumor cells, there are 317 breast tumor cells, 902 chronic myeloid leukemia tumor cells and 1257 malignant melanoma cells in these real datasets. Due to the absence of a ground truth of cancer subtypes, we fail to obtain the clustering information of cancer subtypes. Previous studies have acknowledged that it is difficult to determine the number of clusters in a clustering task [38]. To analyze real tumor cells, the sum of squared error (SSE) is employed to

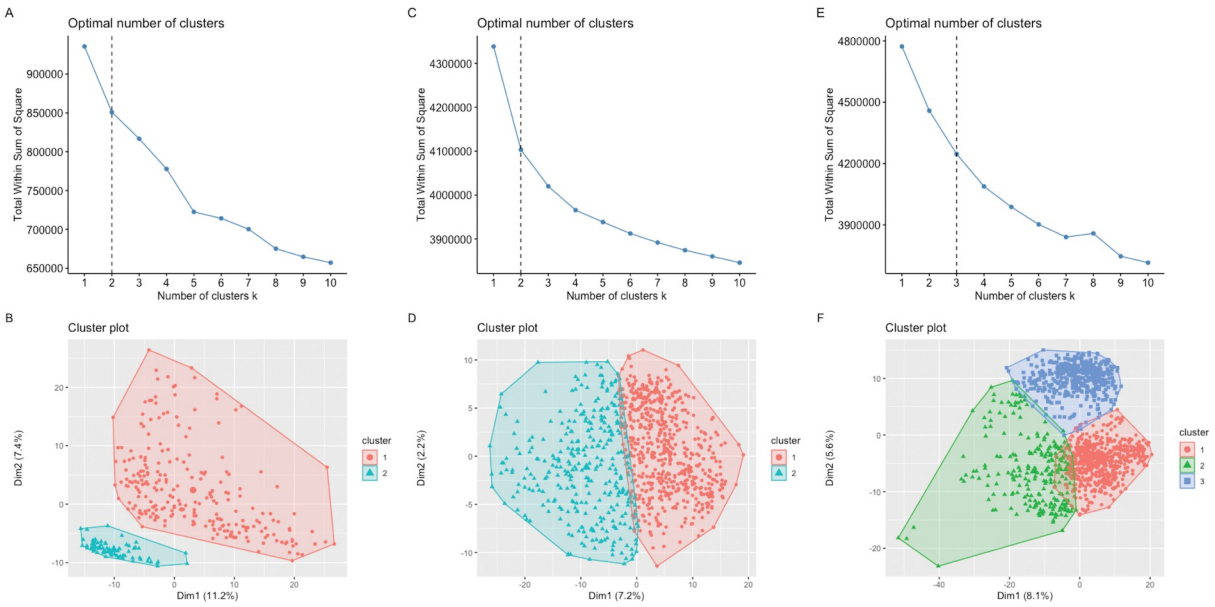

Fig 4. Estimate the number of cancer subtypes for three real tumor cells datasets. (A, B) The optimal value of cancer subtype number and visualization of tumor cells with two possible cancer subtypes in the breast cancer tumor cells. (C, D) The optimal value of cancer subtype number and visualization of tumor cells with two possible cancer subtypes in the chronic myeloid leukemia tumor cells. (E, F) The optimal value of cancer subtype number and visualization of malignant tumor cells with three possible cancer subtypes in the malignant melanoma tumor cells.

https://doi.org/10.1371/journal.pcbi.1011450.g004

identify the number of potential cancer subtypes, where the SSE funnction tool is implemented in the R package (factoextra). The R package "factoextra" is to establish clusters through k-means clustering, a method that aims to minimize the total within-cluster sum of squares (WSS). This WSS metric quantifies the compactness of the clusters, and the objective is to minimize it to achieve more tightly-knit clusters. As shown in Fig 4A, 4C and 4E, we investigate various cancer subtype number ranging from 1 to 10 by using the Sum of Squared Errors (SSE) method about cluster number identification. Fig 4A, 4C and 4E exhibit the elbow point for each tumor cell dataset, which indicates significant changes in the gradient between 1 and 10. Consequently, we select 2, 2, 3 as the optimal value of cancer subtype number, respectively, for the breast cancer tumor cells, chronic myeloid leukemia tumor cells and the malignant melanoma cells. In Fig 4B, 4D and 4F, we also visualize breast tumor cells, chronic myeloid leukemia tumor cells and malignant melanoma cells according to the optimal number. To apply the proposed CSDGI method to three real tumor cell data, we set the parameter k to 2, 2, 3, respectively. For these tumor cells from three real cancer datasets, more detailed information about cancer subtypes can be found in S3 File.

Discovering cancer subtype-specific driver genes

To infer driver genes specific to each cancer subtype, we obtain the learnable low-rank weight in the proposed unsupervised CSDGI framework. Here, we identify top 5% genes as driver genes corresponding to each cancer subtype. As shown in Table 2, we obtain 41 driver genes for each cancer subtype of breast tumor cells, respectively. These underlined genes are genes shared by Subtype 1 and Subtype 2. For the chronic myeloid leukemia tumor cells and the malignant melanoma cells, we infer 52 and 59 driver genes for each cancer subtype, respectively. As shown in S4 File, we list all detailed information of driver genes. Besides, we also highlight the shared driver genes of each cancer subtype by using the bold in S4 File.

Table 2. The identified cancer subtype-specific driver genes in the breast tumor cells.

Cancer subtypes	Driver genes
Subtype 1	*MGP*, *PIP*, *SCGB2A2*, *MUCL1*, *GSTP1*, *HLA-DRA*, *CD74*, *LDHB*, *TFF3*, *SLPI*, *APOD*, *PPP1R1B*, *PPP1R14C*, *FITM2*, *S100P*, *RN7SL56P*, *SEPP1*, *AGR2*, *MRPL45*, *TFF1*, *RBP1*, *PLEKHA5*, *CEACAM6*, *SNCG*, *MIEN1*, *TM4SF1*, *MT1G*, *MGST1*, *RP11-658F2.8*, *AGR3*, *HLA-B*, *GSTM3*, *HLA-A*, *MDK*, *CLU*, *ARHGDIB*, *ORMDL3*, *TMEM45A*, *TCEAL1*, *CISD3*, *AZGP1*
Subtype 2	*MGP*, *CTTN*, *SSR4*, *NDUFC2*, *CLU*, *S100A11*, *NDUFS5*, *AGR2*, *MGST1*, *TFF3*, *SRP9*, *EGR1*, *DYNLT1*, *TXNDC17*, *MIF*, *MT1X*, *SOX4*, *PPFIA1*, *AZGP1*, *HLA-DRA*, *PSENEN*, *KRT19*, *CD9*, *HLA-B*, *TM4SF1*, *KRT7*, *SCGB2A2*, *MT2A*, *S100A16*, *ERGIC3*, *KRTCAP2*, *PPAPDC1B*, *UQCRQ*, *EFNA1*, *ARF4*, *CD24P4*, *PPA1*, *CD63*, *POLR2K*, *ATRAID*, *CD46*

Notes:These underlined genes are genes shared by Subtype 1 and Subtype 2.

https://doi.org/10.1371/journal.pcbi.1011450.t002

To unveil the biological meaning of driver genes, we perform a more detailed analysis. We calculate correlations and p-values according the pearson correlation coefficient, and then the gene co-expression network can be built based on these results. As shown in Fig 5, we obtain driver gene co-expression network of each cancer subtype by we obtain driver gene co-expression network of each cancer subtype by calculating correlation between driver genes. In Table 2, we can find that *MGP*, *SCGB2A2*, *HLA*, *TFF3*, *AGR2*, *TM4SF1*, *MGST1*, *HLA-B*, *CLU* and *AZGP1* are shared in two subtypes for the breast tumor cells. Previous studies show that *MGP* promotes the breast cancer proliferation [39], *SCGB2A2* and *TFF3* can be viewed a

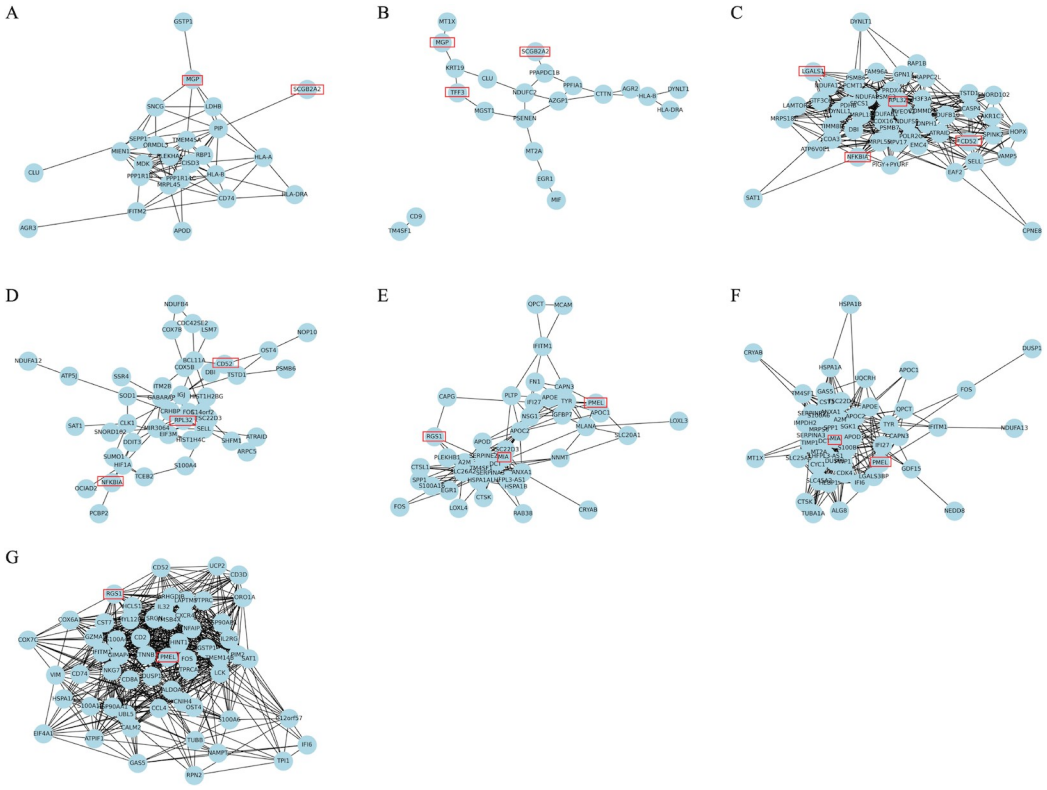

Fig 5. Driver gene co-expression network of each cancer subtype for three real tumor cells datasets. (A, B) Cancer subtype 1 and 2 in the breast cancer tumor cells. (C, D) Cancer subtype 1 and 2 in the chronic myeloid leukemia tumor cells. (E, F, G) Cancer subtype 1, 2 and 3 in the malignant melanoma tumor cells.

https://doi.org/10.1371/journal.pcbi.1011450.g005

valuable predictive biomarker in breast cancer [40,41]. *MUCL1* are identified as driver genes in subtype 1, which have been found to be critical markers of breast cancer [42]. In Fig 5A and 5B, *MGP*, *SCGB2A2* and *TFF3* have been marked with red box. For the chronic myeloid leukemia tumor cells, the shared driver genes between subtype 1 and 2 include *RPL32*, *SAT1*, *SNORD102*, *CD52*, *TSTD1*, *SELL*, *ATRAID*, *PSMB6*, *NFKBIA*, *NDUFA12* and *DBI* in S4 File. Previous evidences illustrate that *RPL32* plays an important role in the SF3B1 mutation for chronic myeloid leukemia (CML) disease progression [43], *CD52* helps identify molecular signature of CML [44] and *NFKBIA* plays a strategic role in CML molecular response [45]. In addition, *LGALS1* is identified as driver genes in the CML cancer subtype 1 and previous works also find that *LGALS1* can be selected as a critical biomarker of CML [46]. In Fig 5C and 5D, *RPL32*, *CD52*, *NFKBIA* and *LGALS1* have been marked with red box. For the malignant melanoma cells, *PMEL*, *IFITM1*, *HSPA1A*, *DUSP1*, *FOS* and *TMSB4X* are the shared driver genes across three subtypes in S4 File. Previous studies have indicated that *PMEL* is a important link between Parkinson's disease and melanoma [47]. Other researchers have also regarded it as a prognostic predictor in skin cutaneous melanoma (SKCM) [48]. Furthermore, the shared *RGS* between Subtype 1 and Subtype 3, and the shared *MIA* between Subtype 1 and Subtype 2 have been viewed as important biomarkers in malignant melanoma [49,50]. In Fig 5E–5G, *PMEL*, *RGS1*and *MIA* have been marked with red box. Overall, these cancer subtype-specific driver genes indicates more significant biological meaning to study the breast tumor cells, chronic myeloid leukemia tumor cells and malignant melanoma cells.

Functional and disease enrichment analysis for CSDGs

To investigate the potential biological significance of CSDGs inferred by our CSDGI method, we perform the functional and disease enrichment analysis. We use Gene Ontology (GO, http://www.geneontology.org/) to analyze the biological function: biological process (BP), cellular component (CC), and molecular function (MF). Besides, the Kyoto Encyclopedia of Genes and Genomes Pathway (KEGG, http://www.genome.jp/kegg/) and Reactome Pathway (Reactome, http://reactome.org/) are also included in functioanal enrichment analysis. For the disease enrichment analysis, we use Disease Ontology (DO, http://disease-ontology.org/), DisGeNET (http://www.disgenet.org/) and Network of Cancer Genes database (NCG, http://ncg.kcl.ac.uk/) to analyze the inferred driver genes of each cancer subtype. As a result, we set the threshold of significance enrichment *p-value* to 0.01. As shown in S5 File, we list comprehensive information on functional and disease enrichment analysis about CSDGs for the breast cancer tumor cells, the chronic myeloid leukemia tumor cells and the malignant melanoma cells. Previous studies show that cell aggregation and mineral absorption have a significant influence on the progression of breast cancer [51,52]. Polyamine biosynthetic process promotes CML tumor growth [53] and ferroptosis may be a novel strategy for chronic myeloid leukemia anti-tumor therapy [54]. Cell aggregation, mineral absorption, polyamine biosynthetic process, and ferroptosis can be viewed as important enrichment analysis results in in S5 File. For the malignant melanoma cells, the enrichment analysis also indicates more important melanoma-related biological processes including melanin biosynthetic process, melanin metabolic process and melanosome organization in S5 File. Furthermore, we find that genes enriched on the same term type are different driver genes of different cancer subtypes for a group of tumor cells. This may be because the informative CSDGs exhibit differences between cancer subtypes. All in all, these inferred CSDGs can serve as valuable references to drive cancer subtype growth and finding new tumor driver markers, which illustrates the biological meaning of tumor development and helps understand the mechanisms of cell transformation driving tumours.

Discussion and conclusions

Cancer is a heterogeneous disease, where cancer driver genes can drive tumorigenesis and the unstable cellular growth. CDGs show the characteristics of tissue-specific or condition-specific. Thus, to deeply understand the cancer progression mechanisam at the cancer subtype level, we need to uncover CSDGs. However, most of the existing computational methods have mainly used the cohort information rather than the cancer subtype-specific information to infer CDGs. Tumors of different subtype are highly heterogeneous. Therefore, we need to identify CDGs from tumor cells of each cancer subtype. In this work, we use the unsupervised learning mechanism to propose a novel CSDGI computational method. At single-cell level, we use CSDGI to infer CSDGs by only considering single-cell transcriptomics data. In the proposed CSDGI, the Encoder-Decoder-Framework helps identify potential cancer subtypes. Furthermore, the application to the breast cancer tumor cells, the chronic myeloid leukemia tumor cells and the malignant melanoma tumor cells indicates that CSDGI may be a new solution to explain cancer subtype molecular mechnism.

One possible limitation of the current CSDGI framework is that these inferred CSDGs rely on the cancer subtype identification. The different tumor cells in the same cancer subtype may result in different CSDGs. Another limitation is that the current CSDGI method may be better suited to identify CDGs in the near 1000-dimension gene space. For the high-dimension gene expression data, such as near 20,000-dimension gene space without gene filtering, the proposed CSDGI could not infer CSDGs more efficiently. In the future, we consider to use genomics data including biological sequence informations to uncover more cancer-related driver genes. The existing methods can be combined into our framework, such as BioSeq-Diabolo [55], BioSeq-BLM [56]. In summary, we believe that CSDGI is fundamental to explore specific biological function involved in tumorigenesis across cancer subtypes, and improve existing driver gene identification methods. The cancer subtype-specific framework may provide a new solution to deeply understand the mechanisms of cell transformation and cancer progression, and give a shortlist of biologically meaningful genes that have potential to promote precision cancer medicine.

Supporting information

S1 File. Differential expression genes in three scRNA-seq datasets.
(XLSX)

S2 File. Selected top 20 genes in three scRNA-seq datasets.
(XLSX)

S3 File. Identified cancer subtypes in three scRNA-seq datasets.
(XLSX)

S4 File. The identified cancer subtype-specific driver genes in the chronic myeloid leukemia tumor cells and the malignant melanoma cells.
(XLSX)

S5 File. Functional enrichment analysis and disease enrichment analysis in three scRNA-seq datasets.
(XLSX)

Author Contributions

Conceptualization: Xiucai Ye.

Data curation: Meng Huang.

Formal analysis: Meng Huang, Jiangtao Ma, Guangqi An, Xiucai Ye.

Investigation: Meng Huang, Jiangtao Ma.

Methodology: Meng Huang, Jiangtao Ma, Xiucai Ye.

Project administration: Xiucai Ye.

Resources: Guangqi An, Xiucai Ye.

Software: Meng Huang, Jiangtao Ma.

Supervision: Xiucai Ye.

Validation: Meng Huang.

Visualization: Meng Huang, Guangqi An.

Writing – original draft: Meng Huang, Jiangtao Ma.

Writing – review & editing: Meng Huang, Jiangtao Ma, Xiucai Ye.

References

1. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458(7239): 719–724. https://doi.org/10.1038/nature07943 PMID: 19360079
2. Stratton MR. Exploring the genomes of cancer cells: progress and promise. Science. 2011; 331 (6024):1553–1558. https://doi.org/10.1126/science.1204040 PMID: 21436442
3. Reddy EP, Reynolds RK, Santos E, Barbacid M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. Nature. 1982; 300 (5888):149–152. https://doi.org/10.1038/300149a0 PMID: 7133135
4. Tabin CJ. et al. Mechanism of activation of a human oncogene. Nature.1982; 300(5888): 143–149. https://doi.org/10.1038/300143a0 PMID: 6290897
5. Chen Fengju, Chandrashekar, Darshan S, Varambally, Sooryanarayana, et al. Pan-cancer molecular subtypes revealed by mass-spectrometry-based proteomic characterization of more than 500 human cancers. Nat Commun. 2019; 10(1):5679. https://doi.org/10.1038/s41467-019-13528-0 PMID: 31831737
6. Tamborero D, Gonzalez-Pere A, Perez-Llama C, Deu-Pons J, Reimand J. Comprehensive identification of mutational cancer driver genes across 12 tumor types. Sci Rep. 2013; 3:2650. https://doi.org/10.1038/srep02650 PMID: 24084849
7. Guo WF, Zhang SW, Zeng T, Akutsu T, Chen L. (). Network control principles for identifying personalized driver genes in cancer. Briefings Bioinf. 2020; 21(5):1641–1662.
8. Sathyanarayanan A, Gupta R, Thompson EW, Nyholt DR, Bauer DC, Nagaraj SH. A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. Briefings Bioinf. 2020; 21(6), 1920–1936. https://doi.org/10.1093/bib/bbz121 PMID: 31774481
9. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrot K, Stuart JM, et al. The cancer genome atlas pan-cancer analysis project. Nat. Genet. 2013; 45(10): 1113–1120. https://doi.org/10.1038/ng.2764 PMID: 24071849
10. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. (2020). A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020; 20(10):555–572. https://doi.org/10.1038/s41568-020-0290-x PMID: 32778778
11. ICGC. International network of cancer genome projects. Nature. 2010; 464(7291):993–998 (2010). https://doi.org/10.1038/nature08987 PMID: 20393554
12. Oesper L, Mahmoody A, Raphael BJ. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biol. 2013; 14(7):1–21. https://doi.org/10.1186/gb-2013-14-7-r80 PMID: 23895164
13. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, et al. An integrated approach to uncover drivers of cancer. Cell. 2010; 143(6):1005–1017. https://doi.org/10.1016/j.cell.2010.11.013 PMID: 21129771

14. Ng S, Collisson EA, Sokolov A, Goldstein T, Gonzalez-Perez A, Lopez-Bigas N, et al. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. Bioinformatics. 2012; 28(18):i640–i646. https://doi.org/10.1093/bioinformatics/bts402 PMID: 22962493

15. Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, Stuart JM. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). Bioinformatics. 2013; 29(21):2757–2764. https://doi.org/10.1093/bioinformatics/btt471 PMID: 23986566

16. Chen Y, Hao J, Jiang W, He T, Zhang X, Jiang T, et al. Identifying potential cancer driver genes by genomic data integration. Sci Rep. 2013; 3(1):3538. https://doi.org/10.1038/srep03538 PMID: 24346768

17. Nawy T. Single-cell sequencing. Nat Methods. 2014; 11(1): 18–18. https://doi.org/10.1038/nmeth.2771 PMID: 24524131

18. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014; 343(6172): 776–779. https://doi.org/10.1126/science.1247651 PMID: 24531970

19. Gierahn TM, Wadsworth MH, Hughes TK, Gierahn TM, Wadsworth MH, Hughes TK, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat Methods. 2017; 14(4): 395–398. https://doi.org/10.1038/nmeth.4179 PMID: 28192419

20. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017; 8(1): 14049. https://doi.org/10.1038/ncomms14049 PMID: 28091601

21. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, et al. PyClone: statistical inference of clonal population structure in cancer. Nat Methods. 2014; 11(4): 396–398. https://doi.org/10.1038/nmeth.2883 PMID: 24633410

22. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011; 472(7341): 90–94. https://doi.org/10.1038/nature09807 PMID: 21399628

23. Pollen AA, Nowakowski TJ, Chen JD, Retallack H, Sandoval-Espinosa C, Nicholas CR, et al. Molecular identity of human outer radial glia during cortical development. Cell. 2015; 163(1): 55–67. https://doi.org/10.1016/j.cell.2015.09.004 PMID: 26406371

24. Qi R, Ma A, Ma Q, Zou Q. Clustering and classification methods for single-cell RNA-sequencing data. Briefings Bioinf. 2020; 21(4): 1196–1208. https://doi.org/10.1093/bib/bbz062 PMID: 31271412

25. Lake BB, Ai R, Kaeser GE, Salathia NS, Yung YC, Liu R, et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science. 2016; 352(6293): 1586–1590. https://doi.org/10.1126/science.aaf1204 PMID: 27339989

26. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014; 344(6190): 1396–1401. https://doi.org/10.1126/science.1254257 PMID: 24925914

27. Guo X, Zhang YY, Zheng LT, Zheng C, Song J, Zhang Q, et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. Nat Med. 2018; 24(7): 978–985. https://doi.org/10.1038/s41591-018-0045-3 PMID: 29942094

28. Peng JY, Sun BF, Chen CY, Zhou JY, Chen YS, Chen H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res. 2019; 29(9): 725–738. https://doi.org/10.1038/s41422-019-0195-y PMID: 31273297

29. Kanter I, Dalerba P, Kalisky T. A cluster robustness score for identifying cell subpopulations in single cell gene expression datasets from heterogeneous tissues and tumors. Bioinformatics. 2019; 35(6): 962–971. https://doi.org/10.1093/bioinformatics/bty708 PMID: 30165506

30. Davis-Marcisak EF, Sherman TD, Orugunta P, Stein-O'Brien GL, Puram SV, Roussos Torres ET, et al. Differential variation analysis enables detection of tumor heterogeneity using single-cell RNA-sequencing data. Cancer Res. 2019; 79(19): 5102–5112. https://doi.org/10.1158/0008-5472.CAN-18-3882 PMID: 31337651

31. Caliński T, Harabasz J. A dendrite method for cluster analysis. Commun Stat-theor M. 1974; 3(1): 1–27.

32. Nabavi S, Schmolze D, Maitituoheti M, Malladi S, Beck AH. EMDomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. Bioinformatics. 2016; 32(4): 533–541. https://doi.org/10.1093/bioinformatics/btv634 PMID: 26515818

33. Huang H, Xue F, Yan W, Wang T, Yoo S, Xu C. Learning Associations between Features and Clusters: An Interpretable Deep Clustering Method. In: 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021; p. 1–10.

34. Huang M, Long C, and Ma J. AAFL: automatic association feature learning for gene signature identification of cancer subtypes in single-cell RNA-seq data. Briefings Funct Genomics.2023; elac047. https://doi.org/10.1093/bfgp/elac047 PMID: 37122141

35. He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; p.770–778.

36. Zorzi M, Chiuso A. A Bayesian approach to sparse plus low rank network identification. 54th IEEE Conference on Decision and Control (CDC). 2015; p.7386–7391.

37. Zorzi M, Chiuso A. Sparse plus low rank network identification: A nonparametric approach. Automatica. 2017; 76: 355–366.

38. Pollard D. Quantization and the method of k-means. *IEEE Trans Inf Theory*. 1982; 28(2): 199–205.

39. Gong C, Zou J, Zhang M, Zhang J, Xu S, Zhu S, et al. Upregulation of MGP by HOXC8 promotes the proliferation, migration, and EMT processes of triple-negative breast cancer. Mol Carcinog. 2019; 58 (10): 1863–1875. https://doi.org/10.1002/mc.23079 PMID: 31264274

40. Talaat IM, Hachim MY, Hachim IY, Ibrahim RAER, Ahmed MAER, Tayel HY. Bone marrow mammaglobin-1 (SCGB2A2) immunohistochemistry expression as a breast cancer specific marker for early detection of bone marrow micrometastases. Sci. Rep. 2020; 10(1): 1–12.

41. May FEB, Westley BR. TFF3 is a valuable predictive biomarker of endocrine response in metastatic breast cancer. Endocr.-Relat. Cancer. 2015; 22(3): 465.

42. Li QH, Liu ZZ, Ge YN, Liu X, Xie XD, Zheng ZD, et al. Small breast epithelial mucin promotes the invasion and metastasis of breast cancer cells via promoting epithelial-to-mesenchymal transition. Oncol Rep. 2020; 44(2): 509–518.

43. Fernández-Martínez JL, deAndrés-Galiana EJ, Sonis ST. Genomic data integration in chronic lymphocytic leukemia. J Gene Med. 2017; 19(1–2): e2936. https://doi.org/10.1002/jgm.2936 PMID: 27928896

44. Zheng C, Li L, Haak M, Zheng C, Li L, Haak M, Brors B, Frank O, Giehl M, et al. Gene expression profiling of CD34+ cells identifies a molecular signature of chronic myeloid leukemia blast crisis. Leukemia. 2006; 20(6), 1028–1034. https://doi.org/10.1038/sj.leu.2404227 PMID: 16617318

45. Pungolino E, D'adda M, Trojani A, Perego A, Elena C, Iurlo A, et al. Jak-2 and Nfkbia Gene Expression Play a Strategic Role in Chronic Myeloid Leukemia (CML) Molecular Response during Early Nilotinib Treatment: The PhilosoPhi34 Data. Blood. 2018; 132: 5118.

46. Xu J, Wu M, Zhu S, Lei J, Gao J. Detecting the stable point of therapeutic effect of chronic myeloid leukemia based on dynamic network biomarkers. BMC Bioinf. 2019; 20(7): 73–81. https://doi.org/10.1186/s12859-019-2738-0 PMID: 31074387

47. Dean DN, Lee JC. Linking Parkinson's Disease and Melanoma: Interplay Between α-Synuclein and Pmel17 Amyloid Formation. Mov Disord. 2021; 36(7): 1489–1498.

48. Zhang S, Chen K, Liu H, Jing C, Zhang X, Qu C, et al. PMEL as a prognostic biomarker and negatively associated with immune infiltration in skin cutaneous melanoma (SKCM). J Immunother (Hagerstown, Md.: 1997). 2021; 44(6): 214. https://doi.org/10.1097/CJI.0000000000000374 PMID: 34028390

49. Sun MY, Wang Y, Zhu J, Lv C., Wu K, Wang XW, et al. Critical role for non-GAP function of Gαs in RGS1-mediated promotion of melanoma progression through AKT and ERK phosphorylation. Oncol Rep. 2018; 39(6): 2673–2680.

50. Li C, Liu J, Jiang L, Xu J, Ren A, Lin Y, et al. The value of melanoma inhibitory activity and LDH with melanoma patients in a Chinese population. Medicine. 2021; 100(8). https://doi.org/10.1097/MD.0000000000024840 PMID: 33663104

51. Liu X, Taftaf R, Kawaguchi M, Chang YF, Chen W, Entenberg D, et al. Homophilic CD44 interactions mediate tumor cell aggregation and polyclonal metastasis in patient-derived breast cancer models. Cancer discovery. 2019; 9(1): 96–113. https://doi.org/10.1158/2159-8290.CD-18-0065 PMID: 30361447

52. Mendes PMV, Bezerra DLC, Dos Santos LR, de Oliveira Santos R, de Sousa Melo SR, Morais JBS, et al. Magnesium in breast cancer: what is its influence on the progression of this disease?. Biol Trace Elem Res. 2018; 184:334–339. https://doi.org/10.1007/s12011-017-1207-8 PMID: 29198048

53. Sari IN, Yang YG, Wijaya YT, Jun N, Lee S, Kim KS, et al. AMD1 is required for the maintenance of leukemic stem cells and promotes chronic myeloid leukemic growth. Oncogene. 2021; 40(3), 603–617. https://doi.org/10.1038/s41388-020-01547-x PMID: 33203990

54. Zhang J, Liu Y, Li Q, Xu A, Hu Y, Sun C. Ferroptosis in hematological malignancies and its potential network with abnormal tumor metabolism. Biomed Pharmacother. 2022; 148:112747. https://doi.org/10.1016/j.biopha.2022.112747 PMID: 35240523

55. Li H and Liu B. BioSeq-Diabolo: Biological sequence similarity analysis using Diabolo, PLOS Computational Biology. 2023; 19(6), e1011214. https://doi.org/10.1371/journal.pcbi.1011214 PMID: 37339155

56. Li H, Pang Y and Liu B. BioSeq-BLM: a platform for analyzing DNA, RNA and protein sequences based on biological language models. Nucleic acids research. 2021; 49(22), e129–e129.