Joint Inference of Clonal Structure using Single-cell Genome and Transcriptome Sequencing Data

Xiangqi Bai[1,2], Zhana Duren[3], Lin Wan [*1,2] and Li C. Xia [*4,5]

[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]Center for Human Genetics and Department of Genetics and Biochemistry, Clemson University, Greenwood, SC 29646, USA

[4]Division of Biostatistics, Department of Epidemiology and Public Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA

[5]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA

Abstract

Latest developments in high-throughput single-cell genome (scDNA-) and transcriptome sequencing (scRNA-seq) technologies enabled cell-resolved investigation of tissue clones. However, it remains challenging to cluster single cells of the same tissue origin across scRNA- and scDNA-seq platforms. In this work, we present a computational framework – CCNMF, which uses a novel Coupled-Clone Non-negative Matrix Factorization technique to jointly infer clonal structure for paired scDNA- and scRNA-seq data of the same specimen. CCNMF clusters single cells through statistically modeling their shared clonal structure and coupling copy number and gene expression profiles by their global correlation. We validated CCNMF using both simulated and real cell mixture benchmarks and fully demonstrated its robustness and accuracy. As real world applications of CCNMF, we analyzed data from a gastric cancer cell line, an ovarian cancer cell mixture, and a triple-negative breast cancer xenograft. We resolved the underlying clonal structures and identified dosage-sensitive genes between co-existing clones. In summary, CCNMF is a coherent computational framework that simultaneously resolves genome and transcriptome clonal structures, facilitating understanding of how cellular gene expression changes along with clonal genome alternations.

Availability: The R package of CCNMF is available at https://github.com/XQBai/CCNMF.

Keywords: Single-cell Genomics, Single-cell Multi-omics Integration, Copy number variants, Nonnegative matrix factorization, Gene dosage effect.

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

Introduction

The understanding of how genomic content changes impact gene expression in individual cells is essential to further understand the clonal development of normal and diseased tissues. In particular, characterizing the clonal gene dosage effect, i.e., the sensitivity of cellular gene expression to the copy number variants (CNVs) within a cell clone, is critical to elucidate the functional consequence of CNVs – a significant challenge in current structural variant research.

To date there is no efficient technology that can effectively and simultaneously measure copy number and expression profiles of the same cell. The closest attempt is scTrio-seq (Hou et al., 2016), which, in a low throughput manner, can measure genomic and transcriptomic contents in up to a few cells per batch. State-of-art high-throughput single-cell sequencing technologies can only measure either the transcriptome (Tang et al., 2009; Macosko et al., 2015; Zheng et al., 2017; Campbell et al., 2017) or the genome content of individual cells but not both (Zahn et al., 2017; Andor et al., 2020; Efremova and Teichmann, 2020).

For examples, popular droplet-based single-cell RNA sequencing (scRNA-seq) is routinely employed to measure cellular expression so as to assess the clonal development states of tissues and cell systems (Andor et al., 2020; McCarthy et al., 2020). Recently introduced droplet-based single-cell DNA sequencing (scDNA-seq) enabled cellular genome-wide measurement of genomic alternations, such as copy number variants, in thousands of cells (Andor et al., 2020; Velazquez-Villarreal et al., 2020). High quality single-cell copy number variants combined with single-cell gene expression profiles promise to further reveal the clonal heterogeneity in complex cell systems (Wang et al., 2014). Realizing the potential, however, still requires highly accurate co-clustering of single cells from the same specimen as represented by their heterogeneous scRNA- and scDNA-seq data.

Addressing the challenge, we developed an efficient computational method termed CCNMF (Coupled-Clone Non-negative Matrix Factorization), that can cluster and couple single cells across scRNA- and scDNA-seq datasets. CCNMF leverages known facts of paired samples, such as their underlying shared clonal structure and the high correlation between gene expression level and copy number state (Campbell et al., 2019; Andor et al., 2020). It models these constraints using an objective function within the coupled NMF framework and employs unsupervised machine learning algorithms to infer the most likely co-clustering solution. The co-clustering concept was first introduced in Duren et al. (2018). In objective optimization, CCNMF simultaneously maximizes intra-technology clone compactness and inter-technology clonal structure coherence. The most coherent underlying clonal structure, i.e. the identity-linked cell clusters between scDNA and scRNA data, is thus inferred as the factorization weight matrix that optimally assigns all cells to their most probable cluster identity. Based on that, CCNMF can accurately estimate gene-wise dosage sensitivity per cell cluster.

Before CCNMF, only a few methods were available for analyzing the combined scDNA- and scRNA-seq data. These methods mostly operate in a map-to-reference mechanism, i.e., data from one technology is mapped to a reference clonal structure derived from another technology (Campbell et al., 2019; Stuart et al., 2019; Zhou et al., 2020). For examples, *clonealign*, an early

attempt of integrative model, statistically assigns scRNA gene expression states to a phylogenetic tree derived by scDNA clones (Campbell et al., 2019); *Seurat*, which mainly integrates multiple scRNA datasets, can also project other single-cell technology data to the scRNA-derived clusters using mutual nearest neighbor search (Stuart et al., 2019); *DENDRO* infers single cell copy numbers from scRNA data and validates the result using the paired scDNA data (Zhou et al., 2020). However, these map-to-reference inference methods risk systematic bias because the choice of reference technology is largely arbitrary, and different choices significantly influence the analysis results. The reference-free approach termed single-cell integrative nonnegative matrix factorization (iNMF) (Welch et al., 2019) assumed tight coupling of the single cell data being integrated. This assumption of data homogeneity is easily violated by the sparse and noisy paired scDNA- and scRNA-seq datasets.

Instead, CCNMF takes a data-driven and loose coupling approach, unbiased toward data and technology sources. CCNMF utilizes the internal coherence of the underlying shared clonal structure and the global concordance between gene expression and copy number in a specimen, to maximize the inference for true clone identity and clonal gene expression effect. We validated that CCNMF can faithfully resolve the underlying clonal structure, accurately recover clone identity for all single cells, and systematically detect gene-wise dosage between clones for both simulated and real cell mixture data. We further developed an integrative visualization scheme to illustrate the most significant clone-differentiating genes. We applied CCNMF to characterize a gastric cancer cell line, an ovarian cancer cell mixture and a patient-derived triple negative breast cancer xenograft. The results showed CCNMF is capable of identifying clonal structure and dosage effect in routinely studied cell systems. Therefore, paired scDNA- and scRNA-seq data combined with CCNMF analysis offers a new way to study the functional consequence of gene dosage change and how it contributes to clonal development.

Results

The CCNMF Toolkit

The CCNMF analytical framework was implemented into an *R* language package/toolkit. The toolkit's workflow was illustrated in Figure 1. As shown, the toolkit accepts paired scRNA- and scDNA-seq datasets in standard formats, such as those provided by 10X Genomics scRNA-/scDNA-seq and DLP scDNA-seq platforms. It also accepts manually created paired scRNA- and scDNA-seq datasets, as long as they follow the same formatting standards. The toolkit then executes the statistical framework and analytical steps as follows:

(1) It prepares the scRNA-seq gene expression data and scDNA-seq copy number segments based on a user provided genome reference; (2) it maps raw scDNA segment bins to genes using genomic location overlapping ($>$ 1bp) – converting them into gene-based coordinates the same as scRNA-seq data; (3) it initializes the coupled term between scRNA-seq (E) and scDNA-seq data (O) using a coupling matrix (A). A is either an identity matrix or a matrix providing with a prior information; (4) it iteratively optimizes the objective function using the gradient descent algorithm until convergence; (5) it identifies the most coherent clonal structure by finding the

maximum weights of the H matrices that represent the most likely cell clonality membership; (6) it identifies clone-differentiating genes and dosage effects by statistically testing gene expression level difference between the inferred cell clones.

The output of the toolkit includes the W matrices that represent the expression or copy number profile centroids of each scRNA or scDNA clusters, the H matrices that represent the cell-wise membership weights toward each cluster for all scRNA and scDNA cells, a list of genes that demonstrated the most significant clonal expression change or dosage sensitivity, and an integrated tSNE and heatmap plot intuitively visualizing these clonal dosage effects.. The toolkit is operating system independent and works with any R installation. It was open source on Github with detailed readme, manual and examples.

CCNMF resolves clonal structures in the *Sim* data

We first validated CCNMF's performance using a simulated *in silico* paired scDNA- and scRNA-seq datasets – *Sim* (see **Methods** for details). *Sim* included 3 coexisting clones in 2 clonal structure scenarios: $\{Bifurcate, Linear\}$, each with 3 parameters: $\{copy_number_fraction, dropout_percentage, outlier_percentage\}$. $Bifurcate$ is the scenario that a parent clone has two child clones, forming a parallel divergence, while *Linear* corresponds to the case with one grand-parent, one parent and one child clones, forming a sequential divergence. These are the only two possible clonal development scenarios given 3 related clones.

$copy_number_fraction$, ranging from 0.1 to 0.5, is the percentage of genome carrying copy number changes, where 0.1 stands for 10% of the genome while 0.5 stands for 50%. $dropout_percentage$, ranging from 0.1 to 0.9, is the percent of cells having no gene expression or copy number values, which represents the combined effect of technology's sensitivity limitation and that of the fact genes are not present or not expressed. $outlier_percentage$, ranging from 0.1 to 0.9, is the percent of cells showing an extreme copy number or expression value, typically because of technical errors.

In each of the scenarios, Adjusted Rand Index (ARI) based accuracy was repeatedly assessed by varying one parameter over its range at a time while keeping the others at default. The default parameter values for copy number fraction, outlier percentage and dropout percentages were 50%, 0% and 0%, respectively. The obtained CCNMF's ARI accuracy was summarized in **Supplementary Tables S1-S3**.

As we can see from **Supplementary Table S1**, for all cases under both scenarios, CCNMF achieved high accuracy in correctly resolving the underlying clonal structure given the varying copy number fractions. It all 100% accurate, except in one case that the accuracy was 98%. It is expected that as the copy number fraction decreases, the clonal copy number difference becomes smaller and the scenario becomes harder for CCNMF to correctly resolve. The results showed that such effect is only modest, as at 10% of genome having clonal copy number difference, CCNMF was still able to correctly resolve the underlying structure, showing its wide applicability.

Similarly, CCNMF achieved high accuracy in recovering the underlying clonal structure given the varying dropout percentages (**Supplementary Table S2**) in all cases under both

scenarios. It had > 98% accuracy except for one case it was 81%. High rates of dropouts are commonly observed in scRNA- and scDNA-seq experiments because of technology limitation, amplification bias and other noises. The exercise demonstrated CCNMF's robust and accurate inference toward highly sparse datasets.

Finally, CCNMF achieved high accuracy given varying outlier percentages (**Supplementary Table S3**). For both scenarios, CCNMF showed ARI > 92% in all tested cases and > 95% when the outlier percentage is < 60% of the input data. Technical errors, as represented by outlier percentage, rarely exceed 60%, thus this exercise proved that CCNMF is robust to technical noises. Putting all together, our comprehensive simulation study demonstrated the good performance of CCNMF in resolving the underlying clonal structures for paired scDNA- and scRNA-seq data in a wide range of technical error, noise and applicability considerations.

CCNMF identifies the cell origins of a cell line mixture

We next analyzed a real cell mixture benchmark dataset – *OV*. The dataset was composed of 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 (Letourneau et al., 2012). The most significant clonality of the mixture is the cells' true cell line identities. We collected these true identities from the dataset's original publication and treated them as the ground truth. We compared CCNMF identified cell identities to the ground truth and found the ARI was 100%, which means CCNMF completely and correctly resolved the underlying clonal structure of *OV*.

The consistent result was also self-evident from Figure 2. In Figures 2A and 2B, we showed their respective tSNE (van der Maaten and Hinton, 2008) plots overlaid with CCNMF cluster identities. In the plots, the two cell lines were clearly separated and all cell identities were correctly assigned. In Figures 2C and 2D, we showed the heatmaps for the scDNA- and scRNA-seq data of the top 14 CCNMF-identified signature genes for differential expression or copy number, when we applied t-test between the clusters.

Interestingly, we observed consistent clone specific gene dosage effects, that the genes had copy number gains in scDNA C2 (TOV-2295(R) cells) (e.g. RPS2, FTH1, CLDN3 and others) all had higher expression levels in the scRNA C2 and vice versa. The only exception may be KRT7 and ID1 for which the effect was opposite.

We identified the most prominent gene differentiating the two cell lines as RPS2, which was significantly amplified in the scDNA cluster C2, and showed highly elevated expression level in the scRNA C2. RPS2 is a well-known biomarker gene for ovarian cancer, down-regulating the let-7 family expression to promote tumor growth (Wang et al., 2011). Therefore RPS2 amplification and the resulted high expression level in the ovarian tissue could be driving determinant that differentiates cancerous cells from adjacent abnormal cells.

CCNMF characterizes clonal dosage sensitivity in a breast cancer xenograft

As a real world application, we analyzed the paired scRNA- and scDNA-seq dataset of a patient xenograft *SA501*. This dataset was composed of xenograft SA501X3F cells derived from a triple-negative breast cancer patient. It was previously analyzed by Eirew et al. (2015), Zahn et al. (2018) and Campbell et al. (2019), however, there is no consensus on ground truth of cell clonal identity. Since we have thoroughly validated CCNMF, We interpreted *SA501* based on CCNMF's findings.

We inferred a three-cluster clonal structure for SA501 as illustrated in Figure 3. We applied t-test to identify the most significantly differential expression or copy number change genes between the clusters. In Figures 3A and 3B, the three clusters' tSNE plots were overlaid with the inferred cell identities. Heatmaps of the top 17 signature genes were shown in Figures 3C and 3D. These genes were selected as a joined set of the top 10 differential genes of the scDNA- and scRNA-seq data.

As we can see, the three-cluster (clone) structure fitted well with the scRNA-seq data. The scRNA profile of C1 was clearly defined by the high expression level of the first 7 genes to the left and the low level of the 6 genes in the middle. The scRNA profile of C2 was defined by the slightly higher expression level of the last 4 genes to the right and the low expression level of all the other genes. Finally, the scRNA profile of C3 was defined by the high expression level of the middle 6 genes and the low expression level of all the other genes.

The structure was however less evident in the scDNA-seq data, where the clusters C1 and C3 seemed indistinguishable from each other. Note that C1 and C3 clusters also had similar copy number profiles for the top signature genes. And the tSNE plots for clusters C1 and C3 were closer as compared to C2. These results suggested that C2 diverged earlier from C1 + C3 and the later divergence of C1 and C3 could have been driven by transcription changes. In scenarios like this, when the scDNA-seq data can only represent a degenerated clonal structure, analytical tools mapping scRNA-seq data onto the clonal structure defined by the scDNA-seq data will only identify two clusters, thus mis-represent the underlying structure. CCNMF avoids such difficulty by its unbiased design.

The most significant signature genes differentiating C2 from C1+C3 are TM4SF1 and TNFR5F12A. As we can see, TM4SF1 was amplified while TNFR5F12A was deleted, both in scDNA C2 only. However, their expression levels were not always affected by their copy number changes. Rather they had low expression levels in both C1 and C2, while a higher level in C3. This is an interesting example of real world clone-specific gene expression dosage sensitivity. It means TM4SF1's copy gain did not increase its expression in C2. The normal copy TNFR5F12A still expressed in C1, as low as that of C2, in which it had lost copies. It suggested that there are dosage insensitive genes for specific clones, that the genes' clonal expression levels were not primarily determined by dosage, but transcriptional or other regulatory mechanisms. While these genes expression levels could still be determined by dosage in the other clones of the same specimen. TM4SF1 and TNFR5F12A were both trans-factors that regulate cell migration and apoptosis. They were consistently associated with tumor development and metastasis (Sun et al., 2015;

Yang et al., 2018).

CCNMF cleans up noisy data and identifies subclones in a heterogeneous cell line

Cells from the same cell line were typically presumed to have the same or at least similar genomic profiles. However this might not be true as it comes to cancer cell lines. To understand such unexpected clonality, we analyzed the *NCI-N87* gastric cell line. The datasets were large-scale and composed of 1005 scDNA-seq cells and 3246 scRNA-seq cells.

Although in high-throughput, it is associated with a new analytical difficulty that a substantial proportion of sequenced cells are likely replicating cells in S phase (Andor et al., 2020). The ongoing DNA replication in these cells drastically alters copy number segments and disguises true copy number variants. To filter out such cells before CCNMF analysis, we made use of the filtering signal given by the chromosome arm-level copy number variance. A replicating cell has higher such variance because the replication activity gives rise to scattered copy number doubling segments easily observable at the arm level (see Figure 4). We used an Expectation-Maximization algorithm to fit a 2-component normal mixture to the obtained variance distribution of intra-arm copy numbers for all scDNA cells (**Supplementary Figure S4**). We then filtered out the noisy replicating cells which were assigned to the larger mean component by their posterior probabilities. Using this effective and efficient filtering procedure, we successfully identified and removed all replicating cells, whose copy numbers are highly fluctuating. As shown in Figure 4, 281 cells (C0 group) in scDNA-seq data were identified as noisy/replicating cells and filtered out.

For the remaining cells, we performed one round of feature selection to reduce the data dimension and retain only relevant genes. To achieve that, we developed a statistical approach to select the most informative copy number segments. That is: we fitted a mixture normal distribution to the inter-cell variances of all copy number segments, as shown in **Supplementary Figure S5**; we then iteratively selected the most informative segments with the highest variance based on the distribution until at least 20% of the total segments were included. The genes residing in the selected informative copy number segments were then used as the scDNA-seq data input to CCNMF.

We applied CCNMF on the filtered and selected scDNA-seq data combined with the paired scRNA-seq data. By CCNMF, the remaining 724 scDNA cells were co-clustered into two subclones (C1 and C2). They are visually distinguishable by the arm-level amplification events on chromosome 3 and chromosome 19, while the 456 cells in C1 consistently had the 19q amplification and the 268 cells in C2 had the 3q amplification. There is a small proportion of cells with chromosome 11q amplification which were also assigned to C2. This clonal structure was independently validated by the analysis in Andor et al. (2020) though they took a more complicated and supervised map-to-refrence approach. These two clusters were also clearly separated in the scDNA and scRNA tSNE plots (Figures 5A and 5B).

We applied t-test to identify the most significantly differential expression or copy number change genes between the two clusters and the top 10 of them were illustrated in heatmaps as

Figures 5C and 5D. We assessed the dosage sensitivity of the top signature genes differentiating the two subclones. Two of these genes (TOP2A and MAF1) were consistently amplified in both C1 and C2. Yet they showed clone specific dosage sensitivity as their expression levels were significantly higher in C1 but stayed low in C2. TOP2A is a DNA topoisomerase gene, which controls the topologic states of DNA during replication, while MAF1 is an effector gene that represses RNA polymerase III. Both TOP2A and MAF1 were found to be frequently amplified biomarkers for increased aggressiveness in cancers (Nielsen et al., 2010; Hokonohara et al., 2019). On the other hand, UBE2S, a ubiquitin-conjugating enzyme gene, was uniquely amplified and highly expressed in C1. High expression of UBE2S was previously identified as a cancer progression marker in breast cancer (Ayesha et al., 2016; Liu et al., 2018). All other genes were relatively highly expressed in C1 but stayed low in C2 regardless of their clonal genomic copies.

These findings suggested that the two coexisting subclones of NCI-N87 may be different in their cancerous potential, while C1 is likely to be a more aggressive form as compared to C2. The high expression levels of TOP2A, MAF1 and UBE2S all suggested a high rate of replication and protein recycling activities in C1, which may explain for its higher aggressiveness. This high level of expression seen in C1 could be a direct result of increased gene dosage of these genes. However, this dosage effect is not necessary universal, as it could be reduced by transcriptional or other regulatory mechanisms, as we can see in C2, when the amplifications of TOP2A and MAF1 did not lead to their increased expression.

Although unexpected to many, given their high mutation burden, cancer cell lines could relatively easily split into heterogenic subclones if such combination is favored by the selection force such as lab culturing, as have already been reported elsewhere (Shen et al., 2020; Hynds et al., 2018). Several studies also already demonstrated the heterogenic nature of NCI-N87 using phenotypical, karyotyping and sequencing evidences (Park et al., 1990; Chang et al., 2016; Andor et al., 2020). CCNMF analysis uniquely offers an in-depth characterization of clone-specific dosage effect, which allows further hypothesis testing regarding driver gene changes that are determinant to the functional divergence of subclones.

Conclusion

Tumoral and other pathological tissues often demonstrate heterogeneous and clonal genetic nature, such as the pervasive structural and copy number alterations as observed in cancer tissues (Xia et al., 2018b, 2016, 2018a). These alterations will lead to clonal dosage effect, which is an important force driving the clonal development of many diseases. To facilitate understanding of the tissue clonal structure and the associated clone specific dosage sensitivity, in the paper, we introduced a new non-negative matrix factorization based co-clustering approach: CCNMF. Using paired scRNA- and scDNA-seq data of the same specimen, CCNMF optimizes an objective function that simultaneously maximizes for intra-technology clonal compactness, inter-technology clonal coherence and global dosage-expression correlation. We designed, implemented and validated the CCNMF toolkit using both simulated and real cell line mixture benchmarks. We demonstrated the utility of CCNMF by delineating the underlying clonal structure and

clonal differential expression in an ovarian tumor cell line mixture, a patient derived breast cancer xenograft and a gastric cancer cell line. We expect CCNMF to serve as a much needed bioinformatics tool for performing single cell level clonal gene dosage analysis for the community.

Methods

Coupled factorization of scDNA and scRNA data

We first derive the coupled-clone nonnegative matrix factorization framework we will use to identify the clonal structure underlying paired scDNA- and scRNA-seq data from the same biological specimen. Note that the input matrix $O \in R^{p \times n_1}$ is the copy numbers of p genes and n_1 cells, and the input matrix $E \in R^{p \times n_2}$ is the gene expression levels of p genes and n_2 cells. In general, these matrices could be any paired scDNA- and scRNA-seq datasets generated from the same biological specimen (Figure 1).

Nonnegative Matrix Factorization (NMF) is one approach to seek latent low-dimensional representation of a given data matrix (Lee and Seung, 1999, 2001). It factorizes a feature profile (e.g. gene expression, CNV) into two matrices, one describes the structure of features, while the other describes the structure among samples (i.e. cells in this study). To couple the nonnegative factorization of matrices O and E, we additionally define $A \in R^{p \times p}$ to represent the linked sensitivity of gene expression to copy number. The matrix A can be estimated by a linear regression model using publicly available paired RNA and DNA bulk sequencing data (Duren et al., 2017), or by using the uninformative identity matrix. Hence, we simultaneously co-factorize O and E by minimizing the following objective function:

$$\mathcal{F}(W,H) = \min_{W_1,H_1,W_2,H_2 \geq 0} \frac{1}{2} \|O - W_1H_1\|_F^2 + \frac{\lambda_1}{2} \|E - W_2H_2\|_F^2 - \lambda_2 tr(W_2^T A W_1) \quad (1)$$

$$subject\ to: \|W_1\|_F^2 = 1, \|W_2\|_F^2 = 1, W_1, W_2, H_1, H_2 \geq 0,$$

where we denote $W_1 \in R^{p \times k}, W_2 \in R^{p \times k}$ and $H_1 \in R^{k \times n_1}, H_2 \in R^{k \times n_2}$ by shorthands W and H, and $tr()$ is the trace of a matrix.

Note that by minimizing the first two terms of the objective Equation (1) we ensured the respective decomposition of O and E, such that $O = W_1H_1$ and $E = W_2H_2$, where W_i $(i = 1, 2)$ is the cluster centroid matrix for the n_i $(i = 1, 2)$ cells and H_i $(i = 1, 2)$ is the weight matrix that softly assigns n_i $(i = 1, 2)$ cells to these clusters. Upon convergence, H_i provides one with the inferred cluster identities for all single cells. Also note that by minimizing the cross term $-tr(W_2^T A W_1)$ we ensured the coherence of the inferred clonal structure between the scRNA- and scDNA-seq data.

Optimization of the objective function

Next, we apply the alternating direction methods of multipliers (ADMM) (Boyd et al., 2010; Duren et al., 2018) to minimize the objective function. Let μ_1 and μ_2 be the matrices containing the Lagrangian multipliers for W_1 and W_2, thus the transformed objective function is as follows:

$$L(W, H, \mu_1, \mu_2) = \mathcal{F}(W, H) + \sum_{n=1}^{2} \mu_n tr(W_n^T W_n) \tag{2}$$

In order to solve the transformed objective function, we first obtained the required gradients by setting the first order derivatives to zeros. Then, we use the obtained gradients with a descent algorithm to iteratively update and optimize the objective function until convergence by the following steps (**Supplementary Methods**):

$$\begin{aligned}
h_{ij}^1 &\leftarrow h_{ij}^1 \frac{(W_1^T O)_{ij}}{(W_1^T W_1 H_1)_{ij}}, \\
w_{ij}^1 &\leftarrow w_{ij}^1 \frac{(O H_1^T + \lambda_2 A^T W_2 + W_1 m_{11})_{ij}}{(W_1 H_1 H_1^T + W_1 m_{12})_{ij}}, \\
h_{ij}^2 &\leftarrow h_{ij}^2 \frac{(W_2^T E)_{ij}}{(W_2^T W_2 H_2)_{ij}}, \\
w_{ij}^2 &\leftarrow w_{ij}^2 \frac{(E H_2^T + \frac{\lambda_2}{\lambda_1} A W_1 + W_2 m_{21})_{ij}}{(W_2 H_2 H_2^T + W_2 m_{22})_{ij}},
\end{aligned} \tag{3}$$

where

$$\begin{aligned}
m_{11} &\leftarrow tr(W_1^T (W_1 H_1 H_1^T)), \\
m_{12} &\leftarrow tr(W_1^T (O H_1^T + \lambda_2 A^T W_2)), \\
m_{21} &\leftarrow tr(W_2^T (W_2 H_2 H_2^T)), \\
m_{22} &\leftarrow tr(W_2^T (E H_2^T + \frac{\lambda_2}{\lambda_1} A W_1)).
\end{aligned} \tag{4}$$

Model inputs and model selection

The objective function has two parameter inputs: λ_1, λ_2, to initialize the iterative computation. Our experience is that: λ_1 and λ_2 are data-dependent, which can be empirically determined from the input data. In practice, we used an automated balancing strategy to determine the λ_1 and λ_2. The approach ensures the initial values of the three terms of the objective function are within the same order, thus all contribute to the optimization.

The coupling matrix A is also expected as an input, for which one can supply an uninformative identify matrix, with the only non-zeros are the diagonal ones. This setup provides none dosage sensitivity information to the objective function. Alternatively, we can estimate A from known associations between copy number and gene expression using paired bulk sequencing data of the same tissue source. It was well known that DNA copy number is highly positively correlated with RNA expression level for the most ($> 99\%$) of human genes (Fehrmann et al., 2015). We thus calculated a diagonal coupling matrix ($A \in R^{p \times p}$) where each diagonal element is estimated by the ratio of gene-wise mean expression to mean copy number, for example, using the paired bulk RNA-seq and microarray expression data from The Cancer Genome Atlas (TCGA) (https://www.cancer.gov/tcga). We used such empirically obtained matrix A in simulation to generate realistic paired scRNA- and scDNA-seq datasets.

If the number of clusters k is unknown, CCNMF runs several models with providing different k, and the optimal k is selected as the objective function reaching its lowest value.

Preprocessing of scDNA-seq data

While the scRNA-seq data is typically presented as an easy-to-input gene expression matrix indexed by cells and standardized genes, the scDNA-seq data is typically presented as copy numbers indexed by cells and non-standardized genome segmental bins. To map genome segmental bins to standardized genes, we took the following preprocessing steps: (1) We aligned both the scRNA- and scDNA-seq to the same human genome assembly (one can use either GRCh38 or hg19 (Campbell et al., 2019)); (2) We found the genomic location overlaps between the gene annotation tracks and the genome segmental bins using the R package IRanges (Lawrence et al., 2013); (3) We retained only genes that were mapped to at least one genome segmental bin. For a gene mapped to multiple segment bins, we took the integer mean copy number of these bins as the gene's copy number value. Finally, we prepared datasets of these remaining genes, with properly formed paired scRNA- and scDNA-seq matrices $E \in R^{p \times n_2}$ and $O \in R^{p \times n_1}$ for CCNMF analysis.

Simulation procedures

We generated paired scRNA- and scDNA-seq data from the same clonal structure preset by the ground truth genetic copy number (**GCN**) (as illustrated in **Supplementary Figure S1**). First, we set the ground truth GCN profile for a specific clonal structure, by specifying the first clone (cluster) as normal cells with a GCN vector $V_1 = [2, \cdots, 2] \in R^m$, where m is the total number of genes of a genome. We obtained the second clone's GCN vector $V_2 \in R^m$, by replacing a fraction of V_1 values with random numbers sampled from $\{0, 1\}$. Similarly, we obtained the third clone's GCN vector $V_3 \in R^m$, by replacing a fraction of V_1 values with random numbers sampled from $\{3, 4\}$. Finally, we used the ground truth GCN matrix $V = \left[V_1 \times \mathbb{1}^T, V_2 \times \mathbb{1}^T, V_3 \times \mathbb{1}^T\right]$ ($\mathbb{1} \in R^n$, n is the number of cells in cluster) as the shared genomic change profile for the underlying clonal structure.

To simulate scDNA-seq data, we estimated sequencing platform parameters and noise using TCGA's bulk CNV datasets downloaded from cBioPortal (Cerami et al., 2012; Gao et al., 2013). Specifically, (1) We estimated the probability transition matrix $P(C|G = t)$ for the observed integer copy number C when given genetic changes G, where C is the observed CNV, G is the genetic CNV, $t = 0, 1, 2, 3, \cdots$ is the value of copy number (see **Supplementary Figure S2**); (2) We simulated each copy number per gene and cell $D_{ij} \sim multinomial(P(C|G) * P(V))$ when given the clonal GCN matrix V (see **Supplementary Figure S3**); (3) We added outlier and dropout events based on *outlier_percentage* and *dropout_percentage* (see **Supplementary Methods**).

For the paired scRNA-seq data, we used the Splatter pipeline (Zappia et al., 2017) with the genetic copy number V as the baseline to simulate library effects, dropout, outlier events (see **Supplementary Methods**). Splatter parameters were estimated from the single-cell RNA-

seq data which was provided by users and belonged to the same tumor with bulk CNV data downloaded from the cBioPortal.

Simulated and real datasets

As the first benchmark, we simulated 46 paired scDNA- and scRNA-seq datasets, which we referred to as the **Sim** data. We simulated 3-cluster *Bifurcate* and *Linear* clonal structure scenarios. We varied essential experimental parameters, such as the percentages of outliers and dropouts. For each simulated dataset, we randomly generated cell-wise scDNA- and scRNA-seq data according to the specified scenario and parameters using the procedure as detailed in the preceding subsection (see **Supplementary Figure S1**). Each one paired dataset in *Sim*, has 1000 cells and 2000 genes for scDNA- and scRNA-seq respectively, and the three composing clusters have 200, 400 and 400 cells in each. The first cluster is designated as normal cells, and the second and third clusters represent deletion and amplification clones respectively. We assumed differentially acquired deletions and amplifications to affect 10 % to 50 % genome regions. We deposited the *Sim* data in GitHub.

For the second benchmark, we downloaded a set of paired scDNA- and scRNA-seq real data from the European Genome-phenome Archive with accession EGAD00001004553 (Campbell et al., 2019), referred to as the *OV* data. The *OV* data was composed of DLP scDNA-seq and 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 tumors (TOV2295R). The scRNA subset had 1717 and 4918, and the scDNA subset had 371 and 394 (OV2295R and TOV2295R) cells, respectively.

As a real application, we used a set of paired scRNA- and scDNA-seq data from a triple-negative breast cancer patient-derived xenograft SA501X3F, referred to as **SA501**. It was downloaded from the European Genome-phenome Archive with accession EGAS00001002170 (Zahn et al., 2017; Campbell et al., 2019). It has 1430 cells of scRNA-seq data and 260 cells of scDNA-seq data and the underlying clonal structure was unknown (Campbell et al., 2019). The *SA501* scRNA-seq 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 sequencing which measured 20651 copy number segments.

As another application we also downloaded the paired scRNA- and scDNA-seq data for the NCI-N87 gastric cancer cell line from the Gene Expression Omnibus (GSE142750) and the National Institute of Health's short read archive (PRJNA498809) (Andor et al., 2020). We processed the scDNA sequencing data using the Cellranger-DNA pipeline with GRCh38 as reference genome. There are 1005 scDNA-seq single cells, and 154423 copy number bins across all chromosomes per cell. We merged the neighboring small bins into 50kb bins. Thus, we obtained a 1005×3103 copy number matrix. The paired scRNA-seq data has 3246 cells with 13513 genes per cell.

Performance evaluations

To evaluate the performance of CCNMF given the ground truth cell cluster labels, we used the Adjusted Rand Index (ARI) (Rand, 1971; Kiselev et al., 2017). The ARI measures the similarity between the labels assigned by any two clustering schemes as follows:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \tag{5}$$

where n_{ij}, a_i, b_j are values from the two-way contingency table describing the overlapping label counts between the two clustering schemes i and j. 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 $b_j = \sum_i n_{ij}$. Given the ground truth label set, ARI ranges from 0 to 1, where 0 means completely mislabeling and 1 means completely correct labeling. Therefore it can be used as an accuracy measure for clustering.

Funding

L.W. was supported by NSFC grant (No.12071466), NCMIS of CAS, LSC of CAS, and the Youth Innovation Promotion Association of CAS. L.C.X. was supported by the American Cancer Society (132922-PF-18-184-01-TBG) and the Innovation in Cancer Informatics Fund. X.Q.B gratefully acknowledges the financial support from China Scholarship Council (NO.201904910816).

Conflict of interest statement

The authors declare no conflict of interests.

References

Andor N, Lau BT, Catalanotti C, Sathe A, Kubit M, Chen J, Blaj C, Cherry A, Bangs CD, Grimes SM, et al.. 2020. Joint single cell DNA-seq and RNA-seq of gastric cancer cell lines reveals rules of in vitro evolution. *NAR Genomics and Bioinformatics* **2**. Lqaa016.

Ayesha AK, Hyodo T, Asano E, Sato N, Mansour MA, Ito S, Hamaguchi M, and Senga T. 2016. Ube2s is associated with malignant characteristics of breast cancer cells. *Tumor Biology* **37**: 763–772.

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

Campbell JN, Macosko EZ, Fenselau H, Pers TH, Lyubetskaya A, Tenen D, Goldman M, Verstegen AMJ, Resch JM, McCarroll SA, et al.. 2017. A molecular census of arcuate hypothalamus and median eminence cell types. *Nature Neuroscience* **20**: 484–496.

Campbell KR, Steif A, Laks E, Zahn H, Lai D, McPherson A, Farahani H, Kabeer F, O'Flanagan C, Biele J, et al.. 2019. clonealign: statistical integration of independent single-cell rna and dna sequencing data from human cancers. *Genome Biology* **20**: 54.

Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, et al.. 2012. The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* **2**: 401–404.

Chang HR, Park HS, Ahn YZ, Nam S, Jung HR, Park S, Lee SJ, Balch C, Powis G, Ku JL, et al.. 2016. Improving gastric cancer preclinical studies using diverse in vitro and in vivo model systems. *BMC cancer* **16**: 200–200. 26955870[pmid].

Duren Z, Chen X, Jiang R, Wang Y, and Wong WH. 2017. Modeling gene regulation from paired expression and chromatin accessibility data. *Proceedings of the National Academy of Sciences* **114**: E4914–E4923.

Duren Z, Chen X, Zamanighomi M, Zeng W, Satpathy AT, Chang HY, Wang Y, and Wong WH. 2018. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proceedings of the National Academy of Sciences* **115**: 7723–7728.

Efremova M and Teichmann SA. 2020. Computational methods for single-cell omics across modalities. *Nature Methods* **17**: 14–17.

Fehrmann RSN, Karjalainen JM, Krajewska M, Westra HJ, Maloney D, Simeonov A, Pers TH, Hirschhorn JN, Jansen RC, Schultes EA, et al.. 2015. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature Genetics* **47**: 115.

Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, et al.. 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Science Signaling* **6**: pl1–pl1.

Hokonohara K, Nishida N, Miyoshi N, Takahashi H, Haraguchi N, Hata T, Matsuda C, Mizushima T, Doki Y, and Mori M. 2019. Involvement of maf1 homolog, negative regulator of rna polymerase iii in colorectal cancer progression. *International journal of oncology* **54**: 1001—1009.

Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, Wu X, Wen L, Tang F, Huang Y, et al.. 2016. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Research* **26**: 304–319.

Hynds RE, Vladimirou E, and Janes SM. 2018. The secret lives of cancer cell lines. *Disease Models & Mechanisms* **11**.

Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, and Green AR. 2017. Sc3: consensus clustering of single-cell rna-seq data. *Nature Methods* **14**: 483–486.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan M, and Carey V. 2013. Software for computing and annotating genomic ranges. *PLoS Computational Biology* **9**.

Lee DD and Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**: 788–791.

Lee DD and Seung HS. 2001. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*, pp. 556–562. MIT Press.

Letourneau IJ, Quinn MC, Wang LL, Portelance L, Caceres KY, Cyr L, Delvoye N, Meunier L, de Ladurantaye M, Shen Z, et al.. 2012. Derivation and characterization of matched cell lines from primary and recurrent serous ovarian cancer. *BMC Cancer* **12**: 379.

Liu X, Jin G, Qian J, Yang H, Tang H, Meng X, and Li Y. 2018. Digital gene expression profiling analysis and its application in the identification of genes associated with improved response to neoadjuvant chemotherapy in breast cancer. *World Journal of Surgical Oncology* **16**: 82.

van der Maaten L and Hinton G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**: 2579–2605.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al.. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**: 1202–1214.

McCarthy DJ, Rostom R, Huang Y, Kunz DJ, Danecek P, Bonder MJ, Hagai T, Lyu R, Kilpinen H, Goncalves A, et al.. 2020. Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes. *Nature Methods* **17**: 414–421.

Nielsen KV, Müller S, Møller S, Schønau A, Balslev E, Knoop AS, and Ejlertsen B. 2010. Aberrations of erbb2 and top2a genes in breast cancer. *Molecular oncology* **4**: 161–168. 19945923[pmid].

Park JG, Frucht H, LaRocca RV, Bliss DP, Kurita Y, Chen TR, Henslee JG, Trepel JB, Jensen RT, Johnson BE, et al.. 1990. Characteristics of cell lines established from human gastric carcinoma. *Cancer Research* **50**: 2773–2780.

Rand WM. 1971. Objective criteria for the evaluation of clustering methods. *Publications of the American Statistical Association* **66**: 846–850.

Shen Y, Schmidt BUS, Kubitschke H, Morawetz EW, Wolf B, Käs JA, and Losert W. 2020. Detecting heterogeneity in and between breast cancer cell lines. *Cancer Convergence* **4**: 1.

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, and Satija R. 2019. Comprehensive integration of single-cell data. *Cell* **177**: 1888–1902.e21.

Sun Y, Xu Y, Xu J, Lu D, and Wang J. 2015. 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.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, and Siddiqui A. 2009. mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**: 377–382.

Velazquez-Villarreal EI, Maheshwari S, Sorenson J, Fiddes IT, Kumar V, Yin Y, Webb MG, Catalanotti C, Grigorova M, Edwards PA, et al.. 2020. Single-cell sequencing of genomic dna resolves sub-clonal heterogeneity in a melanoma cell line. *Communications Biology* **3**: 318.

Wang M, Hu Y, Amatangelo MD, and Stearns ME. 2011. Role of ribosomal protein rps2 in controlling let-7a expression in human prostate cancer. *Molecular Cancer Research* **9**: 36–50.

Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H, et al.. 2014. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**: 155–160.

Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, and Macosko EZ. 2019. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**: 1873–1887.e17.

Xia LC, Ai D, Lee H, Andor N, Li C, Zhang NR, and Ji HP. 2018a. Svengine: an efficient and versatile simulator of genome structural variations with features of cancer clonal evolution. *GigaScience* **7**: giy081.

Xia LC, Bell JM, Wood-Bouwens C, Chen JJ, Zhang NR, and Ji HP. 2018b. Identification of large rearrangements in cancer genomes with barcode linked reads. *Nucleic acids research* **46**: e19.

Xia LC, Sakshuwong S, Hopmans E, Bell J, Grimes S, Siegmund D, Ji H, and Zhang N. 2016. A genome-wide approach for detecting novel insertion-deletion variants of mid-range size. *Nucleic Acids Research* **44**: e126.

Yang J, Min KW, Kim DH, Son BK, Moon KM, Wi YC, Bang SS, Oh YH, Do SI, Chae SW, et al.. 2018. 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**: e0202113–e0202113.

Zahn H, Steif A, Laks E, Eirew P, VanInsberghe M, Shah SP, Aparicio S, and Hansen CL. 2017. Scalable whole-genome single-cell library preparation without preamplification. *Nature Methods* **14**: 167–173.

Zappia L, Phipson B, and Oshlack A. 2017. Splatter: simulation of single-cell rna sequencing data. *Genome Biology* **18**: 174.

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, and Zhu J. 2017. Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**: 14049.

Zhou Z, Xu B, Minn A, and Zhang NR. 2020. Dendro: genetic heterogeneity profiling and subclone detection by single-cell rna sequencing. *Genome Biology* **21**: 10.

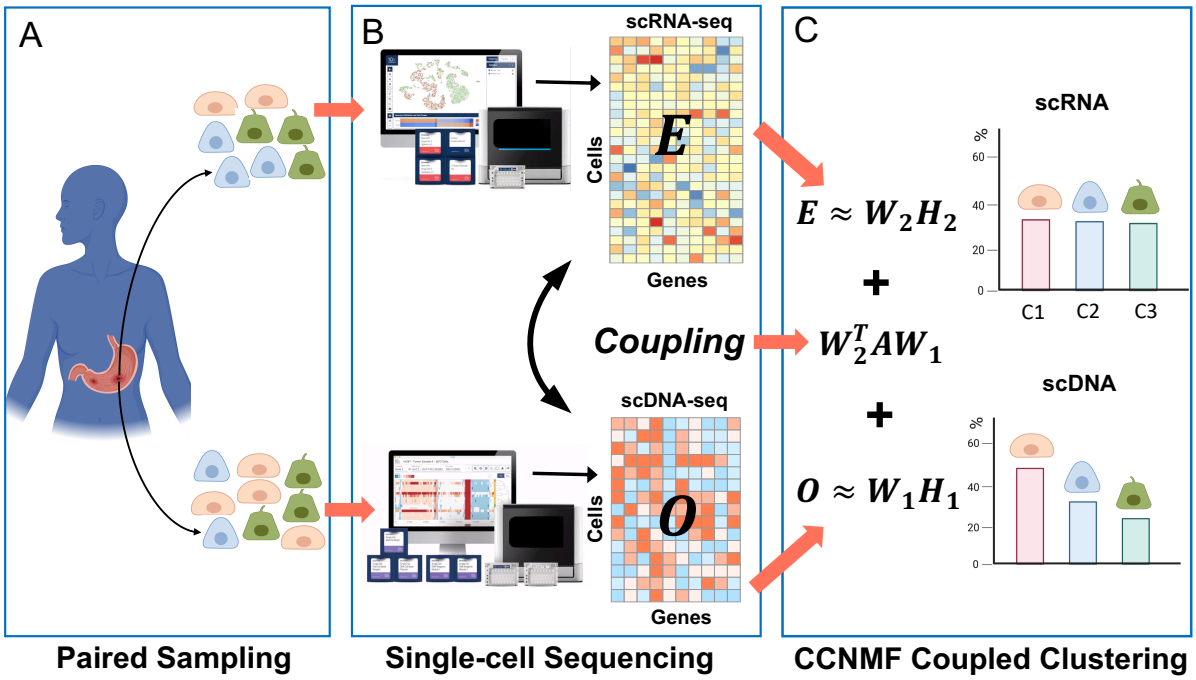

Figure 1: The workflow of coupled-clone nonnegative matrix factorization (CCNMF) for inferring shared clonal structure in paired scDNA- and scRNA-seq data. In order to link paired matrices $E \in R^{n_1 \times p}$ and $O \in R^{n_2 \times p}$, we firstly construct the coupling matrix $A \in R^{p \times p}$ which represents a global concordance between DNA copy number and gene expression. CCNMF unitizes nonnegative matrix factorization (NMF) to jointly infers the shared scRNA and scDNA subclones by maximizing the coupling term $tr(W_2^T AW_1)$.

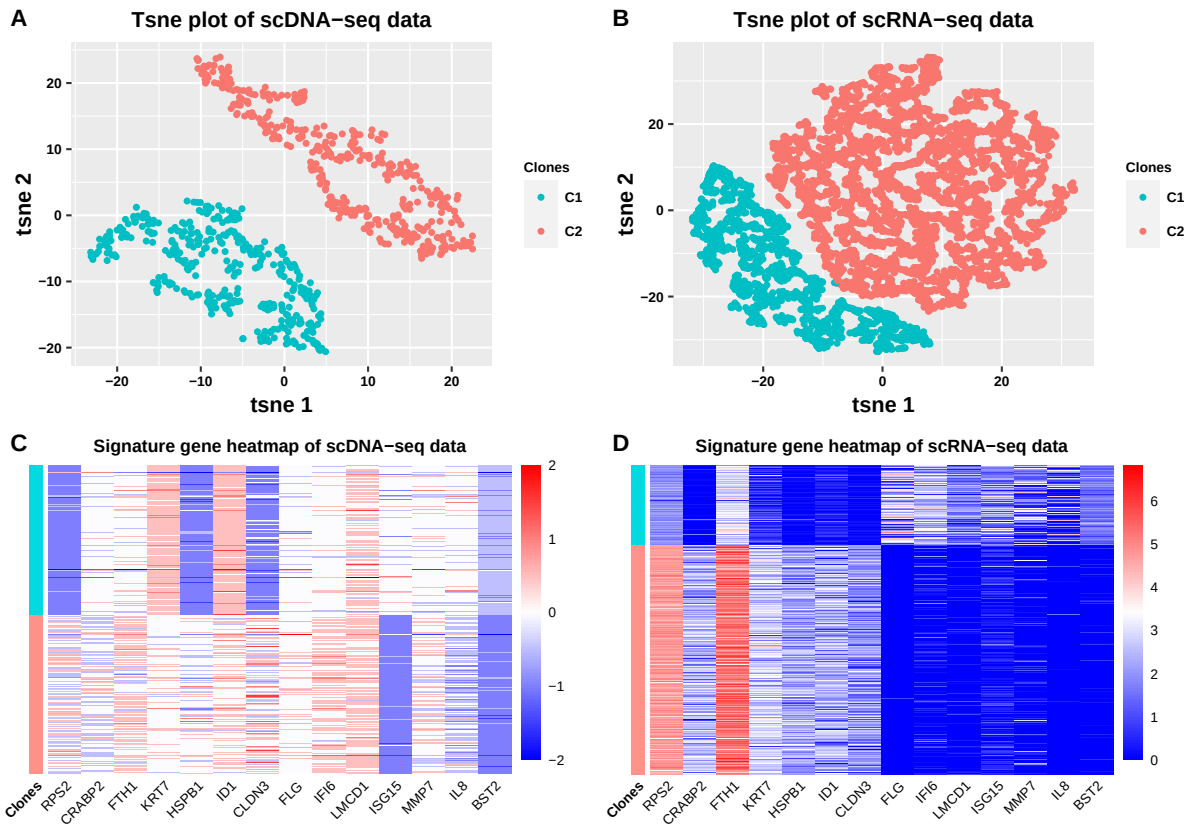

Figure 2: The scDNA and scRNA shared clonal structure of the *OV* cell line data, including tSNE plots of (A) scDNA and (B) scRNA clones, heatmaps of (C) copy number ratios and (D) gene expression of the top 14 clonal signature genes. Cell clusters were color coded as: green for C1/Cluster-1, red for C2/Cluster-2.

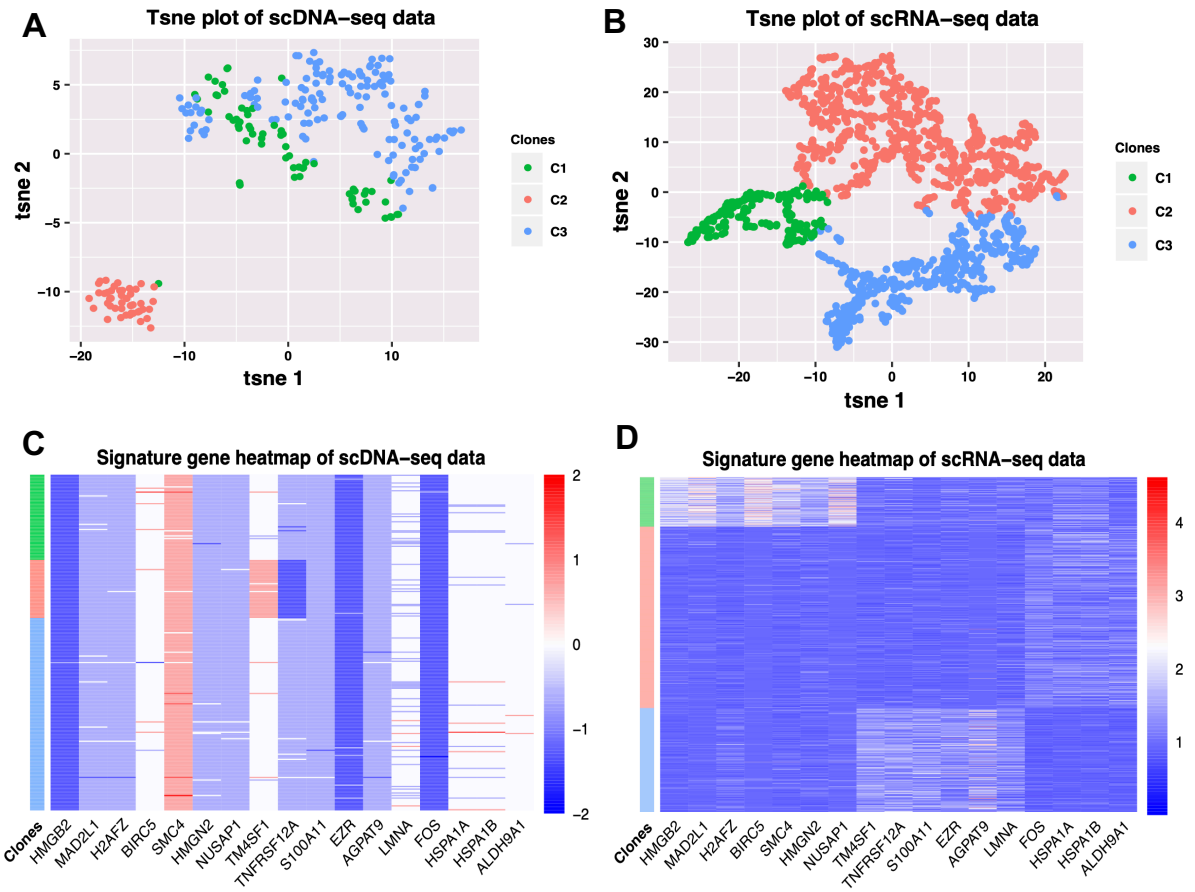

Figure 3: The scDNA and scRNA shared clonal structure of the *SA501* xenograft data, including tSNE plots of (A) scDNA and (B) scRNA clones, heatmaps of (C) copy number ratios and (D) gene expression of the top 17 clonal signature genes. Cell clusters were color coded as: green for C1/Cluster-1, red for C2/Cluster-2 and blue for C3/Cluster-3.

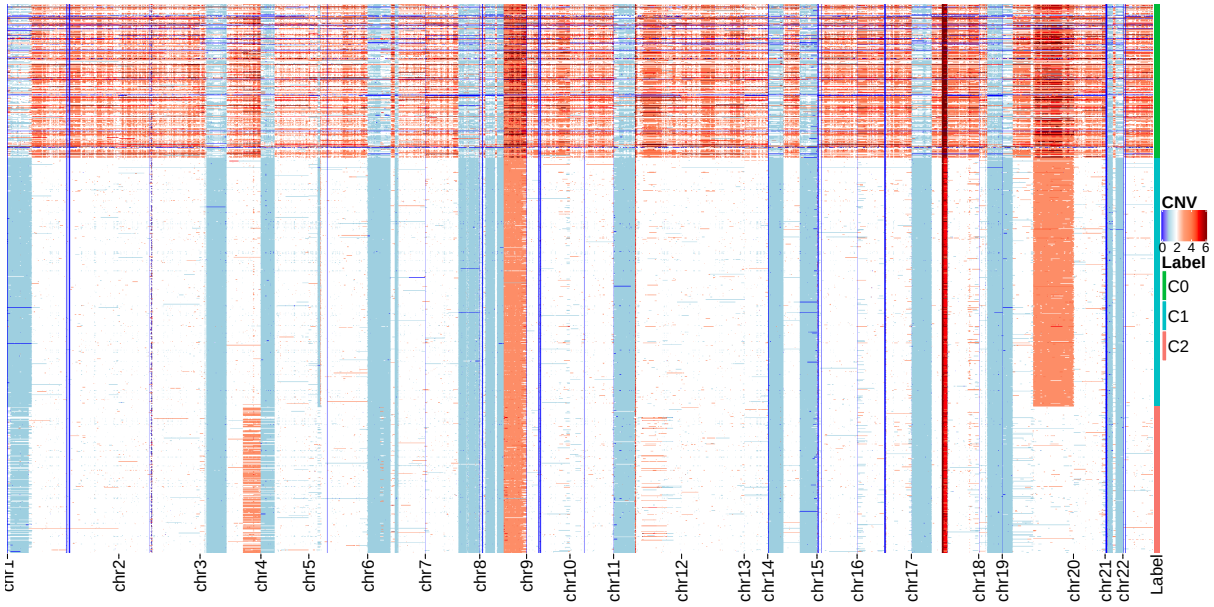

Figure 4: The scDNA subclones of NCI-N87 cell line by CCNMF. Each row represents the genome-wide copy number for a single-cell. The three clusters are respectively colored as green, blue, red. C0: highly variable replicating cells; C1: chr19q amplification; and C2: chr3q amplification.

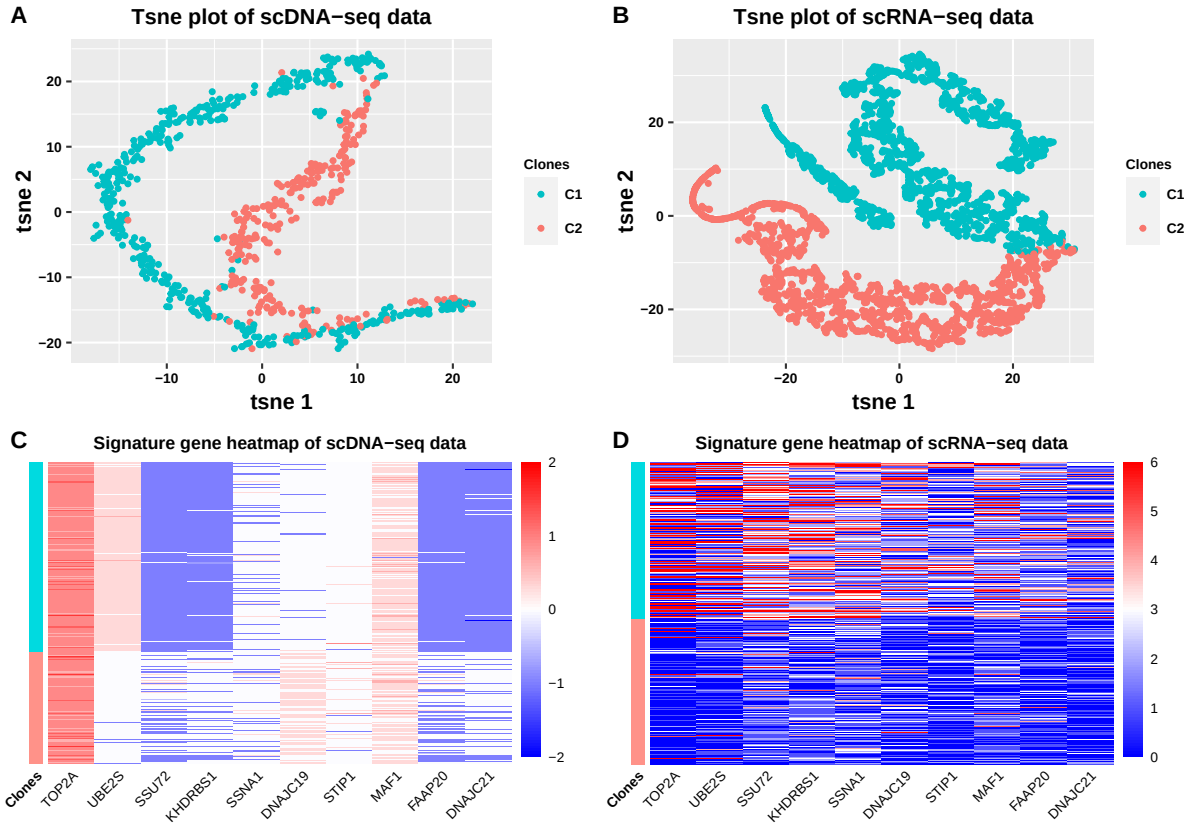

Figure 5: The scDNA and scRNA shared clonal structure of the *NCI-N87* cell line data, including tSNE plots of (A) scDNA and (B) scRNA clones, heatmaps of (C) copy number ratios and (D) gene expression of the top10 clonal signature genes. Cell clusters were color coded as: green for C1/Cluster-1, red for C2/Cluster-2.