Joint Inference of Clonal Structure using Single-cell DNA-Seq and RNA-Seq data

Xiangqi Bai[1,2] Lin Wan[1,2,*]
Li C. Xia[3, *]
[1]NCMIS, LSC, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Division of Oncology, Department of Medicine,
Stanford University School of Medicine, Stanford, CA 94305, United States

Abstract

Understanding how genome changes shape gene expression in individual cells is essential to understand complex genetic diseases such as cancers. Latest high-throughput single-cell RNA- (scRNA-) and DNA-sequencing (scDNA-seq) technologies enabled cell-resolved investigation of pathological tissue clones. However, it is still technically challenging to simultaneously measure the genome and transcriptome content of a single cell. In this work, we developed CCNMF – a new computational tool utilizing the Coupled-Clone Non-negative Matrix Factorization technique to jointly infer clonal structures in single-cell genomics and transcriptomics data. We benchmarked CCNMF using both simulated and real cell mixture derived datasets and fully demonstrated its robustness and accuracy. We also applied CCNMF to the paired scRNA and scDNA data from a triple-negative breast cancer xenograft, resolved its underlying clonal structures, and identified differential genes between cell clusters. In summary, CCNMF presents a joint and coherent approach to resolve the clonal genome and transcriptome structures, which will facilitate a better understanding of the cellular and tissue changes associated with disease development. CCNMF is freely available at https://github.com/XQBai/CCNMF.

Introduction

Understanding how genome changes shape gene expression in individual cells is essential to understand complex genetic diseases such as cancers. In particular, characterizing clonal gene dosage effect, i.e. the sensitivity of each gene's expression to its clonal copy number state, is critical to elucidate the functional consequence of may disease associated copy number variations (CNVs). However, there is still no technology that efficiently and simultaneously measures copy number and expression profiles from the same cell. Recent progress in single-cell RNA sequencing (scRNA-seq) [1, 2, 3, 4] and single-cell DNA sequencing (scDNA-seq) [5, 6] enabled efficient cellular genomic and transcriptomic profiling of the same sample. Capitalizing on these advances, we propose a computational method CCNMF (Coupled-Clone Non-negative Matrix Factorization), that jointly clusters paired scRNA gene expression and scDNA copy number data from the same sample, faithfully recovers the underlying clonal structure, accurately identifies clonal identity for all single cells, and statistically infers gene-wise dosage and expression changes that differentiate the clones.

The accurate inference of clones underlying paired scDNA and scRNA data is central to study clonal dosage effect. Only a few methods working with both scDNA and scRNA data

*Corresponding Authors. Emails: lwan@amss.ac.cn (Lin Wan), l.c.xia@stanford.edu, xia.stanford@gmail.com (Li C. Xia).

were available, which mostly operate in a reference-mapping fashion, i.e., one type of data is mapped to the clonal structure defined by the other type of data [7, 8, 9, 10]. For examples, clonealign statistically assigns scRNA gene expression states to a phylogenetic tree representing scDNA-derived clones in a Bayesian approach [8]. Seurat [7], which mainly integrates multiple scRNA datasets, can project other types of single cell data to the scRNA-derived reference based on nearest neighbor search. In a related work, DENDRO [9] inferred single cell copy numbers from scRNA data and validated the result using the paired scDNA data. However, reference-based inference methods risk systematic bias because the anchor technology was arbitrary chosen without objective justification. These methods also did not exploit the coherence of underlying clonal structure between data types to maximize the use of available information.

Instead, we took a coherent joint inference approach to identify the underlying clonal structure and derive the clonal dosage effect. Our coupled-clone nonnegative matrix factorizations framework (CCNMF) followed the coupled clustering concept by Zhana et., al [11]. The framework is based on optimizing an objective function that simultaneously maximizes intra-technology clone compactness and inter-technology clone coherence for the paired scRNA- and scDNA-seq data. We validated CCNMF's performance using both simulated and real cell-line mixture data and CCNMF achieved high accuracy. In the simulation, we simulated gene dosage effects by empirically modeling the large bulk public RNA and DNA-seq datasets. These simulated clonal structures were successfully recovered by CCNMF given with non-informative prior input. We further developed statistical tests and visualization schemes to intuitively show these most significant clone-differentiating genes. We finally applied CCNMF to characterize a patient-derived triple negative breast cancer xenograph and showed CCNMF is capable of discovering clonal structure and dosage effect in real biological applications. We foresee paired scDNA and scRNA-seq analyses combined with CCNMF analysis offer a novel paradigm to study the functional consequence of clonal gene dosage changes and how it contributes to disease progression.

Materials and Methods

Coupled factorization of scDNA and scRNA data to identify clonal structure

Here, we derive the coupled-clone nonnegative matrix factorization framework we will use to identify the clonal structure underlying paired scDNA and scRNA data from the same tumor sample. Note that the input matrix $O \in R^{p \times n_1}$ is the copy numbers of p genes and n_1 cells from the scDNA data, and the input matrix $E \in R^{p \times n_2}$ is the gene expression levels of p genes and n_2 cells from the scRNA data. In general, these inputs matrices could be from any paired scDNA- and scRNA-seq datasets generated from the same sample but with unpaired cells (Figure 1). To couple up the nonnegative factorization of matrices O and E, we additionally define a matrix $A \in R^{p \times p}$ to represent the prior knowledge of gene-wise dosage effect linking expression to copy number. The matrix A can be estimated priorly either by a linear regression model using public paired RNA and DNA bulk sequencing data [12], or by using an uninformative prior as an identity matrix. Hence, we simultaneously cluster the datasets O and E by minimizing the following objective function $\mathcal{F}(W, H)$:

$$\min_{W,H \geq 0} \mathcal{F}(W,H) = \min_{W_1,H_1,W_2,H_2 \geq 0} \frac{1}{2} \|O - W_1 H_1\|_F^2 + \frac{\lambda_1}{2} \|E - W_2 H_2\|_F^2 - \lambda_2 tr(W_2^T A W_{1sd}) + \mu(\|W_1\|_F^2 + \|W_2\|_F^2) \quad (1)$$

where we denote $\{W_1 \geq 0, W_2 \geq 0\}$ and $\{H_1 \geq 0, H_2 \geq 0\}$ by their shorthands W and H.

Note that by minimizing the first two terms of the objective we ensured the respective decompositions of O and E by NMF, such that $O = W_1 H_1$ and $E = W_2 H_2$, where W_i $(i = 1, 2)$ is the mean matrix of clusters for n_i $(i = 1, 2)$ cells and H_i $(i = 1, 2)$ is the weight matrices that

softly assign n_i $(i = 1, 2)$ single cells to the underlying common clusters. Upon convergence, H_i provides one with the inferred cluster identities for all single cells. Also note that by minimizing the third term $-tr(W_2^T AW_1)$ we ensured the coherence of the inferred underlying clone structure between the scRNA and scDNA data. Finally, by minimizing the last term, we controlled the growth rates of W_1 and W_2 to avoid overfitting.

Optimization solution of CCNMF

Next, we applied the alternating direction methods of multipliers (ADMM) [13, 11] to find the gradients of Eq 2. We let Φ and Ψ be the matrices containing the Lagrangian multipliers for W and H, thus we had the transformed objective function as follows:

$$L(W, H, \Phi, \Psi) = \mathcal{F}(W, H) + \sum_{k=1}^{2} tr(\Phi_k W_k^T) + \sum_{k=1}^{2} tr(\Psi_k H_k^T), subject\ to: \Phi \geq 0, \Psi \geq 0, k = 1, 2 \tag{2}$$

We set all its first order derivatives to zeros and obtained the following equations:

$$\begin{aligned}
\frac{\partial L(W, H, \Phi, \Psi)}{\partial W_1} &= (W_1 H_1 H_1^T + 2\mu W_1) - (OH_1^T + \lambda_2 A^T W_1) + \Phi_1 = 0 \\
\frac{\partial L(W, H, \Phi, \Psi)}{\partial H_1} &= -W_1^T O + W_1^T W_1 H_1 + \Psi_1 = 0 \\
\frac{\partial L(W, H, \Phi, \Psi)}{\partial W_2} &= (\lambda_1 W_2 H_2^T H_2 + 2\mu W_2) - (\lambda_1 EH_2^T + \lambda_2 AW_1) + \Phi_2 = 0 \\
\frac{\partial L(W, H, \Phi, \Psi)}{\partial H_2} &= \lambda_1 (W_2^T W_2 H_2 - W_2^T E) + \Psi_2 = 0
\end{aligned}$$

Solving these equations, we obtained the values for Lagrangian multipliers, as follows:

$$\begin{aligned}
\Phi_1 &= (OH_1^T + \lambda_2 A^T W_1) - (W_1 H_1 H_1^T + 2\mu W_1) & \Psi_1 &= W_1^T O - W_1^T W_1 H_1 \\
\Phi_2 &= (\lambda_1 EH_2^T + \lambda_2 AW_1) - (\lambda_1 W_2 H_2^T H_2 + 2\mu W_2) & \Psi_2 &= \lambda_1 (W_2^T E - W_2^T W_2 H_2)
\end{aligned}$$

which, when plugged back into Eq (2), gave us the required gradients of the objective function.

Finally, we used the obtained gradients with a descent algorithm to iteratively update and optimize the objective function until convergence by the following steps:

$$\begin{aligned}
w_{ij}^1 &\leftarrow w_{ij}^1 \frac{(OH_1^T + \lambda_2 A^T W_2)_{ij}}{(W_1 H_1 H_1^T + 2\mu W_1)_{ij}} & h_{ij}^1 &\leftarrow h_{ij}^1 \frac{(W_1^T O)_{ij}}{(W_1^T W_1 H_1)_{ij}} \\
w_{ij}^2 &\leftarrow w_{ij}^2 \frac{(EH_2^T + \frac{\lambda_2}{\lambda_1} A^T W_1)_{ij}}{(W_2 H_2 H_2^T + \frac{2\mu}{\lambda_1} W_2)_{ij}} & h_{ij}^2 &\leftarrow h_{ij}^2 \frac{(W_2^T E)_{ij}}{(W_2^T W_2 H_2)_{ij}}
\end{aligned}$$

Parameter choices and estimation of the coupling matrix A

The model has three parameter inputs: λ_1, λ_2 and μ, to initialize the iterative computation. These parameters can be empirically determined from the input data. In practice, we used an automatic balancing strategy to determine the parameters, which ensured the initial values of the four terms of the objective function are within the same order. Our experiences were that: λ_1 and λ_2 are variable; while μ should be set equal to 1.

The coupling matrix A is also expected as input, for which one can supply an uninformative identify matrix, in which the only non-zero elements are the diagonal ones. To provide a more informative prior, we can estimate A from known associations between copy number and gene

expression using paired bulk sequencing data of the same type of tissue source. It was well known that DNA copy number is highly positively correlated with expression levels for most ($> 99\%$) of expressed human genes [14]. We thus calculated a diagonal coupling matrix ($A \in R^{m \times m}$) where each diagonal element is estimated by the ratio of gene-wise mean expression to mean copy number using the paired bulk RNA-seq and microarray data on The Cancer Genome Atlas (TCGA). This empirically obtained matrix A was used in simulation to generate realistic paired scRNA and scDNA datasets.

Alignment of scDNA copy number with scRNA gene expression

While the scRNA data is typically presented as a gene expression matrix by cell and gene, the scDNA data is typically presented as copy numbers by cell and genome segmental bins. To associate the genome segmental bins to the corresponding genes, we took these preprocessing steps: (1) We aligned both the scRNA and scDNA data to the same human genome assembly (we used hg19 [8]); (2) We found the genomic location overlaps between the gene annotation tracks and the genome segmental bins (R package IRanges [15]); (3) We retained only genes that had mapped to one unique genome segmental bin and excluded any ambiguous genes with multi-mappings because of spanning segment breakpoints. After that, each of the p remaining copy number segmental bins were one-to-one mapped to the remaining p genes. Finally, we extracted from raw input the data of these remaining genes and copy number segmental bins, which gave us a properly formed paired scRNA and scDNA data $E \in R^{p \times n_2}$ and $O \in R^{p \times n_1}$ for CCNMF analysis.

Simulating paired scDNA and scRNA datasets

We first evaluated CCNMF using simulated paired scDNA and RNA data following the procedure as illustrated in Figure S1. The simulation principle is to coherently generate scRNA and scDNA data from the same ground truth genetic copy number and clonality while also allowing adding sequencing platform specific noises. To simplify the simulation, we set the total number of clones to be $k = 3$ in all simulated scenarios. We always specified that the first clone (cluster) as normal cells with a genetic copy number profile vector $V_1 = [2, \cdots, 2] \in R^m$, where m enumerates over all genome segmental bins.

We specified the second cluster to represent clonal deletions. We obtained its associated genetic copy number vector $V_2 \in R^m$ by replacing fractional components of V_1 with the absolute copy number values randomly sampled from $\{0, 1\}$ according to parameters. Similarly, we specified the third cluster to represent clonal amplifications and obtained $V_3 \in R^m$ by replacing fractional components of V_1 with copy number randomly sampled from $\{3, 4\}$. We also recorded the ground truth clonal genetic copy numbers as G_i^{CN}.

Next, we defined the observed copy number per gene and cell as O_i^{CN}, which is the experimentally observed scDNA copy number data. We recognized that various batch, sequencing and platform noises can affect the genome segmentation results from experiments and cause O_i^{CN} to deviate from G_i^{CN}. To realistically simulate O_i^{CN}, we used a Markovian model, which we estimated the transition probability matrix $P(O^{CN}|G^{CN})$ from the bulk copy number data of the TCGA project. To simplify the computation, the dimension of $P(O^{CN}|G^{CN})$ was set to $C_{max} + 1$, such that the copy number states can range from 0 to C_{max} . In practice, we chose $C_{max} = 4$ as the maximum cut-off for copy number, which means any copy number larger than 4 (inclusive) were grouped into the state C_{max}.

Specifically, we estimated the transition matrix as follows: we downloaded the TCGA genetic copy number difference G_{diff}^{CN} data from cBioPortal [16, 17] with 171 triple-negative breast cancer basal samples on paired bulk RNA-seq and DNA-seq data. where $G_{diff}^{CN} = \{-2, -1, 0, 1, 2\}$

and 0 means diploid/normal. We transformed G^{CN}_{diff} to the integral copy number by $G^{CN} = G^{CN}_{diff} + 2$. We also downloaded the TCGA tumor purities for all samples from Butte et al.[18] and denoted them by $Purity = \{p_1, \cdots, p_n\}$. To estimate $P(O^{CN}|G^{CN})$: (1) We compensated for the associated purity and arrived at the raw copy number number $R^{CNV}_n = (2 \times C^{NR}_{diff})/p_n + 2$; (2) We grouped the R^{CNV} according to their underlying genetic copy number G^{CN} status (see Figure S2); (3) We fitted a Gaussian mixture model to each grouped R^{CNV} (see Figure S3); (4) We calculated the $P(O^{CN}|G^{CN})$ by non-parametric binning of the histogram from the fitted Gaussian mixture per G^{CN} status.

Note that $P(O^{CN}) = P(O^{CN}|G^{CN}) * P(G^{CN})$ is the empirically estimated multinomial probability vector we will use to simulate observed copy number O^{CN} given the underlying genetic G^{CN}. We therefore simulated per gene per cell scDNA data D_{ij} by randomly sampling from these multinomial distributions, such that $D_{ij} \sim multinomial(P(O^{CN}|G^{CN}) * P(V_i))$. As the last step, we added technology, batch and platform specific outliers and dropouts to the simulated scDNA data following the same procedure as for simulating the scRNA data that we described immediately below.

We simulated the scRNA data based on their associated clonal copy number profiles using the Splatter pipeline [19]. Specifically: (1) We simulated the i-th clonal gene expression background with multiplying the copy number profile V_i by the dosage effect [20], such that the gene-wise expression mean of the i-th cluster is $\lambda_i^{'} = \lambda_i * V_i$; (2) We proportionally adjusted the gene-wise means for each cell using every cell's library sizes (L_j) which can be fitted by a log normal function with the estimated parameters from real data (see details in [19]), where $\lambda_{ij}^{'} = L_j(\lambda_i^{'}/\sum(\lambda_i^{'})$; (3) We generated reads for each gene and each cell where their counts followed a Poisson mixture with an outlier component, such that $X_{ij}^{'} = 1^O_{ij}(X^O_{ij}) + (1 - 1^O_{ij})X_{ij}$, $X_{ij} \sim Poission(\lambda_{ij}^{'})$, $1^O_{ij} \sim Bernoulli(\pi^O)$, π^O is the probability of outlier occurrence, and X^O_{ij} is the outlier's expression; (4) We simulated cell-wise gene dropout events by randomly replacing fractions of the generated gene expression with zeros, such that $G_{ij} = 1_{ij}X_{ij}^{'}$ mimicking a dropout effect $1_{ij} \sim Bernoulli(1/(1 + \lambda_i^{'}))$.

Simulated and real paired scDNA and scRNA datasets

As our first benchmark, we simulated 46 paired scDNA- and scRNA-seq datasets, which we referred to as the **Sim** data. In the simulation, we used 3-cluster and linear/bifurcated clone structure scenarios. We varied common experimental parameters, such as the percentages of outliers, dropouts, and genes showing dosage effect. For each simulated dataset, we randomly generated cell-wise scDNA and scRNA data according to the specified scenario and parameters using the procedure as detailed previously (also see Figure S1). Each of the obtained dataset in *Sim*, has 1000 cells and 2000 genes/CNV bins, and the three composing clusters have 200, 400 and 400 cells each. The first cluster is always normal cells, and the second and third clusters representing deletion and amplification clones respectively. We set the percentages of differentially acquired deletions and amplifications to affect 10 % to 50 % chromosome regions. We deposited the *Sim* data into GitHub.

For the second benchmark, we downloaded a set of paired scDNA and scRNA real data from the public domain, which we referred to as the **OV** data. The *OV* data was composed of the DLP scDNA-seq and the 10X genomics scRNA-seq data generated from a mixture of high grade serous carcinoma (HGSC) cell lines. The mixture was made up of cells from ascites (OV2295R) and solid (TOV2295R) tumors. The scRNA subset had 1717 and 4918, and the scDNA subset had 371 and 394 (OV2295R and TOV2295R) cells, respectively.

To demonstrate the full utility of CCNMF, we also used another set of paired scRNA- and scDNA-seq data referred to as **SA501**. The data was generated from a triple-negative breast

cancer patient derived xenograft SA501X3F. It had 1430 cells of scRNA data and 260 cells of scDNA data, respectively and the underlying clonal structure was unknown [8]. The *SA501* scRNA data was generated using the 10X genomics chromium platform, which measured 32738 genes per cell. The *SA501* scDNA data was generated using the single-cell DLP DNA-seq which measured 20651 copy number segments.

Performance Evaluations

To evaluate the performance of CCNMF given the ground truth cell cluster labels, we used the Adjusted Rand Index (ARI) [21, 22]. The ARI measures the similarity between the labels assigned by any two clustering schemes as follows:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \tag{3}$$

where n_{ij}, a_i, b_j are values from the contingency table describing the overlapping label counts between the two clustering schemes. Here n_{ij} is the number of overlapping label counts between the cluster i of the first scheme and the cluster j of the second scheme. Note $a_i = \sum_j n_{ij}$, and $b_j = \sum_i n_{ij}$.

Results and Discussion

The CCNMF tool for identifying coherent scDNA and scRNA clone structure

Our CCNMF analytical framework was implemented into an *R* language package. The workflow of the tool was illustrated in Figure 1. As shown, the tool accepts standard scRNA- and scDNA-seq data formats, as provided by 10X genomics scRNA/scDNA and DLP scDNA platforms. It also accepts manually curated scDNA and scRNA data, as long as it follows the standard formats. The tool then executes the statistical framework and analytical steps as already detailed in the Methods section.

Briefly: (1) it aligns the scRNA gene expression and scDNA copy number bins to the provided reference genome. (2) it establishes the one-to-one correspondence between genes and bins using location overlapping ($>$ 1bp); (3) it initializes the coupled NMF between scRNA and scDNA data using with a non-informative prior, or optionally an user provided informative coupling matrix A; (4) it iteratively optimizes the objective using the CCNMF algorithm until convergence, thus simultaneously identifies the most coherent clonal structure and most probable cell clonality membership; and (5) it identifies genes exhibiting clonal differential expression and dosage effects by statistical testing using the inferred cell clonal identities.

The output of the tool includes the W matrices which represent the expression or copy number profile centroids of each RNA or DNA clusters, the H matrices that represent the cell-wise membership weights toward each cluster for all RNA and DNA cells, and a list of genes that demonstrated most significant expression change and shift in dosage sensitivity. The tool is operating system independent thus works with any R installation. It is publicly available from the Github (https://github.com/XQBai/CCNMF) with detailed readme, manual and example pages.

CCNMF recovers the underlying clonal structures in the *Sim* dataset

We performed systematic benchmark studies using the simulated paired scRNA and scDNA datasets. The performance of CCNMF as measured by ARI was presented in Tables S1-S3. These simulations included 6 different scenarios, which are $\{Bifurcate, Linear\} \times \{copy_number_fraction,$

$outlier_percentage, dropout_percentage\}$. In each scenario the performance of CCNMF was repeatedly assessed by varying the parameter of interest in its range while keeping the other parameters at default and unchanged. The default parameter values for copy number fraction, outlier percentage and dropout percentages were 50%, 0% and 0% respectively.

Table S1 showed the simulation results for varying the copy number fraction under the linear and bifurcate clonal structure scenarios. Copy number fraction, ranging from 0.1 to 0.5, was defined as the percentage of genome undergoing copy number changes, where 0.1 stands for 10% genome region had such changes while 0.5 stands for 50% of the genome. As we can see from the table, for all cases under both scenarios, CCNMF achieved high accuracy in recovering the underlying clonal structure. It was almost all 100% accurate, except for one case that it was 98%. It is expected that as the copy number fraction reduces, the clonal copy number difference becomes smaller and the scenario becomes harder for CCNMF to resolve. The results showed that such effect is very mild, as at only 10% of genome having copy number difference between the clones, CCNMF was still able to correctly resolve the underlying clone structure.

Table S2 showed the simulation results for the dropout percentage under the linear and bifurcate clonal structure scenarios. Dropout percentage, ranging from 0.1 to 0.9, was defined as the percentage of gene expression or copy number values that are zero cell-wise, either because of limited sensitivity of technology or because of it is non-presence/expressed. This type of noise are also very common in scRNA and scDNA experiments because of amplification bias and other random events. Dropout percentage at 0.1 means that 10% (0.9 for 90%) of all simulated cell-wise gene expression or copy numbers were perturbed to be zeros. As we can see from the table, for all cases under both scenarios, CCNMF achieved high accuracy in recovering the underlying clonal structure. It had $> 98\%$ accuracy, except for one case that it was 81%.

Table S3 showed the simulation results for the outlier percentage under the linear and bifurcate clonal structure scenarios. Outlier percentage, ranging from 0.1 to .9, was defined as the percentage of cells having an extreme copy number or expression value. These data points are typically deemed technical errors and are excluded from downstream analysis. Outlier percentage at 0.1 stands for 10% (while 0.9 for 90%) of all simulated scDNA and scRNA cells were perturbed to be outliers. As we can see from the table, when the outlier percentage is $<$ 60%, CCNMF achieved high accuracy (all ARI $> 92\%$ with the majority of them $> 95\%$) under both scenarios. Putting all together, our comprehensive simulation study had demonstrated the high performance of CCNMF in coherently resolving the underlying clonal structures across the paired scDNA and scRNA data.

CCNMF identified the underlying co-clusters in the OV dataset

We next performed an additional benchmark using the a real paired scRNA and scDNA dataset. The dataset was composed a cell mixture involving OV-2295(R), an ascites site cell line (abnormal build up but non-cancerous adjacent tissue) and TOV-2295(R), a high-grade serous ovarian cancer cell line from the same patient [23]. The CCNMF result and visualization was presented in Figure 2. For this dataset, the most significant separation of the cells are their cell line identities. We had collected these true identities from the original publication and regarded them as ground truth. By comparing CCNMF identified clusters to the ground truth, we found the ARI is 1, which means CCNMF completely recovered this underlying mixture structure.

The good result was also self-evident by looking at Figure 2, in which cell clusters in all four sub figures are same color coded by cluster to ease the interpretation. In Figures 2A and 2B, we showed the heatmap clusters for the scDNA and scRNA data for CCNMF identified differentially expressed or copy number varied genes respectively. In Figures 2C and 2D we showed their respective tSNE [24] plots overlaid with CCNMF cluster identity. In all these plots, two cell

clusters were clearly separated and their cluster membership correctly assigned. In particular, we observed consistent cluster-wise gene dosage effect, whereas the genes showing copy number gains in Cluster 2 (C2 or the TOV-2295(R) cells) in the scDNA data (e.g. MAP3K13, AMECK1, MGST1, TCEB3 and others) all had higher expression levels in scRNA data and vice versa. The only exception may be GUCY1B3 for which the effect was not the same direction.

These plotted genes were selected as the top 10 ranked by their difference significance after we applied t-test to identify the most differentially expressed or copy number changed genes between the clusters for both the scDNA and the scRNA data. Among them, the most obvious signature gene differentiated the two cell lines was MAP3K13, which was significantly amplified in the tumoral scDNA Cluster 2 cells, thus showed highly elevated expression level in the corresponding scRNA Cluster 2 cells. We further validated this finding by literature search [25, 26, 27]. We learned that MAP3K13 was a well-known biomarker gene for ovarian cancer, which was identified as a positive regulator of the Myc gene to promote tumor development. Therefore MAP3K13 amplification and resulted high expression level in the ovarian tissue is an important determinant that differentiates adjacent abnormal tissue from further cancerous development.

Clonal differential expression and dosage sensitivity in the *SA501* dataset

As real a world application, we performed a complete analysis of the paired scRNA and scDNA *SA501* dataset. The dataset was composed of xenograft SA501X3F cells derived from a triple-negative breast cancer patient. The CCNMF output and visualization was presented in Figure 3. This dataset were previously analyzed by Campbell, J. N. et al. [8], Zahn, H. et al. [5] and Eirew, P. et al. [28], however there were no consensus on cell cluster identities. Since we have thoroughly validated the CCNMF's performance with both simulated and real cell line mixture data, here, we will elucidate *SA501* clonal structure based on CCNMF's findings.

Using CCNMF, we first obtained a three cluster clonal structure as depicted in Figure 3 for *SA501*. All the four subfigures were same color coded for the same CCNMF identified cluster (green for C1/Cluster-1, red for C2/Cluster-2 and blue for C3/Cluster-3). In Figures 3A and 3B, we showed the heatmap clusters for the scDNA and scRNA data for identified differentially expressed or copy number varied genes. These plotted genes were selected as the top 7 ranked by their difference significance after we applied t-test to identify the most differentially expressed or copy number changed genes pairwise between all scDNA and scRNA clusters. The number 7 was chosen only for the succinctness our discussion, while a complete ranked list was computed by CCNMF. We finally arrived at total 17 selected genes after consolidating redundant ones. In Figures 3C and 3D we showed the three clusters' tSNE plots overlaid with identity.

As we can see, the three cluster (clone) structure was evident in the scRNA data. In particular, the scRNA profile of Cluster-1 was clearly defined by the high expression levels of the first 7 genes to the left and the low levels of the 6 genes in the middle. The scRNA profile of Cluster-2 was defined by the slightly higher expression levels of the last 4 genes to the right and the low expression levels of all the other genes. Finally, the scRNA profile of Cluster-3 was defined by the high expression levels of the middle 6 genes and the low expression levels of all the other genes.

The clonal structure was however less obvious in the scDNA data, where the clusters C1 and C3 seemed to be very similar to each other. This trend is clear for both heatmaps, where C1 and C3 had very similar copy number profiles for their top ranked differentially expressed genes. It is also observed in the tSNE plots that the clusters C1 and C3 were much closer together as compared to C2. These results suggested that C2 diverges earlier from C1 and C3 and the later divergence of C1 and C3 could have been dominated by gene regulation changes. In

situations like this, where the scDNA data showing a degenerated clonal structure, tools relying on mapping the scRNA data onto the clonal structure derived from the scDNA data would only identify two expression clusters, thus mis-represented the underlying structure. CCNMF completely avoids such complicity thanks to its co-clustering design.

The most evident clonal signature gene differentiates C2 and C1+C3 is TM4SF1, which was significantly amplified in the C2 cells only. However, its expression was not significantly affected by the amplification. Rather it remained low for both C1 and C2 cells, while its expression level is slightly higher in C3. In fact, this observed clone specific dosage sensitivity suggested that gene expression levels can be mitigated or compensated via transcriptional regulations even if the gene dosage was not the same. By literature search, we found TM4SF1 is a trans-factor that regulates cell migration and apoptosis. It was reported [29] to contribute to the development and metastasis of advanced breast cancer.

Another evident clonal signature gene we identified for C2 is TNFR5F12A, which was significantly deleted in the C2 cells only. Its expression, however, was also not significantly affected by this deletion. Rather it remained low for both C1 and C2 cells, while the expression level is slightly higher in C3. By literature search, we found TNFR5F12A is a trans-factor that also regulates cell migration and apoptosis. It was also reported [30] to contribute to the development and metastasis of advanced breast cancer. Therefore, the CCNMF analysis of the *SA501* identified clonal-defining genes such as TM4SF1 and TNFR5F12A that may have contributed to its development into metastatic cells.

Conclusions

Tumoral and other pathological tissues often demonstrate heterogeneous and clonal nature, which could carry important gene dosage effect information for elucidating the genetic cause and etiology of these diseases. To facilitate the understanding of the tissue clonal structure and the associated gene dosage effect, we proposed a coherent and elegant non-negative matrix factorization based co-clustering approach CCNMF. CCNMF operates to optimize an objective function that simultaneously maximizes for intra-technology clonal compactness, inter-technology clonal coherence and expected dosage effect consistence, using paired scRNA- and scDNA-seq data. We developed, implemented and validated the CCNMF tool with both simulated and real cell line mixture datasets achieving high accuracy. We further demonstrated the utility of CCNMF by identifying the underlying clonal structure and clonal differential expression in a patient derived breast cancer sample. We expect CCNMF to serve as a much needed bioinformatics tool for performing single cell level clonal dosage effect analysis for the community.

References

[1] Tang, F. *et al.* mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009).

[2] Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049 (2017).

[3] Macosko, E. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

[4] Campbell, J. N. *et al.* A molecular census of arcuate hypothalamus and median eminence cell types. *Nature Neuroscience* **20**, 484–496 (2017).

[5] Zahn, H. *et al.* Scalable whole-genome single-cell library preparation without preamplification. *Nature Methods* **14**, 167–173 (2017).

[6] Andor, N. *et al.* Joint single cell dna-seq and rna-seq of gastric cancer reveals subclonal signatures of genomic instability and gene expression. *bioRxiv* (2018). `https://www.biorxiv.org/content/early/2018/10/17/445932.1.full.pdf`.

[7] Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).

[8] Campbell, K. R. *et al.* clonealign: statistical integration of independent single-cell rna and dna sequencing data from human cancers. *Genome Biology* **20**, 54 (2019).

[9] Zhou, Z., Xu, B., Minn, A. & Zhang, N. R. Dendro: genetic heterogeneity profiling and subclone detection by single-cell rna sequencing. *Genome Biology* **21**, 10 (2020).

[10] Welch, J. D. *et al.* Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887.e17 (2019).

[11] Duren, Z. *et al.* Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proceedings of the National Academy of Sciences* **115**, 7723–7728 (2018).

[12] Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W. H. Modeling gene regulation from paired expression and chromatin accessibility data. *Proceedings of the National Academy of Sciences* **114**, E4914–E4923 (2017).

[13] Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* **3**, 1–122 (2010).

[14] Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature Genetics* **47**, 115 (2015).

[15] Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Computational Biology* **9** (2013).

[16] Correction: The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* **2**, 960–960 (2012). URL `https://cancerdiscovery.aacrjournals.org/content/2/10/960`. `https://cancerdiscovery.aacrjournals.org/content/2/10/960.full.pdf`.

[17] Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Science Signaling* **6**, pl1–pl1 (2013). URL `https://stke.sciencemag.org/content/6/269/pl1`. `https://stke.sciencemag.org/content/6/269/pl1.full.pdf`.

[18] Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nature Communications* **6** (2015).

[19] Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell rna sequencing data. *Genome Biology* **18**, 174 (2017).

[20] Parris, T. Z. *et al.* Clinical implications of gene dosage and gene expression patterns in diploid breast carcinoma. *Clinical Cancer Research* **16**, 3860–3874 (2010).

[21] Rand, W. M. Objective criteria for the evaluation of clustering methods. *Publications of the American Statistical Association* **66**, 846–850 (1971).

[22] Kiselev, V. Y. *et al.* Sc3: consensus clustering of single-cell rna-seq data. *Nature Methods* **14**, 483–486 (2017).

[23] Létourneau, I. J. *et al.* Derivation and characterization of matched cell lines from primary and recurrent serous ovarian cancer. *BMC Cancer* **12**, 379 (2012).

[24] van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

[25] Zhang, Q. *et al.* The map3k13-trim25-fbxw7a axis affects c-myc protein stability and tumor development. *Cell Death and Differentiation* **27**, 420–433 (2020).

[26] Nam, S. W. *et al.* Identification of large-scale characteristic genes of müllerian inhibiting substance in human ovarian cancer cells. *International Journal of Molecular Medicine* **23**, 589–596 (2009).

[27] Snijders, A. M. *et al.* Genome-wide-array-based comparative genomic hybridization reveals genetic homogeneity and frequent copy number increases encompassing ccne1 in fallopian tube carcinoma. *Oncogene* **22**, 4281–4286 (2003).

[28] Eirew, P. *et al.* Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* **518**, 422–426 (2015).

[29] Yonghong, S., Yahong, X., Jie, X., Dan, L. & Jianyu, W. Role of tm4sf1 in regulating breast cancer cell migration and apoptosis through pi3k/akt/mtor pathway. *International journal of clinical and experimental pathology* **8**, 9081–9088 (2015).

[30] Jungho, Y. *et al.* High tnfrsf12a level associated with mmp-9 over expression is linked to poor prognosis in breast cancer: Gene set enrichment analysis and validation in large-scale cohorts. *PLoS One* **13(8)**, e0202113 (2018).

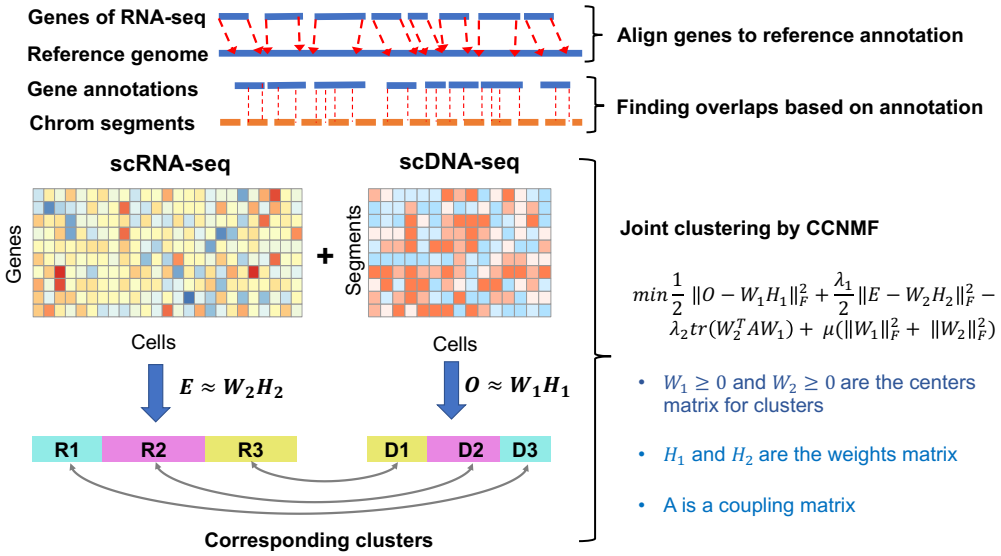

Figure 1: The analytical work of coupled-clone nonnegative matrix factorization (CCNMF) for inferring shared clonal structure in paired scDNA- and scRNA-seq data

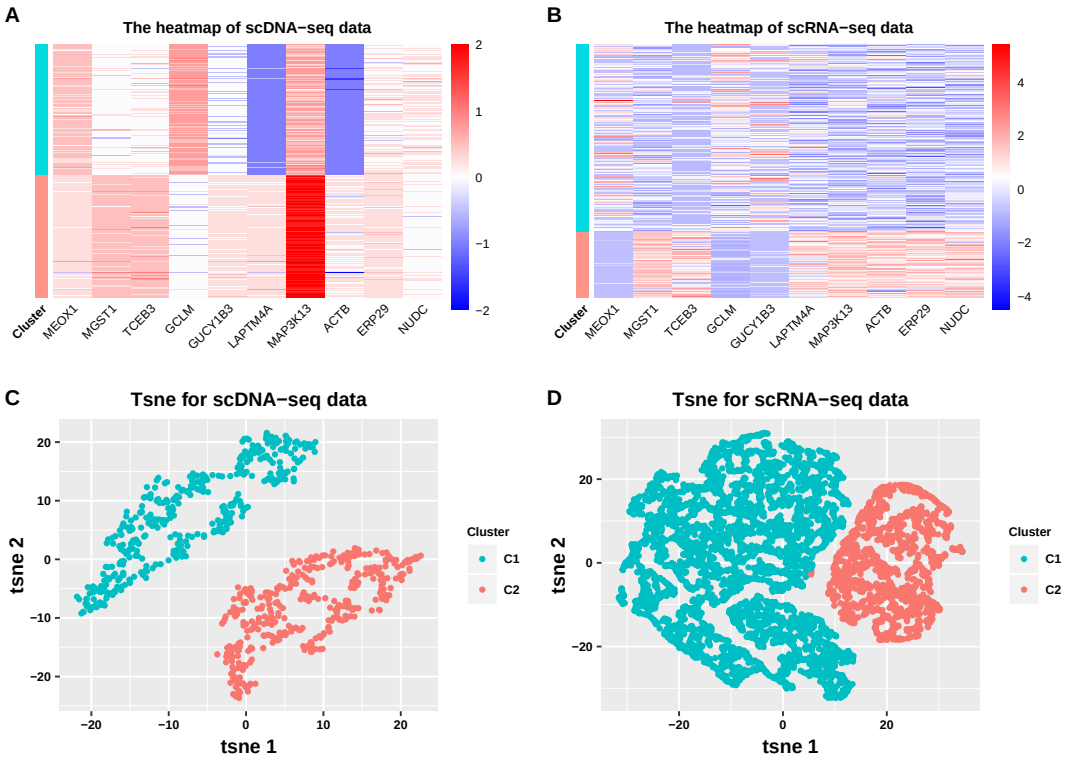

Figure 2: The scDNA and scRNA clonal structure of the *OV* dataset, including heatmaps of (A) copy number ratios and (B) gene expression of clonal signature genes and tSNE plots of (C) scDNA and (D) scRNA clones.

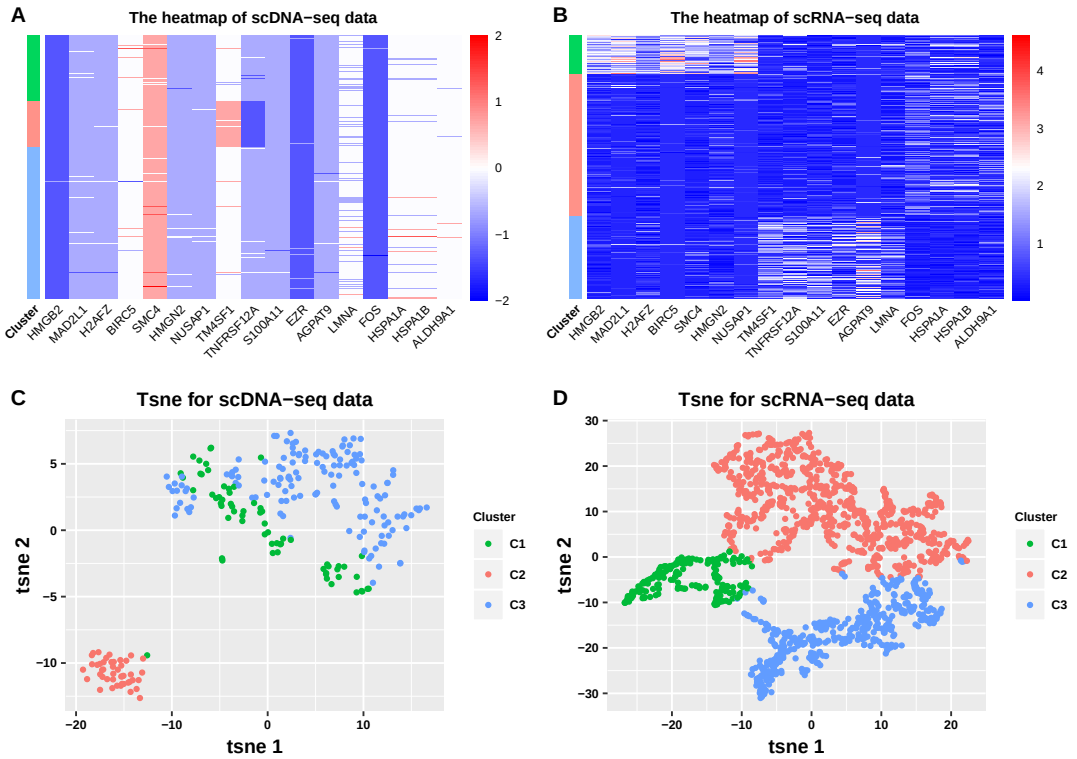

Figure 3: The scDNA and scRNA clonal structure of the *SA501* dataset, including heatmaps of (A) copy number ratios and (B) gene expression of clonal signature genes and tSNE plots of (C) scDNA and (D) scRNA clones.