Published online 27 July 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 3 **1**
https://doi.org/10.1093/nargab/lqac056

SMGR: a joint statistical method for integrative analysis of single-cell multi-omics data

Qianqian Song [1,2,*], Xuewei Zhu[3], Lingtao Jin[4], Minghan Chen[5], Wei Zhang [1,2] and Jing Su [6,7,*]

[1]Center for Cancer Genomics and Precision Oncology, Wake Forest Baptist Comprehensive Cancer Center, Atrium Health Wake Forest Baptist, Winston-Salem, NC 27157, USA, [2]Department of Cancer Biology, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA, [3]Department of Internal Medicine, Section on Molecular Medicine, Wake Forest School of Medicine, Winston-Salem, NC 27101, USA, [4]Department of Molecular Medicine, UT Health San Antonio, San Antonio, TX 78229, USA, [5]Wake Forest University, Department of Computer Science, Winston-Salem, NC 27109, USA, [6]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN 46202, USA and [7]Section on Gerontology and Geriatric Medicine, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA

Received January 04, 2022; Revised June 16, 2022; Editorial Decision July 11, 2022; Accepted July 20, 2022

ABSTRACT

Unravelling the regulatory programs from single-cell multi-omics data has long been one of the major challenges in genomics, especially in the current emerging single-cell field. Currently there is a huge gap between fast-growing single-cell multi-omics data and effective methods for the integrative analysis of these inherent sparse and heterogeneous data. In this study, we have developed a novel method, Single-cell Multi-omics Gene co-Regulatory algorithm (SMGR), to detect coherent functional regulatory signals and target genes from the joint single-cell RNA-sequencing (scRNA-seq) and single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) data obtained from different samples. Given that scRNA-seq and scATAC-seq data can be captured by zero-inflated Negative Binomial distribution, we utilize a generalized linear regression model to identify the latent representation of consistently expressed genes and peaks, thus enables the identification of co-regulatory programs and the elucidation of regulating mechanisms. Results from both simulation and experimental data demonstrate that SMGR outperforms the existing methods with considerably improved accuracy. To illustrate the biological insights of SMGR, we apply SMGR to mixed-phenotype acute leukemia (MPAL) and identify the MPAL-specific regulatory program with significant peak-gene links, which greatly enhance our understanding of the regulatory mechanisms and potential targets of this complex tumor.

INTRODUCTION

Single-cell multi-omics technologies are emerging for measuring multiple molecule types at individual cell such as scNMT-seq (1) and sci-ATAC-seq (2). These technological developments allow profiling multiple molecular layers at single-cell resolution and assaying cells from multiple samples under different conditions. Single-cell multi-omics technologies are increasingly used to provide deep insights into the complex cellular ecosystem and biological processes. Integrative analysis of single-cell multi-omics data have offered many exciting biological opportunities and revealed the molecular determinants of human diseases (3–5). For example, Stuart *et al*. (6) revealed the putative mechanisms of cell-type-specific epigenomic regulation within their defined mouse cortical cell types. Nativio *et al*. (7) identified molecular pathways and epigenetic alterations underlying late-onset Alzheimer's disease, by integrating transcriptomic and epigenomic profiling of human brains. Bian *et al*. (8) reconstructed genetic lineages and traced the epigenomic and transcriptomic dynamics through single-cell multi-omics. Granja *et al*. (9) identified both patient-shared malignant signatures and patient-specific regulatory features such as RUNX1-linked regulatory elements, via the integrative analysis of single-cell transcriptomic and chromatin-accessibility profiles in acute leukemia. These studies highlight the significance of single-cell multi-omics integration in accelerating the investigations of cell-type definition, gene regulation and illuminat-

*To whom correspondence should be addressed. Tel: +1 336 926 4972; Email: qsong@wakehealth.edu
Correspondence may also be addressed to Jing Su. Tel: +1 317 278 2950; Email: su1@iu.edu

ing the causes and underlying mechanisms of human diseases especially cancers.

The fast advance of single-cell technologies leads to the rapid growth of single-cell multi-omics data. Although there are approaches such as LIGER (10), Seurat v3 (11,12), Conos (13), MNN (14), etc. that analyze single-cell multi-omics data at the cell-level, including batch effects removal and cell integration, none of those can be directly applied to reveal the feature-level characteristics (e.g. genes and peaks). Therefore, there is a clear need to develop a tailored and effective method for analyzing single-cell multi-omics data from the feature-based angle. A reliable and accurate feature-based method needs to overcome the following challenges: (i) the unique technical issues of single-cell data (e.g. dropouts and dispersion) (15–17). For example, single-cell RNA-seq (scRNA-seq) is well acknowledged for sparsity with abundant zeros, meanwhile, single-cell ATAC-seq (scATAC-seq) is also affected by dropout events due to the loss of DNA material during library preparation, i.e. open chromatin regions with no reads due to loss of DNA material during the scATAC-seq protocol. (ii) Batch effects arisen from different operators, experimental protocols (18), and technical variation (19–21), especially that current scRNA-seq data and scATAC-seq data are often generated by different labs.

To address the above challenges, we propose our novel Single-cell Multi-omics Gene co-Regulatory (SMGR) method, for integrative analysis of scRNA-seq and scATAC-seq data. SMGR explicitly disentangles and detects coherent scRNA-seq genes and scATAC-seq peaks, i.e. co-regulatory programs, to gain insights into the transcriptional regulators and targeted genes and further reveal cell-type specific gene regulatory networks. SMGR is validated on both simulation data and experimental data for its capability and accuracy in detecting co-regulatory programs. With the accurate and reliable integrative analysis of single-cell multi-omics data, our SMGR method provides comprehensive insights into cell type-specific gene regulation, thus uncovers the intrinsic molecular underpinnings and enhances the understanding of underlying mechanisms.

MATERIALS AND METHODS

In this work, the co-regulatory program is defined as a set of genes sharing both similar expression profiles and similar chromatin accessibility profiles. That is, for the genes within one co-regulatory program, not only is the chromatin accessibility of those gene regions similar across the cells, but also is the expression pattern similar across the cells. Therefore, genes within a co-regulatory program are more likely to be co- regulated, which are coincide with regulatory hubs controlling their expression. Moreover, in the identified co-regulatory program, the variability of chromatin accessibility will exhibit good concordance with the variation in gene expression levels. This coherence of joint profiles indicates a higher probability that these genes are simultaneously regulated/co-regulated than grouping on accessibility or gene expression alone, showing the advantage of integrative analysis across single-cell multi-omics data.

To reveal the co-regulatory programs, we propose a multi-joint statistical method to explicitly identify consistent patterns across scRNA-seq and scATAC-seq, while removing the effects of different sources of variability. In order to use enough values of scATAC-seq data, we quantify scATAC-seq peak counts by summing all counts within the gene body (22). Thus, the scATAC-seq data is converted to gene-based activity matrix, with the mapped gene and peaks being easily retrieved. As multiple studies (15–17) show that scRNA-seq data gene expression and scATAC-seq peaks can be sufficiently captured by Negative Binomial (NB) distribution or Zero-Inflated NB (ZINB) distribution, we choose either NB or ZINB model to formulate scRNA-seq and scATAC-seq data in our multi-joint statistical SMGR method. An empirical test can help users decide and choose the appropriate model for their respective data.

The SMGR model

Specifically, let x_{ij_k} denotes the gene i's expression (x_{ij_1}) in cell j_1, or gene i's activities (x_{ij_2}) quantified from peaks in cell j_2, from the single-cell multi-omics data ($k = 1$ refers to scRNA-seq, $k = 2$ refers to scATAC-seq; $i \in \{1, \ldots, m\}$; $j_k \in \{1, \ldots, n_k\}$). When considering the negative binomial distribution, without loss of generality, we assume that the measured counts x_{ij_k} for cell j_k in each of the single-cell multi-omics dataset k follows the NB distribution $NB(u, \varphi)$, which has the probability function as:

$$f\left(x_{ij_k}; \varphi, u\right) = \frac{\Gamma\left(x_{ij_k} + \varphi\right)}{\Gamma(\varphi)\,\Gamma\left(1 + x_{ij_k}\right)} \left(\frac{\varphi}{\varphi + u}\right)^{\varphi} \times \left(\frac{u}{\varphi + u}\right)^{x_{ij_k}}$$

As u and φ are different regarding different genes ($i \in \{1, \ldots, m\}$) across single-cell data, we have:

$$f\left(x_{ij_k}; \varphi_{ik}, u_{ik}\right) = \frac{\Gamma\left(x_{ij_k} + \varphi_{ik}\right)}{\Gamma(\varphi_{ik})\,\Gamma\left(1 + x_{ij_k}\right)} \times \left(\frac{\varphi_{ik}}{\varphi_{ik} + u_{ik}}\right)^{\varphi_{ik}} \left(\frac{u_{ik}}{\varphi_{ik} + u_{ik}}\right)^{x_{ij_k}}$$

Here, u_{ik} represents the estimation for the intrinsic signals across cells, φ_{ik} is the dispersion parameter, and $\sigma_{ik}^2 = u_{ik} + \frac{{u_{ik}}^2}{\varphi_{ik}}$ represents the square deviation of the observed values across cells.

When considering the Zero-Inflated NB distribution $ZINB(\rho, u, \varphi)$ distribution, which is a mixture distribution assigning ρ_{ik} to extra zeros and $(1 - \rho_{ik})$ to a negative binomial distribution, where $0 \le \rho_{ik} \le 1$. As ρ_{ik}, u_{ik} and φ_{ik} are different regarding different genes ($i \in \{1, \ldots, m\}$) across single-cell multi-omics data, we have:

$$f\left(x_{ij_k}; r_{ik}, \varphi_{ik}, u_{ik}\right) = \sum_{x_{ij_k}=0} \rho_{ik} + (1 - \rho_{ik}) \cdot f\left(x_{ij_k} = 0; \varphi_{ik}, u_{ik}\right) + \sum_{x_{ij_k}>0} (1 - \rho_{ik})\, f\left(x_{ij_k}; \varphi_{ik}, u_{ik}\right)$$

Let $y_i = (y_{i1}, \ldots, y_{iz})'$ be a vector consisting of z unobserved latent representation that are shared by the

scRNA-seq and scATAC-seq datasets. As Generalized Linear Model (GLM) has been frequently used (23–27) in single cell data to alleviate extreme value effects, we then use GLM model formulated as:

$$\log u\left(x_{ij_k}|y_i\right) = \beta_{j_k} + \gamma_{j_k} y_i,$$

to distinguish the intrinsic biological signals y_i from the extrinsic variability (β_{j_k} and γ_{j_k}) including the technical variances at the cell-level (j_k) and batch effects across different omics layers (k). In this way, $u(x_{ij_k}|y_i)$ is the conditional mean of x_{ij_k} given y_i, which is composed of the intrinsic biological signals of genes and peaks captured by latent representation y_i without confounding variabilities across cells and molecular layers. Technical variances and batch effects are captured by offsets β_{j_k} and scale factors γ_{j_k}.

In this way, the original single-cell multi-omics data is projected as a z-dimensional latent representation Y by this generalized linear model, with technical biases and batch effects removed during the projection. In this latent representation, the biological levels of scRNA-seq genes and scATAC-seq peaks are represented as y_i, the coherent patterns of genes and peaks can be identified by clustering similar y_i in the latent representation.

Optimization of SMGR

To estimate the parameters of NB or ZINB, we use the maximum likelihood approach. As is assumed above that x_{ijk} follows the NB distribution or ZINB distribution, the conditional log-likelihood function of x_{ij_k} of NB distribution can be written as

$$\begin{aligned} log f\left(x_{ij_k}|y_i, \beta_{j_k}, \gamma_{j_k}, \varphi_{ik}\right) &= \log\left(\Gamma\left(x_{ij_k} + \varphi_{ik}\right)\right) \\ &\quad - \log\left(\Gamma\left(\varphi_{ik}\right)\right) - \log\left(\Gamma\left(x_{ij_k} + 1\right)\right) \\ &\quad + x_{ij_k} \cdot \log\left(u_{ik}\right) + \varphi_{ik} \cdot \log\left(\varphi_{ik}\right) \\ &\quad - \left(\varphi_{ik} + x_{ij_k}\right) \cdot \log\left(\varphi_{ik} + u_{ik}\right) \end{aligned}$$

For zero-inflated negative binomial distribution $ZINB(r_{ik}, u_{ik}, \varphi_{ik})$, in which $u_{ik} = \exp(\beta_{j_k} + \gamma_{j_k} y_i)$, we have:

$$\begin{aligned} &log f\left(x_{ij_k}|\boldsymbol{y}_i, \beta_{j_k}, \gamma_{j_k}, \varphi_{ik}\right) \\ &= \sum_{x_{ij_k}=0} \left[\log\left(\rho_{ik} + (1-\rho_{ik}) \cdot \left(\frac{\varphi_{ik}}{\varphi_{ik}+u_{ik}}\right)^{\varphi_{ik}}\right)\right] \\ &+ \sum_{x_{ij_k}>0} \left[\log\left(1-\rho_{ik}\right) + \log\left(\Gamma\left(x_{ij_k}+\varphi_{ik}\right)\right) - \log\left(\Gamma\left(\varphi_{ik}\right)\right)\right. \\ &- \log\left(\Gamma\left(x_{ij_k}+1\right)\right) + x_{ij_k} \cdot \log\left(u_{ik}\right) + \varphi_{ik} \cdot \log\left(\varphi_{ik}\right) \\ &\left. - \left(\varphi_{ik} + x_{ij_k}\right) \cdot \log\left(\varphi_{ik} + u_{ik}\right)\right]. \end{aligned}$$

For the latent representation y_i, $f(y_i)$ represents the density function of the standard multivariate normal distribution $N(0, I)$. We assume that, given y_i, x_{ij_k} are conditionally independent. Therefore, the joint log-likelihood of (x_{ij_k}, y_i) can be written as

$$\begin{aligned} &l\left(x_{ij_k}, y_i; \beta_{j_k}, \gamma_{j_k}\right) \\ &= \sum_{i=1}^{m} \left\{ \sum_{k=1}^{K} \sum_{j_k=1}^{n_k} \left\{\log f(x_{ij_k}|y_i, \beta_{j_k}, \gamma_{j_k}) + \log f\left(y_i\right)\right\} \right\}, \end{aligned}$$

with log-likelihood as $\log f(x_{ij_k}|y_i, \beta_{j_k}, \gamma_{j_k})$ shown as above and the $\log f(y_i)$ is the log-likelihood of normal distribution. Now we have the joint log-likelihood as $l(x_{ij_k}, y_i; \beta_{j_k}, \gamma_{j_k})$. Through maximizing this joint log-likelihood, that is,

$$\begin{aligned} &\max_{\beta_{j_k}, \gamma_{j_k}} l(x_{ij_k}, y_i, \beta_{j_k}, \gamma_{j_k}) \\ &= \max_{\beta_{j_k}, \gamma_{j_k}} \sum_{i=1}^{m} \left\{ \sum_{k=1}^{K} \sum_{j_k=1}^{n_k} \left\{\log f(x_{ij_k}|y_i, \beta_{j_k}, \gamma_{j_k}) + \log f\left(y_i\right)\right\} \right\}. \end{aligned}$$

Since the latent representation y_i and the model parameters (β_{j_k} and γ_{j_k}) are conditionally dependent, we use a two-step iterative approach to solve the above optimization problem.

- Step 1: parameter estimation. The parameters β_{j_k} and γ_{j_k} in the omics-type specific ZINB models are estimated conditional on y_i. Specifically, the maximization problem of $\max_{\beta_{j_k}, \gamma_{j_k}} \sum_{i=1}^{m} \sum_{j_k=1}^{n_k} \log f(x_{ij_k}|y_i, \beta_{j_k}, \gamma_{j_k})$ is solved for each omics type k using the iteratively reweighted least squares (IWLS) algorithm (28).
- Step 2: latent representation update. Conditional on the estimated parameters β_{j_k} and γ_{j_k} from Step 1, the latent representation y_i is updated to maximize the log-likelihood function. Since y_i is not observable, the Markov Chain Monte Carlo (MCMC) simulation with Metropolis-Hastings sampler (29) is used during optimization. Briefly, an L-step random walk is performed, with the $(l+1)$'th step as:

$$\begin{cases} y^* & \text{, if } l\left(x_{ij_k}, y^*; \beta_{j_k}, \gamma_{j_k}\right) - l\left(x_{ij_k}, y_i^{(l)}; \beta_{j_k}, \gamma_{j_k}\right) > \ln u \\ y_i^{(l)} & \text{, if other wise} \end{cases}$$

, where $l = 1, \cdots, L$, the vector y^* is drawn from the joint posterior distribution

$$\begin{aligned} f\left(y_i, x_{ij_k}\right)|_{y_i^{(l)}} &\propto f\left(y_i\right)|_{y_i^{(l)}} \prod_{j_k=1}^{n_k} \prod_{k=1}^{K} \\ &\times f(x_{ij_k}|y_i^{(l)}, \beta_{j_k}, \gamma_{j_k}), \end{aligned}$$

and $u \in [0, 1]$ is a uniform random number. Then the mean of all random walk steps, $\sum_l y_i^{(l)}/L$, is used to update the latent representation. In this step, we set $L = 200$ draws, with another 200 burn-ins. The optimization algorithm is summarized in Supplement file 1.

With the estimated latent representation y_i, we cluster them in the latent space by k-means to identify the co-regulatory programs that present coherent patterns of gene expression and chromatin activities. The dimension of the latent space and the number of clusters can be determined with the lowest Bayesian information criterion (BIC). This BIC statistic is calculated as $\log(\hat{n}) * \hat{k} - 2 * l(x_{ij_k}, y_i; \beta_{j_k}, \gamma_{j_k})$, where $\hat{n}$ is the data size and $\hat{k}$ is the number of parameters. With the number of clusters in the latent space, we obtain the concordant patterns, i.e. the co-regulatory programs, between scRNA-seq data and scATAC-seq data.

In this work, SMGR identifies such co-regulatory programs across scATAC-seq and scRNA-seq data simultaneously, based on the co-variation between chromatin ac-

cessibility and gene expression, while addressing the biological difference between the chromatin accessibility and the transcriptomics profiles, to provide information about gene co-regulation. Such identified co-regulatory program can be applied to predict gene co-regulatory networks and enriched transcription factors. The co-regulatory programs identified through combining gene expression with chromatin accessibility is useful for identifying functional regulators and recovering the regulatory mechanisms in diseases.

Benchmark methods and evaluation indices

To compare the SMGR's performance, we used the benchmarking methods including SOMatic (30) and SCENIC (31). The SOMatic method first identifies gene clusters and peak clusters in scRNA-seq and scATAC-seq data separately by self-organizing map (SOM), and then associate the two types of clusters by a linking function to identify the gene clusters where their regions have similar peak activities. Different with SOMatic, SMGR joint models the gene expression and peak activities from integrated scRNA-seq and scATAC-seq data profiling. Each of the comparing method is evaluated by the following evaluation indices: the adjusted Rand index (ARI) (32), the Davies-Bouldin (DB) index (33), the Dunn's index (34), the Calinski-Harabasz (CH) index (35) and the Silhouette index (36). These evaluation metrics provide evaluations of the similarity of genes within a co-regulatory program, as well as their differences between different co-regulatory programs. Thus, we use these metrics to evaluate the performance of SMGR. Larger values of the ARI, Dunn's index, CH indices, Silhouette, and smaller values of the DB index, indicate better results.

Identification of transcriptional factor. We use the GENIE3 (37) method to identify the upstream regulatory transcriptional factors (TFs) and regulatory networks based on the co-regulatory programs. Specifically, GENIE3 uses a tree-based regression model to predict regulators of genes and reveal the TF-target interactions, which has been shown with superior performance in different gene expression data (38,39). In this work, the input of GENIE3 are genes within the identified co-regulatory programs by SMGR, and the output is the upstream regulator and regulatory network respectively.

RESULTS

Schematic overview of SMGR

We propose the Single-cell Multi-omics Gene co-Regulatory (SMGR) method, to detect coherent functional patterns of genes and peaks, i.e. co-regulatory programs, from scRNA-seq and scATAC-seq data, which links regulatory elements with target genes that have significant biological context and enables the best exploit of single-cell multi-omics data. We hypothesize the existence of latent representation that captures the intrinsic signals of scRNA-seq gene expression and scATAC-seq peaks, which are not affected by extrinsic variances from different molecular layers. With the assumption that scRNA-seq and scATAC-seq data shows the characteristics of negative binomial distributions or zero-inflated negative binomial distributions, we utilized a generalized linear model with latent representation to formulate the single-cell multi-omics data. In this way, the co-regulatory programs can be identified through clustering of latent representation, thus enables the identification of regulatory mechanisms. Importantly, SMGR preserves biological variations without being influenced by technical variances from different omics layers. For biological discovery, we apply SMGR to the scRNA-seq and scATAC-seq data from peripheral blood mononuclear cell (PBMC) and mixed-phenotype acute leukemia (MPAL). Detailed explanations of SMGR are included in the Materials and Methods. Figure 1 provides an illustrative overview of the SMGR method and its featured analysis. The software for implementing SMGR is available at https://github.com/QSong-github/SMGR.

Evaluation of SMGR using simulation data

We first compared SMGR with SOMatic to identify the co-regulatory programs based on the simulation data of scRNA-seq and scATAC-seq. Details of simulation data generation were provided in the Data availability. As shown in Figure 2A, the box plots present the evaluation scores of the co-regulatory programs identified by SMGR and SOMatic respectively on 10 simulation datasets. Each dataset consists of both simulated scRNA-seq and scATAC-seq. The evaluation score is calculated by averaging the corresponding evaluation index in simulated scRNA-seq and scATAC-seq. The results show that SMGR accurately identifies each co-regulatory program in 10 simulation datasets and demonstrates higher CH index in $\log_{10}$ value (mean $\pm$ SE: 2.85 $\pm$ 0.06) than SOMatic (mean $\pm$ SE: 2.76 $\pm$ 0.07), which suggests that the programs identified by SMGR are denser and better separated in both scRNA-seq and scATAC-seq data than those identified by SOMatic. Moreover, SMGR also shows higher Silhouette (mean $\pm$ SE: 0.27 $\pm$ 0.03), Dunn scores (mean $\pm$ SE: 0.62 $\pm$ 0.1), and lower DB Index (mean $\pm$ SE: 3.56 $\pm$ 2.3), suggesting better inter-program separation and intra-program compactness in both simulated scRNA-seq and scATAC-seq data. Notably, given the ground truth for these simulation data, SMGR shows significantly better ARI scores (mean $\pm$ SE: 0.84 $\pm$ 0.01) than SOMatic (ARI, mean $\pm$ SE: 0.77 $\pm$ 0.02).

Performance evaluation on experimental data

To further demonstrate the performance of SMGR, we compared it with SCENIC on experimental data. Since SCENIC aims to identify gene regulatory network from scRNA-seq data, here we systematically compared the performance of SCENIC and SMGR using 14 benchmarking scRNA-seq data (40). Specifically, the co-regulatory programs identified by SMGR are compared with the regulatory networks identified by SCENIC. As shown in Figure 2B, the box plots present the evaluation scores of the co-regulatory programs identified by SMGR and the gene regulatory networks identified by SCENIC, respectively. Based on the 14 benchmarking datasets, the co-regulatory programs identified by SMGR show higher CH index in

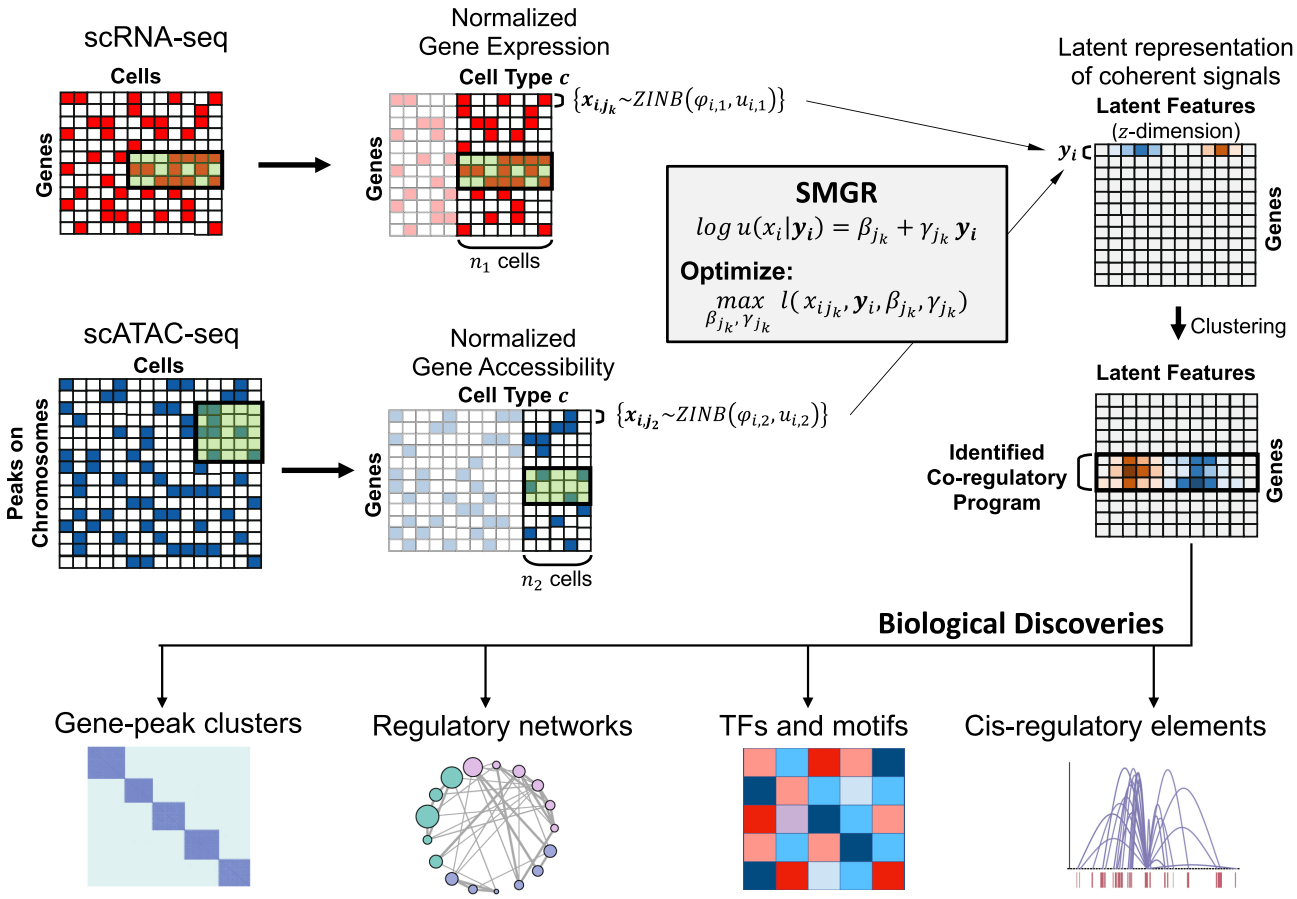

Figure 1. Schematic overview of SMGR. SMGR identifies the latent representation and the co-regulatory programs from scRNA-seq and scATAC-seq data. SMGR enables featured downstream analysis including the visualization of co-regulatory programs, construction of regulatory networks, identification of enriched TFs, motifs, and the cis-regulatory elements.

$\log_{10}$-value (mean ± SE: 3.15 ± 0.61) than the regulatory networks by SCENIC (mean ± SE: 1.06 ± 1.61). Moreover, SMGR also shows higher Silhouette (mean ± SE: 0.047 ± 0.05), Dunn scores (mean ± SE: 0.055 ± 0.02), and lower DB Index (mean ± SE: 5.56 ± 3.19) than SCENIC. These results suggest better inter-program separation and intra-program compactness in the SMGR-identified co-regulatory programs than the SCENIC's regulatory networks.

Moreover, we also evaluate the performance of SMGR and SOMatic on the real experimental data of MPAL dataset. As shown in Figure 3A and B, based on healthy-like and the lymphoid-like cells from MPAL dataset, the co-regulatory programs identified by SMGR show higher CH index than SOMatic, in both healthy-like cells (2.39 versus 2.25) and lymphoid-like cells (5.60 versus 3.98). Moreover, these co-regulatory programs by SMGR show higher Silhouette and Dunn index than SOMatic, and lower DB index, in both healthy-like cells (Silhouette: 0.16 versus 0.14; Dunn: 0.13 versus 0.10; DB: 9.45 versus 11.92) and lymphoid-like cells (Silhouette: 0.35 versus 0.28; Dunn: 1.54 versus 1.09; DB: 12.36 versus 19.60). Collectively, SMGR showed superior performance on real experimental datasets and proved to achieve the best identification of co-regulatory programs from joint scRNA-seq and scATAC-seq data.

Functionality evaluation on experimental data

As a co-regulatory program is likely to be enriched with biological functions (41,42), we compare the extent to which different methods affect the functional discovery. As shown in Figure 3C, the corresponding bar plots present the enriched functions of co-regulatory programs for the lymphoid-like cells (L-) and healthy-like (H-) cells. Notably, SMGR shows more enriched pathways than SOMatic for lymphoid-like (SMGR: 44 versus SOMatic: 32) and healthy-like cells (SMGR: 47 versus SOMatic: 35). In terms of GO enrichment, SMGR also shows more enriched terms than SOMatic for lymphoid-like (SMGR: 37 versus SOMatic: 24) and healthy-like cells (SMGR: 44 versus SOMatic: 32). Further details can be found in Supplementary File S1.

Moreover, we use GENIE3 (37) to identify the upstream TFs based on the co-regulatory programs from SOMatic and SMGR respectively, since genes in a regulatory program are often regulated by the same upstream TFs. Based on the top co-regulatory program identified by SMGR and

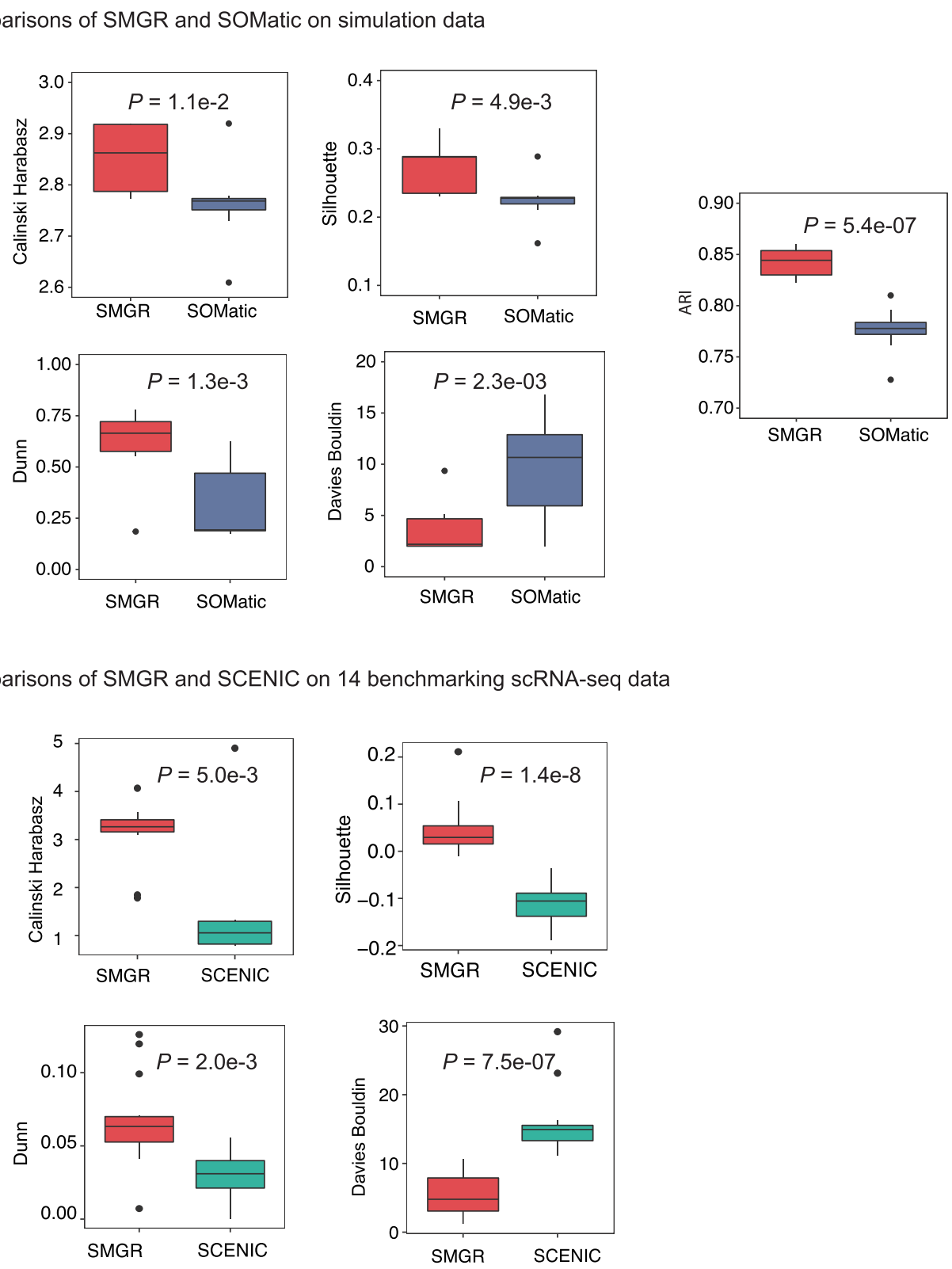

Figure 2. Benchmark SMGR with SOMatic and SCENIC. (**A**) Comparisons of SMGR and SOMatic on 10 simulation data. (**B**) Comparisons of SMGR and SCENIC on 14 benchmarking scRNA-seq data. *P*-values of the comparisons are provided using t-test. Horizontal lines in the middle of boxes indicate median values.

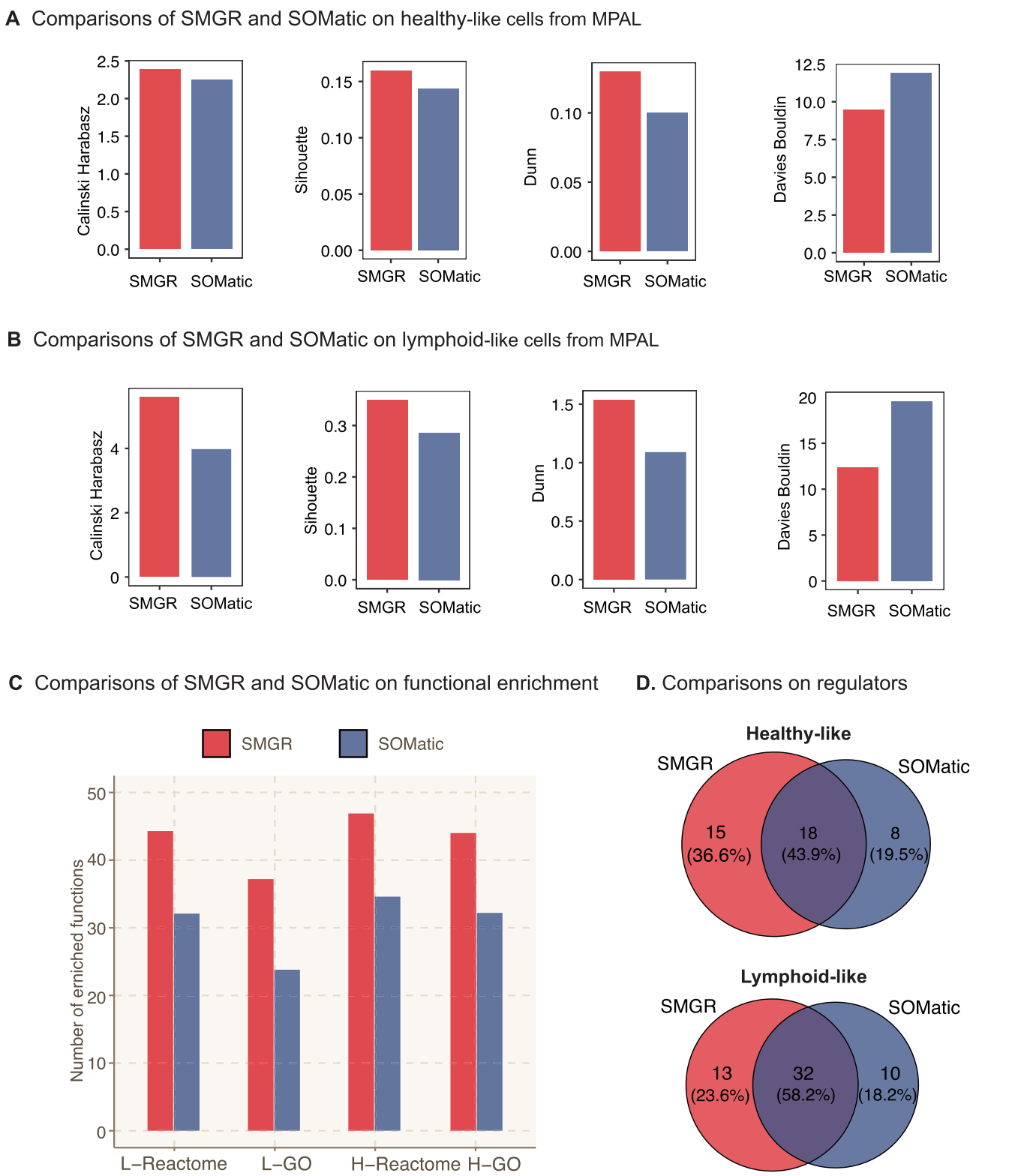

Figure 3. Evaluation of SMGR using experimental data of scRNA-seq and scATAC-seq. (**A**) Comparisons of SMGR and SOMatic on healthy-like cells from MPAL. (**B**) Comparisons of SMGR and SOMatic on lymphoid-like cells from MPAL. (**C**) Comparisons of SMGR and SOMatic on functional enrichment. L-Reactome and L-GO represents the average number of the enriched Reactome pathways and GO terms (adjusted P value < 0.05) in the lymphoid-like cells. H-Reactome and H-GO represent the average enriched Reactome pathways and GO terms in the healthy-like cells. (**D**) Comparisons between SMGR and SOMatic in revealing upstream transcriptional factors.

SOMatic, respectively, the Venn diagrams (Figure 3D) show the number of predicted TFs. For the healthy-like cells, SMGR shows 18 common TFs which are also identified by SOMatic, and 15 specific TFs which are only identified by SMGR. For the lymphoid-like cells, SMGR shows 32 common TFs with SOMatic, and 13 specific TFs. The common TFs shared by SMGR' and SOMatic's program, such as RUNX1 (43) and NFE2 (44,45), are known TFs that play important roles in leukemia. Meanwhile, some TFs which are known to be important in acute leukemia can only be identified by the SMGR's program but not by SOMatic, for example, RUNX3 (46,47), SNAI1 (48) and THAP1 (49). Of note, SMGR reveals more regulators based on the co-regulatory program, suggesting the validity and superior performance of SMGR.

SMGR identifies cell-type specific co-regulatory programs

Biological samples in clinical or experimental studies are often heterogeneous mixtures with different cell populations and cellular states. To fundamentally characterize cell populations, it is necessary to unveil the heterogeneity through the integrative analysis of both transcriptional level and epigenomics level. Herein, we applied SMGR to the scRNA-seq and scATAC-seq data of PBMC (Figure 4A) that consists of 12 different cell types including CD4 naïve T cell and memory T cell. SMGR identified the co-regulatory programs in each of the cell types, with the corresponding gene expression and gene-level chromatin activity shown in Figure 4B and C respectively. Here we only show the top expressed co-regulatory programs in each cell type. Specifically, for CD14+ monocytes, we identified the top co-regulatory program across scRNA-seq (Figure 4B) and scATAC-seq data (Figure 4C), including RAC2 that contributes to the activated NADPH oxidase in monocytes (50), CD44 that mediates cell-cell interaction and participates in monocyte differentiation (51), and KLF4 as a critical regulator of monocyte differentiation (52). These genes present similar patterns in both scRNA-seq and scATAC-seq data of CD14+ monocytes. Similarly, for CD8 naïve T cells, we identified the top co-regulatory program including CD8A, TRBC2 that is involved in T-cell antigen receptor (TCR) complex, and HMGB1 that induces cytokine secretion (53), which consistently expressed across scRNA-seq and scATAC-seq data of CD8 naïve T cells. We also listed the co-regulatory program of Double negative T cells and pDC cells based on the scRNA-seq (Figure 4B) and scATAC-seq data (Figure 4C). For Double negative T cells, we identified the co-regulatory program with genes such as GAPDH, which is a key player in T cell development and function (54), and CD247 that regulates T-cell activation (55). For pDC cells, we obtained the co-regulatory program with CXCR4 that regulates dendritic cell location and activation (56) and CDC37. Full tables of top co-regulatory programs of each cell type are listed in Supplementary Table S1.

To further quantify the coherent patterns of SMGR's co-regulatory programs, we characterize them in the latent representation and observed clearly consistent patterns (Figure 4D). Additionally, we did enrichment analysis of CD4 naïve T cells (red) and CD4 memory T cells (blue), respectively (Figure 4E). CD4 naïve T cells are shown to be enriched with JAK-STAT and platelet signaling, while CD4 memory T cells are enriched with activation of CSF3 (G-CSF) signaling and different interleukin signaling pathways. These analyses suggest that our identified co-regulatory programs contribute to the elucidation of functions in different cell phenotypes through incorporating scRNA-seq and scATATC-seq by SMGR.

Application of SMGR to single-cell multi-omics data of MPAL cells

To demonstrate the functional utility, we next extended the application of SMGR to the scRNA-seq and scATAC-seq data (Figure 5A) of mixed-phenotype acute leukemia (MPAL) (9). For revealing the underlying malignancy mechanisms in MPAL, here we applied SMGR to the healthy-like and lymphoid-like cells respectively. SMGR detected eight co-regulatory programs in healthy-like cells and ten co-regulatory programs in lymphoid-like cells (Supplementary Table S2). Based on these programs, we then identified the most differential ones between lymphoid-like cells and healthy-like using Wilcox test. The co-regulatory programs in lymphoid-like cells that were mostly differential with healthy-like cells were L1, L2 and L3. L1 included genes such as SIGLEC7 (57), SLC8A1 (58) and RUNX1 (43) that play important roles in AML. L2 and L3 consisted of genes including MDM2 (59) and CD36 (60) respectively. In contrast, the co-regulatory programs in healthy-like cells that were most distinct with lymphoid-like cells were H1, H2, H3, H4 (Figure 5B). Example genes identified in each program were listed accordingly. These differential co-regulatory programs suggested potential regulatory mechanisms involved in MPAL. To link genes with regulatory TFs, we then identified significant regulatory networks using GENIE3 (37) (Figure 5C), based on the three differential co-regulatory programs in lymphoid-like cells. The hub regulatory network was shown with RUNX1, FOS and NFE2. Specifically, RUNX1 has well known function in blood cells (43). The full network table was listed in Supplementary Table S3.

With the specific TFs in lymphoid-like cells, we investigated their scATAC-seq coverage and found that NFE2 showed strongly distinct chromatin accessibility between lymphoid-like cells and healthy-like cells (Figure 5D). The bottom peak-to-gene links colored by pearson correlation of peak accessibility and gene expression suggested that NFE2 bound to the linked distal regulatory region of CBX5. CBX5 encodes the HP1α protein, which is a key heterochromatin protein and is critical in chromatin condensation and chromosome segregation. HP1α has been determined to be required for leukemia cell maintenance (61). Meanwhile, NFE2 was also shown to putatively regulate HNRNPA1, which is suggested as a diagnostic marker and therapeutic target for chronic myeloid leukemia (44,45).

Moreover, we performed differential chromatin accessibility analysis and footprinting analysis. Differential gene accessibility analysis comparing lymphoid-like cell with healthy-like cells observed all three TFs were identified with high increase in accessibility in lymphoid-like cells (Figure 5E). That is, RUNX1, FOS and NFE2 have overall lower

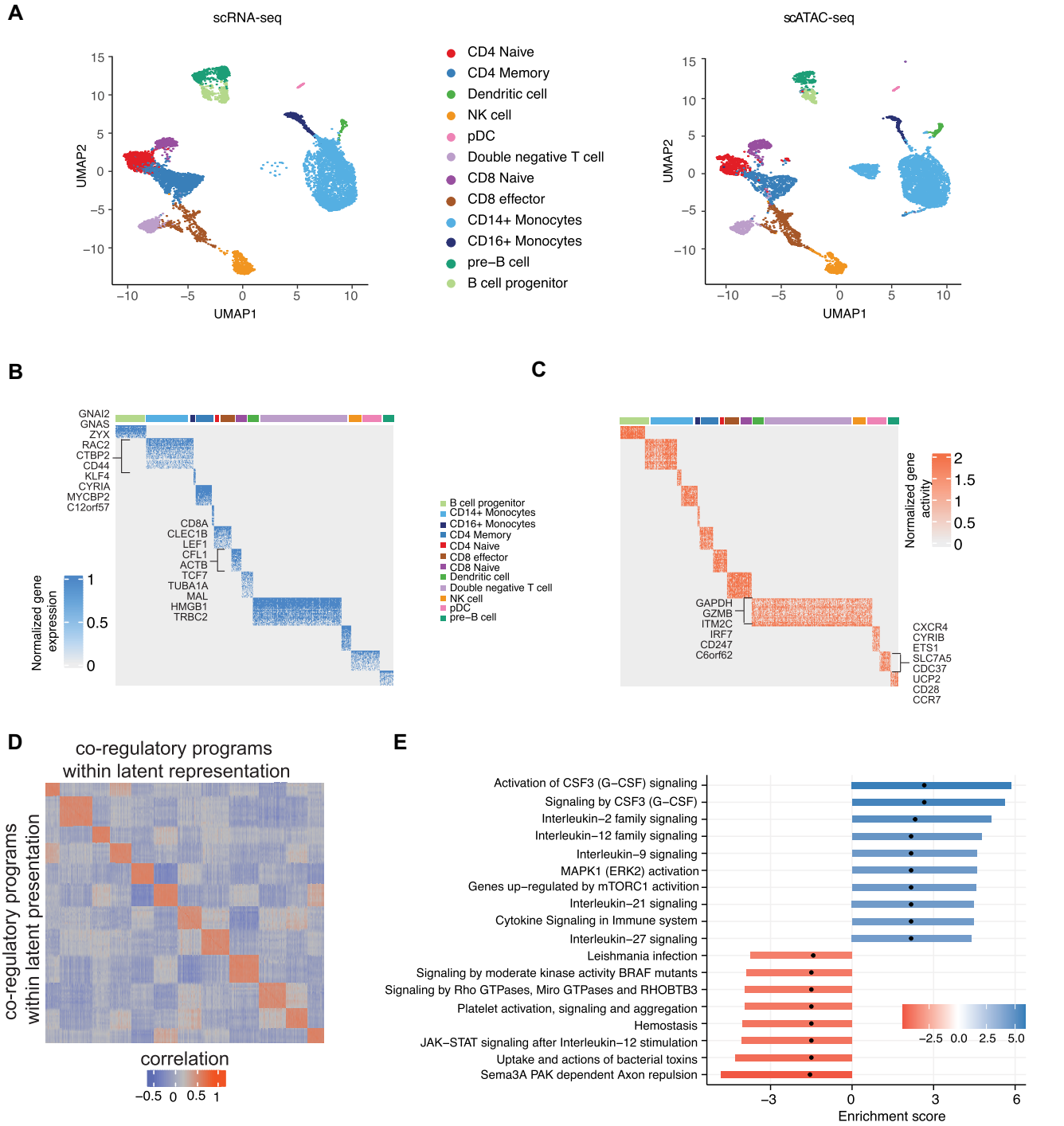

Figure 4. SMGR identifies cell-type specific co-regulatory programs. (**A**) UMAP visualization of the scRNA-seq and scATAC-seq data of peripheral blood mononuclear cell (PBMC). Different colors represent different cell types. (**B**) Heatmap shows the expression of genes within the top co-regulatory program for each cell type. Color scale represents the normalized gene expression in scRNA-seq data. (**C**) Heatmap shows the gene-level based chromatin activities within the top co-regulatory programs for each cell type. Color scale represents the normalized gene activities in scATAC-seq data. (**D**) Heatmap shows the association of gene expression and chromatin activities within a co-regulatory program in the latent representation. Color scale represents the Pearson correlation. (**E**) Enrichment analysis of the top co-regulatory programs of CD4 naïve T cells (red) and CD4 memory T cells (blue), respectively.

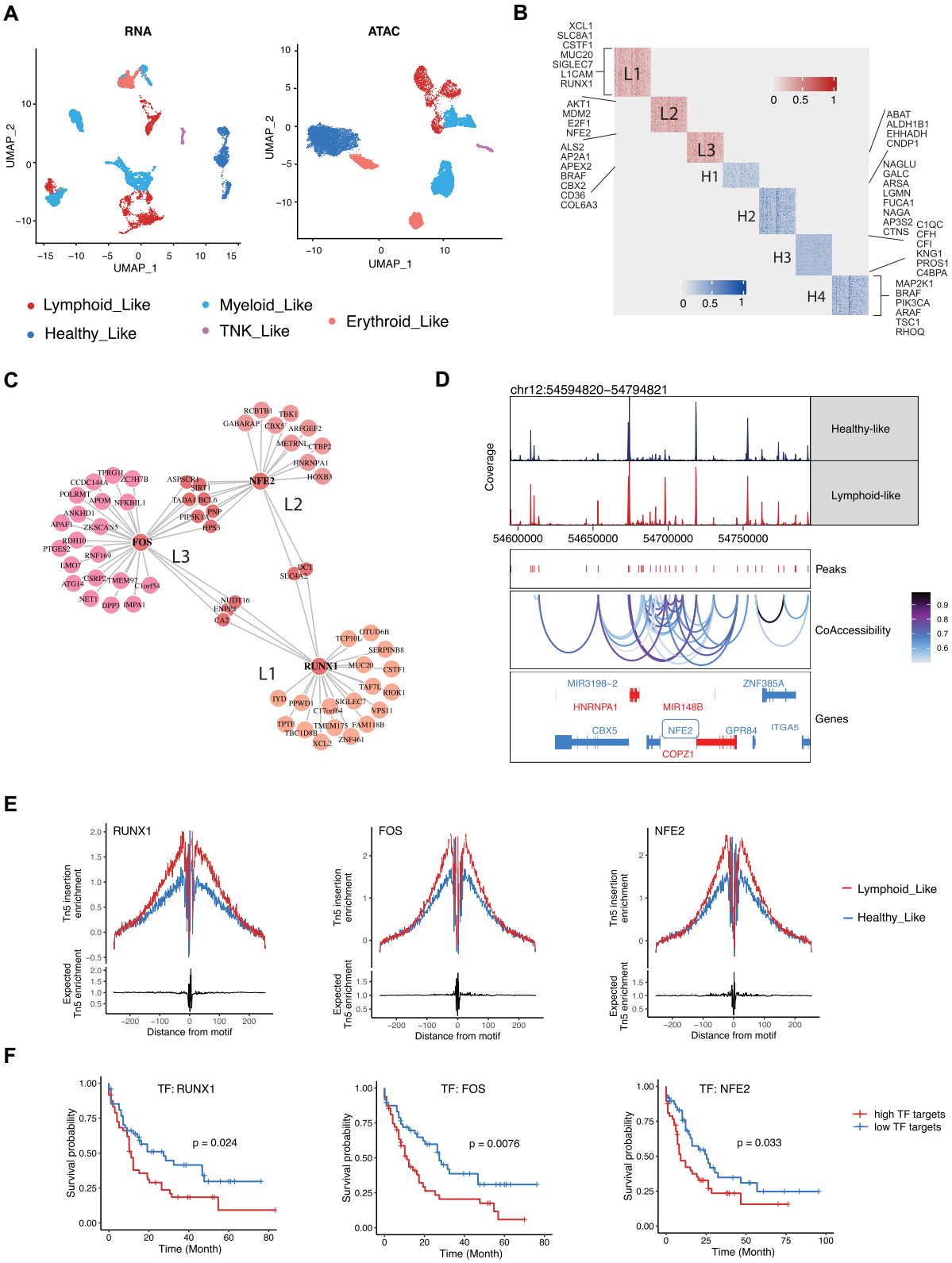

Figure 5. Application of SMGR to single-cell multi-omics data of MPAL cells. (**A**) UMAP visualization of the scRNA-seq and scATAC-seq data of mixed-phenotype acute leukemia. Different colors represent different cell types. (**B**) Heatmap shows the co-regulatory programs that differentiate between lymphoid-like cells and healthy-like. (**C**) The hub regulatory network of RUNX1, FOS, and NFE2 based on the co-regulatory programs of lymphoid-like cells. (**D**) scATAC-seq coverage reveals distinct chromatin accessibility between healthy-like cells and lymphoid-like cells. Peak-to-gene links are colored by Pearson correlation of the peak accessibility and gene expression. (**E**) Transcription factor footprints (average ATAC-seq around predicted binding sites) for three TFs for healthy-like cells and lymphoid-like cells respectively. (**F**) Kaplan–Meier curve for patients with AML from TCGA stratified by putative TF–target genes. *P*-value is calculated by log-rank test.

TF activity in healthy-like cells than the lymphoid-like cells. Interestingly, these TFs also presented lower gene activity in scRNA-seq data of healthy-like cells in contrast to lymphoid-like cells (Supplementary Figure S1), indicating these TFs are important regulatory signatures in MPAL. Further survival analysis using these TFs' targeted genes to stratify The Cancer Genome Atlas (TCGA) AML patients observed significantly decreased survival ($P = 0.024$ for RUNX1; $P = 0.0076$ for FOS; $P = 0.033$ for NFE2). Altogether, these results suggest that our identified regulatory programs are important that putatively upregulate the leukemic signaling cascade in MPAL.

DISCUSSION

In this study, we have developed a novel statistical method that is tailored and effective for identifying co-regulatory programs in single-cell multi-omics data including scRNA-seq and scATAC-seq. Applications of SMGR in the scRNA-seq and scATAC-seq data of MPAL identified the aberrant molecular features for MPAL development, which provide mechanistic insights into gene regulation at the cellular resolution. SMGR offers to investigate multiple regulatory layers that control cellular heterogeneity and complex biological mechanisms, which provides tremendous clinical value for identifying mechanisms, targets, and predictors for enhancing translational therapy.

One major distinction separating our method from published methods is that we focus on identifying co-regulatory features rather than cell groups. A common approach of revealing co-regulatory features is to integrate cells by removing cell-level batch effects first, and then identify the co-expressed or differentially expressed genes and peaks. However, our method provides a straightforward way to investigate the concordant genes and peak values by removing feature-level batch effects, which achieves better performance than available methods. On the other hand, though there are approaches studying coherent features based on bulk multi-omics data, none of these approaches can be directly applied to single cell data as they are not designed to account for the unique characteristics of single-cell multi-omics data as well as the technical noise and extrinsic variance among multiple single-cell samples. Collectively, SMGR is anticipated to be a very useful tool for identifying potential biomarkers and novel hypothesis for experimental validation.

Our SMGR method is provided as a freely available and open-source R package in GitHub, with detailed tutorials and workflows. We anticipate SMGR to unleash the power of emerging extensive single-cell multi-modal data and provide data-driven bioinformatics methods as well as open-source tools to the research community for better biological hypothesis testing and experimental design. Given the merits of SMGR, we also acknowledge that there are several limitations and caveats that warrant further study. First, though SMGR identified strongly coherent co-regulatory programs with genes from scRNA-seq and peaks from scATAC-seq data, it is still unknown of the hierarchical relations among these co-regulatory programs. Second, though SMGR is majorly designed for scRNA-seq and scATAC-seq data, we will further work on adapting our method to include more omics layers such as single-cell methylation or single-cell proteomics profiles. Third, we have used the original cellular annotations provided by the published data, which can potentially affect the performance of SMGR. We recommend use high-quality cell annotations of scRNA-seq and scATAC-seq data for applications, thus to improve the quality of the identified co-regulatory programs.

CONCLUSION

We have developed a novel method, Single-cell Multi-omics Gene co-Regulatory algorithm (SMGR), to detect coherent functional regulatory program and target genes from the joint scRNA-seq and scATAC-seq data. SMGR demonstrates high accuracy and robust performance in both simulation data and experimental single-cell multi-omics data. SMGR is available as a ready-to-use open-source software for revealing regulatory elements and mechanisms in complex diseases.

DATA AVAILABILITY

Simulation data. Based on the data characteristics, we have used the zero-inflated negative binomial distribution to generate simulation data. For simulation data, to obtain the ground truth of co-regulatory programs, we generate the synthetic scRNA-seq data with N clusters of genes and the synthetic scATAC-seq data with N clusters of peaks. We choose N as 3. For a certain gene cluster and its corresponding peak cluster, we generate the gene counts and peak counts across cells by the zero-inflated negative binomial distribution $ZINB(\rho, u, \varphi)$ with same parameters. Genes and peaks of different clusters are generated by zero-inflated negative binomial distributions with different parameters. Specifically, u, φ are chosen randomly between 1 and 5. In this way, we build the ground truth of co-regulatory programs, where genes with similar expressions and peak activities (generated by same distributions) should be captured as one regulatory program. In Figure 2A, we use 10 simulation datasets, with each dataset consisting of simulated scRNA-seq and scATAC-seq.

Real experimental data. (i) 14 benchmarking scRNA-seq datasets generated using both droplet and plate-based protocols are downloaded from Tian *et al.* (40). (ii) *PBMC data*. The PBMC scRNA-seq and scATAC-seq datasets are downloaded from 10× Genomics website through https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k. (iii) *MPAL data*. The MPAL scRNA-seq and scATAC-seq datasets are downloaded from Gene Expression Omnibus (GEO) with the accession code GSE139369.

All the functions mentioned above are implemented as a R package, which is accessible at https://github.com/QSong-github/SMGR.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

The authors acknowledge the DEMON high performance computing (HPC) cluster, the Texas Advanced Computing Center (TACC) at The University of Texas at Austin (http://www.tacc.utexas.edu), and the Extreme Science and Engineering Discovery Environment (XSEDE, which is supported by National Science Foundation grant number ACI-1548562) for providing HPC resources that have contributed to the research results reported within this paper.

FUNDING

Q.S. is supported in part by the Bioinformatics Shared Resources under the NCI Cancer Center Support Grant to the Comprehensive Cancer Center of Wake Forest University Health Sciences [P30CA012197]; X.Z. is supported by the NIH grant [NIH R01 HL132035]; J.S. was partially financially supported by the Indiana University Precision Health Initiative and by the Indiana University Melvin and Bren Simon Comprehensive Cancer Center Support Grant from the National Cancer Institute [P30 CA 082709].
Conflict of interest statement. None declared.

REFERENCES

1. Clark,S.J., Argelaguet,R., Kapourani,C.-A., Stubbs,T.M., Lee,H.J., Alda-Catalinas,C., Krueger,F., Sanguinetti,G., Kelsey,G., Marioni,J.C. *et al.* (2018) scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.*, **9**, 781.
2. Cusanovich,D.A., Daza,R., Adey,A., Pliner,H.A., Christiansen,L., Gunderson,K.L., Steemers,F.J., Trapnell,C. and Shendure,J. (2015) Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, **348**, 910–914.
3. Baron,M., Veres,A., Wolock,S.L., Faust,A.L., Gaujoux,R., Vetere,A., Ryu,J.H., Wagner,B.K., Shen-Orr,S.S., Klein,A.M. *et al.* (2016) A single-cell transcriptomic map of the human and mouse pancreas reveals Inter- and Intra-cell population structure. *Cell Syst.*, **3**, 346–360.
4. Puram,S.V., Tirosh,I., Parikh,A.S., Patel,A.P., Yizhak,K., Gillespie,S., Rodman,C., Luo,C.L., Mroz,E.A., Emerick,K.S. *et al.* (2017) Single-Cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, **171**, 1611–1624.
5. Athanasiadis,E.I., Botthof,J.G., Andres,H., Ferreira,L., Lio,P. and Cvejic,A. (2017) Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis. *Nat. Commun.*, **8**, 2045.
6. Welch,J.D., Kozareva,V., Ferreira,A., Vanderburg,C., Martin,C. and Macosko,E.Z. (2019) Single-Cell Multi-omic integration compares and contrasts features of brain cell identity. *Cell*, **177**, 1873–1887.
7. Nativio,R., Lan,Y., Donahue,G., Sidoli,S., Berson,A., Srinivasan,A.R., Shcherbakova,O., Amlie-Wolf,A., Nie,J. and Cui,X. (2020) An integrated multi-omics approach identifies epigenetic alterations associated with alzheimer's disease. *Nat. Genet.*, **52**, 1024–1035.
8. Bian,S., Hou,Y., Zhou,X., Li,X., Yong,J., Wang,Y., Wang,W., Yan,J., Hu,B., Guo,H. *et al.* (2018) Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science*, **362**, 1060–1063.
9. Granja,J.M., Klemm,S., McGinnis,L.M., Kathiria,A.S., Mezger,A., Corces,M.R., Parks,B., Gars,E., Liedtke,M., Zheng,G.X.Y. *et al.* (2019) Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.*, **37**, 1458–1465.
10. Welch,J.D., Kozareva,V., Ferreira,A., Vanderburg,C., Martin,C. and Macosko,E.Z. (2019) Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, **177**, 1873–1887.
11. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.
12. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck,W.M. 3rd, Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.
13. Barkas,N., Petukhov,V., Nikolaeva,D., Lozinsky,Y., Demharter,S., Khodosevich,K. and Kharchenko,P.V. (2019) Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods*, **16**, 695–698.
14. Haghverdi,L., Lun,A.T., Morgan,M.D. and Marioni,J.C. (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, **36**, 421–427.
15. Lun,L.A.T., Bach,K. and Marioni,J.C. (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, **17**, 75.
16. Vieth,B., Ziegenhain,C., Parekh,S., Enard,W. and Hellmann,I. (2017) powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*, **33**, 3486–3488.
17. Grün,D., Kester,L. and van Oudenaarden,A. (2014) Validation of noise models for single-cell transcriptomics. *Nat. Methods*, **11**, 637–640.
18. Lun,A.T. and Marioni,J.C. (2017) Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics*, **18**, 451–464.
19. Cao,J., Packer,J.S., Ramani,V., Cusanovich,D.A., Huynh,C., Daza,R., Qiu,X., Lee,C., Furlan,S.N. and Steemers,F.J. (2017) Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, **357**, 661–667.
20. Rosenberg,A.B., Roco,C.M., Muscat,R.A., Kuchina,A., Sample,P., Yao,Z., Graybuck,L.T., Peeler,D.J., Mukherjee,S. and Chen,W. (2018) Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, **360**, 176–182.
21. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N. and Martersteck,E.M. (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.
22. Stuart,T., Srivastava,A., Madad,S., Lareau,C.A. and Satija,R. (2021) Single-cell chromatin state analysis with signac. *Nat. Methods*, **18**, 1333–1341.
23. Brennecke,P., Anders,S., Kim,J.K., Kołodziejczyk,A.A., Zhang,X., Proserpio,V., Baying,B., Benes,V., Teichmann,S.A., Marioni,J.C. *et al.* (2013) Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods*, **10**, 1093–1095.
24. McCarthy,D.J., Chen,Y. and Smyth,G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.
25. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
26. Hafemeister,C. and Satija,R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.*, **20**, 296.
27. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
28. Zeileis,A., Kleiber,C. and Jackman,S. (2008) Regression models for count data in R. *J. Stat. Softw.*, **27**, 1–25.
29. Chib,S. and Greenberg,E. (1995) Understanding the metropolis-hastings algorithm. *Am Stat*, **49**, 327–335.
30. Jansen,C., Ramirez,R.N., El-Ali,N.C., Gomez-Cabrero,D., Tegner,J., Merkenschlager,M., Conesa,A. and Mortazavi,A. (2019) Building gene regulatory networks from scATAC-seq and scRNA-seq using linked self organizing maps. *PLoS Comput. Biol.*, **15**, e1006555.
31. Aibar,S., González-Blas,C.B., Moerman,T., Huynh-Thu,V.A., Imrichova,H., Hulselmans,G., Rambow,F., Marine,J.-C., Geurts,P., Aerts,J. *et al.* (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*, **14**, 1083–1086.
32. Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.
33. Davies,D.L. and Bouldin,D.W. (1979) A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, **1**, 224–227.
34. Dunn,J.C. (1974) Well-separated clusters and optimal fuzzy partitions. *J. Cybern.*, **4**, 95–104.
35. Caliński,T. and Harabasz,J. (1974) A dendrite method for cluster analysis. *Commun. Stat*, **3**, 1–27.

36. Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.
37. Huynh-Thu,V.A., Irrthum,A., Wehenkel,L. and Geurts,P. (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**, e12776.
38. Greenfield,A., Madar,A., Ostrer,H. and Bonneau,R. (2010) DREAM4: combining genetic and dynamic information to identify biological networks and dynamical models. *PloS One*, **5**, e13397.
39. Marbach,D., Costello,J.C., Küffner,R., Vega,N.M., Prill,R.J., Camacho,D.M., Allison,K.R., Kellis,M., Collins,J.J. and Stolovitzky,G. (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.
40. Tian,L., Dong,X., Freytag,S., Lê Cao,K.-A., Su,S., JalalAbadi,A., Amann-Zalcenstein,D., Weber,T.S., Seidi,A. and Jabbari,J.S. (2019) Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods*, **16**, 479–487.
41. Abu-Jamous,B. and Kelly,S. (2018) Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. *Genome Biol.*, **19**, 172.
42. Yin,W., Mendoza,L., Monzon-Sandoval,J., Urrutia,A.O. and Gutierrez,H. (2021) Emergence of co-expression in gene regulatory networks. *PloS One*, **16**, e0247671.
43. Wilkinson,A.C., Ballabio,E., Geng,H., North,P., Tapia,M., Kerry,J., Biswas,D., Roeder,R.G., Allis,C.D., Melnick,A. *et al.* (2013) RUNX1 is a key target in t(4;11) leukemias that contributes to gene activation through an AF4-MLL complex interaction. *Cell Rep.*, **3**, 116–127.
44. Noto,P.B., Sikorski,T.W., Zappacosta,F., Wagner,C.D., Montes de Oca,R., Szapacs,M.E., Annan,R.S., Liu,Y., McHugh,C.F., Mohammad,H.P. *et al.* (2020) Identification of hnRNP-A1 as a pharmacodynamic biomarker of type i PRMT inhibition in blood and tumor tissues. *Sci. Rep.*, **10**, 22155.
45. Li,S.-Q., Liu,J., Zhang,J., Wang,X.-L., Chen,D., Wang,Y., Xu,Y.-M., Huang,B., Lin,J., Li,J. *et al.* (2020) Transcriptome profiling reveals the high incidence of hnRNPA1 exon 8 inclusion in chronic myeloid leukemia. *J. Adv. Res.*, **24**, 301–310.
46. Müller,C.I., Luong,Q.T., Shih,L.Y., Jones,L.C., Desmond,J.C., Kawamata,N., Tcherniantchouk,O., Liu,Q., Ito,K., Osato,M. *et al.* (2008) Identification of marker genes including RUNX3 (AML2) that discriminate between different myeloproliferative neoplasms and normal individuals. *Leukemia*, **22**, 1773–1778.
47. Zhang,W., Ma,Q., Long,B., Sun,Z., Liu,L., Lin,D. and Zhao,M. (2021) Runt-related transcription factor 3 promotes acute myeloid leukemia progression. *Front. Oncol.*, **11**, 725336.
48. Carmichael,C.L., Wang,J., Nguyen,T., Kolawole,O., Benyoucef,A., De Mazière,C., Milne,A.R., Samuel,S., Gillinder,K., Hediyeh-Zadeh,S. *et al.* (2020) The EMT modulator SNAI1 contributes to AML pathogenesis via its interaction with LSD1. *Blood*, **136**, 957–973.
49. Wang,L., Zhao,H., Li,J., Xu,Y., Lan,Y., Yin,W., Liu,X., Yu,L., Lin,S., Du,M.Y. *et al.* (2020) Identifying functions and prognostic biomarkers of network motifs marked by diverse chromatin states in human cell lines. *Oncogene*, **39**, 677–689.
50. Hordijk,P.L. (2006) Regulation of NADPH oxidases: the role of rac proteins. *Circ Res.*, **98**, 453–462.
51. Zhang,G., Zhang,H., Liu,Y., He,Y., Wang,W., Du,Y., Yang,C. and Gao,F. (2014) CD44 clustering is involved in monocyte differentiation. *Acta Biochim. Biophys. Sin. (Shanghai)*, **46**, 540–547.
52. Feinberg,M.W., Wara,A.K., Cao,Z., Lebedeva,M.A., Rosenbauer,F., Iwasaki,H., Hirai,H., Katz,J.P., Haspel,R.L., Gray,S. *et al.* (2007) The Kruppel-like factor KLF4 is a critical regulator of monocyte differentiation. *EMBO J.*, **26**, 4138–4148.
53. Messmer,D., Yang,H., Telusma,G., Knoll,F., Li,J., Messmer,B., Tracey,K.J. and Chiorazzi,N. (2004) High mobility group box protein 1: an endogenous signal for dendritic cell maturation and Th1 polarization. *J. Immunol.*, **173**, 307–313.
54. Mondragón,L., Mhaidly,R., De Donatis,G.M., Tosolini,M., Dao,P., Martin,A.R., Pons,C., Chiche,J., Jacquin,M., Imbert,V. *et al.* (2019) GAPDH overexpression in the t cell lineage promotes angioimmunoblastic t cell lymphoma through an NF-κB-Dependent mechanism. *Cancer Cell*, **36**, 268–287.
55. Ye,W., Zhou,Y., Xu,B., Zhu,D., Rui,X., Xu,M., Shi,L., Zhang,D. and Jiang,J. (2019) CD247 expression is associated with differentiation and classification in ovarian cancer. *Medicine*, **98**, e18407.
56. Gallego,C., Vétillard,M., Calmette,J., Roriz,M., Marin-Esteban,V., Evrard,M., Aknin,M.-L., Pionnier,N., Lefrançois,M., Mercier-Nomé,F. *et al.* (2021) CXCR4 signaling controls dendritic cell location and activation at steady state and in inflammation. *Blood*, **137**, 2770–2784.
57. Yang,L., Feng,Y., Wang,S., Jiang,S., Tao,L., Li,J. and Wang,X. (2021) Siglec-7 is an indicator of natural killer cell function in acute myeloid leukemia. *Int. Immunopharmacol.*, **99**, 107965.
58. Huang,S., Zhang,B., Fan,W., Zhao,Q., Yang,L., Xin,W. and Fu,D. (2019) Identification of prognostic genes in the acute myeloid leukemia microenvironment. *Aging (Albany NY)*, **11**, 10557–10580.
59. Khurana,A. and Shafer,D.A. (2019) MDM2 antagonists as a novel treatment option for acute myeloid leukemia: perspectives on the therapeutic potential of idasanutlin (RG7388). *Onco Targets Ther.*, **12**, 2903–2910.
60. Zhang,T., Yang,J., Vaikari,V.P., Beckford,J.S., Wu,S., Akhtari,M. and Alachkar,H. (2020) Apolipoprotein C2 - CD36 Promotes leukemia growth and presents a targetable axis in acute myeloid leukemia. *Blood Cancer Discov.*, **1**, 198–213.
61. Prieto,C., Nguyen,D., Vu,L.P., Perez,A., Gourkanti,S., Schurer,A., Chou,T., Chow,A., Taggart,J., Barlowe,T.S. *et al.* (2018) RNA binding protein rbmx is required in acute myeloid leukemia by regulating the transcriptional activity of the heterochromatin protein HP1α. *Blood*, **132**, 883–883.