Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations

Zhana Duren[a,b,1], Xi Chen[a,b,1], Mahdi Zamanighomi[a,b,c,1], Wanwen Zeng[a,b,d], Ansuman T. Satpathy[c], Howard Y. Chang[c], Yong Wang[e,f], and Wing Hung Wong[a,b,c,2]

[a]Department of Statistics, Stanford University, Stanford, CA 94305; [b]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; [c]Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA 94305; [d]Ministry of Education Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic & Systems Biology, Department of Automation, Tsinghua University, 100084 Beijing, China; [e]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 100080 Beijing, China; and [f]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, 650223 Kunming, China

Contributed by Wing Hung Wong, June 14, 2018 (sent for review April 4, 2018; reviewed by Andrew D. Smith and Nancy R. Zhang)

When different types of functional genomics data are generated on single cells from different samples of cells from the same heterogeneous population, the clustering of cells in the different samples should be coupled. We formulate this "coupled clustering" problem as an optimization problem and propose the method of coupled nonnegative matrix factorizations (coupled NMF) for its solution. The method is illustrated by the integrative analysis of single-cell RNA-sequencing (RNA-seq) and single-cell ATAC-sequencing (ATAC-seq) data.

coupled clustering | NMF | single-cell genomic data

Biological samples of interest in clinical or experimental studies are often heterogeneous mixtures; i.e., a sample may consist of many subpopulations of cells with distinct cellular states. To resolve the heterogeneity and to characterize the constituent subpopulations, it is necessary to generate functional genomic data at the single-cell level. An exciting recent development in genomics technology has been the emergence of methods for single-cell (sc) measurements; for example, scRNA sequencing (scRNA-seq) (1) enables transcription profiling, scATAC sequencing (scATAC-seq) (2) identifies accessible chromatin regions, and sc-bisulfite sequencing (3) measures DNA methylation, all at the single-cell level.

Often, the first step in the analysis of single-cell data is clustering, that is, to classify cells into the constituent subpopulations. Clustering methods for scRNA-seq data are discussed in refs. 4 and 5, and clustering of scATAC-seq data is described in ref. 6. Existing methods, however, do not address the increasingly common situation where two or more types of sc-genomics experiments are performed on different subsamples from the same cell population. For example, Fig. 1*A* depicts the situation when one subsample is analyzed by scRNA-seq while another is analyzed by scATAC-seq. Although the clustering methods developed for scRNA-seq and scATAC-seq were each shown to be capable of identifying distinct cell types, the association of gene expression changes to chromatin accessibility dynamics better defines cell types and lineages, especially in complex tissues (7, 8). To connect these two assays, one might suggest to separately cluster each data type, followed by an integration afterward. However, such approaches can be problematic because scATAC-seq and scRNA-seq data do not always possess a similar power for detection of cell types (9). Furthermore, clusters may be data type specific due to technical noise. So it is advantageous to systematically couple the two clustering processes in such a way that the clustering of the cells in the scRNA-seq sample can also make use of information from the scATAC-seq sample, and vice versa (10). In this paper, we formulate this "coupled clustering" problem as an optimization problem and introduce a method, named coupled nonnegative matrix factorizations (coupled NMF), for its solution.

Approach

Coupled NMF. We first introduce our approach in general terms. Let O be a p_1 by n_1 matrix representing data on p_1 features for n_1 units in the first sample; then a "soft" clustering of the units in this sample can be obtained from a nonnegative factorization $O = W_1H_1$ as follows: The ith column of W_1 gives the mean vector for the ith cluster of units, while the jth column of H_1 gives the assignment weights of the jth unit to the different clusters. Similarly, clustering of the second sample can be obtained from the factorization $E = W_2H_2$, where E is the p_2 by n_2 matrix of data on p_2 features (which are different from the features measured in the first sample) for n_2 units in the second sample. To couple two matrix factorizations, we introduce a term $tr(W_2^TAW_1)$, where A is a "coupling matrix." The construction of A is application specific but depends on the assumption that, based on scientific understanding or prior data, it is possible to identify a subset of features in one sample that are linearly predictable from the features measured in the other sample. In such a situation, we can take A to be the matrix representation of the linear prediction operator.

Significance

Biological samples are often heterogeneous mixtures of different types of cells. Suppose we have two single-cell datasets, each providing information on a different cellular feature and generated on a different sample from this mixture. Then, the clustering of cells in the two samples should be coupled as both clusterings are reflecting the underlying cell types in the same mixture. This "coupled clustering" problem is a new problem not covered by existing clustering methods. In this paper, we develop an approach for its solution based on the coupling of two nonnegative matrix factorizations. The method should be useful for integrative single-cell genomics analysis tasks such as the joint analysis of single-cell RNA-sequencing and single-cell ATAC-sequencing data.

Author contributions: H.Y.C., Y.W., and W.H.W. designed research; Z.D., X.C., W.Z., and A.T.S. performed research; Z.D., M.Z., and W.H.W. analyzed data; and Z.D., M.Z., and W.H.W. wrote the paper.

Reviewers: A.D.S., University of Southern California; and N.R.Z., University of Pennsylvania.

The authors declare no conflict of interest.

Data deposition: The single- cell gene expression data and chromatin accessibility data of RA induction reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, https://www.ncbi.nlm.nih.gov/geo (accession nos. GSE115968 and GSE115970).

[1]Z.D., X.C., and M.Z. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: whwong@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1805681115/-/DCSupplemental.

Published online July 9, 2018.

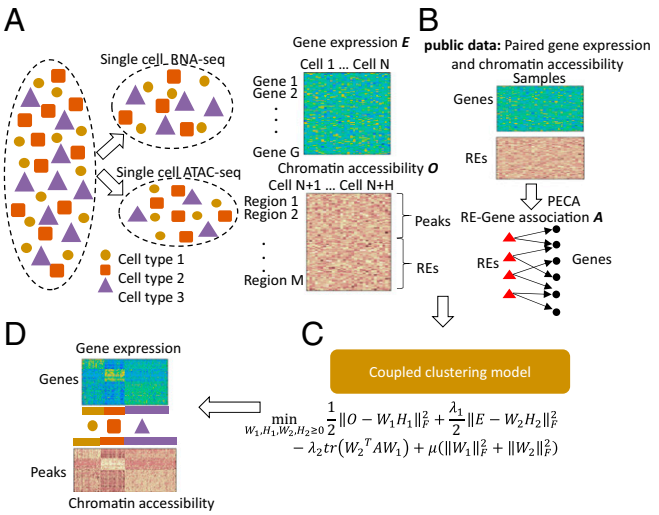


Fig. 1. Overview of the coupled-clustering method. (*A*) Single-cell gene expression and single-cell chromatin accessibility data. (*B*) Learning coupling matrix from public data. (*C*) Coupled clustering model. (*D*) Cluster-specific gene expression and chromatin accessibility.

Once the coupling is defined this way, we can obtain the factorizations of the two data matrices by solving the following optimization problem (Fig. 1*C*):

$$\min_{W_1,H_1,W_2,H_2\geq 0} \frac{1}{2}\|O-W_1H_1\|_F^2+\frac{\lambda_1}{2}\|E-W_2H_2\|_F^2-\lambda_2 tr(W_2^T A W_1) + \mu\left(\|W_1\|_F^2+\|W_2\|_F^2\right). \quad [1]$$

There are three tuning parameters: λ_1, λ_2, and μ. The first two terms are clustering the two samples. The third term is to induce the consistency of features from the second sample with linear transformed features from the first sample. The fourth term controls the growth of W_1 and W_2. After solving the optimization, the cluster profile and the cluster assignments for the kth cluster in the first sample can be obtained respectively from the kth column of W_1 and the kth row of H_1. Similarly, the clustering in the second sample can be obtained from W_2 and H_2 (Fig. 1*D*).

Application in Single-Cell Genomic Data. We apply the coupled NMF approach to cluster scRNA-seq and scATAC-seq data. In this application, O_{ij} denotes the degree of openness (i.e., accessibility) of the ith region in the jth cell (6). By a region we mean either a predefined regulatory element (RE) or a peak called from merged scATAC-seq data. From scRNA-seq data, we compute the data matrix E where E_{gh} denotes the expression level of the gth gene in the hth cell (11). Details are given in *Materials and Methods*, Data Processing. Note that the scATAC-seq and the scRNA-seq data are not measured in the same cell (Fig. 1*A*).

To get the coupling matrix A, we take advantage of our recent work on modeling paired gene expression and chromatin accessibility data (7) and use a diverse panel of cell lines with both expression and accessibility data to train a prediction model of gene expression from accessibility (see *Materials and Methods* for details). After fitting the model, we select a set of "well-predicted" genes (named gene set S) and use this set of genes' RE–gene associations to couple the two types of data. In this application, AW_1 gives the cluster-specific predictions of the expression of genes based on the cluster-specific accessibilities of REs, and hence the trace term enforces our expectation that the expression of genes should be consistent with the predictions based on accessibility of REs. As the coupling matrix A is noisy, we can refine the coupling iteratively, as follows, to get a better result. We assign single cells to clusters according to the assignment weights given by H_1 and H_2. After getting the cluster results, we choose cluster-specific genes based on scRNA-seq clustering. We restrict the rows of A and W_2 to the cluster-specific genes and get a submatrix of $\tilde{A}$ of A and $\widetilde{W_2}$ of W_2. Then we replace the A and W_2 in the trace term by the submatrices and recluster the cells by optimizing the objective function in Eq. **1**. We continue until the cluster assignments are not changed by further iterations.

Results

Results on Simulation Data. We first evaluate the performance of our method in a simulation study. Single-cell datasets are simulated by sampling reads from a bulk dataset (*Materials and Methods*). The bulk datasets used in our simulation study are from two very similar cell types from a hematopoietic differentiation process, namely common myeloid progenitor (CMP) and megakaryocyte erythroid progenitor (MEP) (12). For each of these two cell types, we first generated 100 scRNA-seq datasets and 100 scATAC-seq datasets (*Materials and Methods*).

To simulate a scRNA-seq dataset from a mixed population with two cell types, we simply mix the 200 scRNA-seq data from two cell lines together and treat them as a single scRNA-seq dataset. We apply k-means and NMF to cluster the mixed cells. We run k-means 50 times with different random initial values and choose the result that gives the minimum total sum of within-cluster distances. Similarly, we run NMF 50 times and choose the result that gives the minimum approximation error in the Frobenius norm. The results of all 50 runs on scRNA-seq and scATAC-seq data by k-means and NMF are shown in *SI Appendix*, Fig. S1. Finally, we perform coupled NMF clustering based on both the 200-cell mixture of the scRNA-seq sample and the 200-cell mixture of the scATAC-seq sample (*SI Appendix*, Fig. S3). The performance of the three clustering results (k-means on scRNA-seq only, NMF on scRNA-seq only, and coupled NMF on both scRNA-seq and scATAC-seq) is presented in Fig. 2*A*. A similar comparison on the clustering results of scATAC-seq is illustrated in Fig. 2*B*. The convergence of coupled NMF is given in *SI Appendix*, Fig. S2. It is seen that coupling leads to greatly improved results, reducing the assignment error rate by more than threefold over the other two methods (Fig. 2*C*).

Assessment of Prediction Model Before Coupling. We are interested in applying coupled NMF to analyze data generated from differentiation of a mouse embryonic stem cell, namely scRNA-seq and scATAC-seq at day 4 after retinoic acid (RA) treatment (*Materials and Methods*). Before analyzing the single-cell data, we want to assess whether the model learned from the diverse panel (i.e., matrix A) provides reliable predictive power to connect chromatin accessibility and gene expression in this biological context. We thus generated bulk RNA-seq and ATAC-seq at day 4 of the RA treatment (named RA day 4). Using the model trained on the diverse panel, we predicted the expression of genes in set S at RA day 4. *SI Appendix*, Fig. S3 shows the observed vs. predicted gene expressions. It is seen that genes in S were predicted with high accuracy ($R^2 = 0.75$, $r = 0.87$). This gives us confidence in using the model to initiate the coupling.

Results on Real Single-Cell Data. Next, we use coupled NMF to analyze RA day-4 scRNA-seq and scATAC-seq data. We first perform coupled NMF with $K = 2$ (i.e., two clusters) and then visualize the clustering result on Spearman correlation-based t-distributed stochastic neighbor embedding (t-SNE) plots. There are clearly two separated clusters in both t-SNE plots of scATAC-seq and scRNA-seq. Increasing the number of clusters to three, we can see three well-separated clusters in t-SNE plots. However, when K is increased to 4 or 5, the separation among clusters is no longer clear (Fig. 3*A* and *SI Appendix*, Fig. S4). We also calculate clustering stability based on the method in Brunet et al.

A

K-means (50 replicates)

RNA-seq	CMP	MEP
Cluster 1	59	36
Cluster 2	41	64

NMF (50 replicates)

RNA-seq	CMP	MEP
Cluster 1	74	24
Cluster 2	26	76

Coupled clustering

RNA-seq	CMP	MEP
Cluster 1	5	93
Cluster 2	95	7

B

K-means (50 replicates)

ATAC-seq	CMP	MEP
Cluster 1	1	14
Cluster 2	99	86

NMF (50 replicates)

ATAC-seq	CMP	MEP
Cluster 1	24	99
Cluster 2	76	1

Coupled clustering

ATAC-seq	CMP	MEP
Cluster 1	11	100
Cluster 2	89	0

C

	K-means	NMF	Coupled clustering
RNA-seq err	77	50	12
ATAC-seq err	87	25	11
Error rate	41.00%	18.75%	5.75%

Fig. 2. (*A*) Clustering results of *k*-means, NMF, and our coupled clustering on simulation scRNA-seq data of CMP and MEP. (*B*) Clustering results of *k*-means, NMF, and our coupled clustering on simulation scATAC-seq data of CMP and MEP. (*C*) Comparison of *k*-means, NMF, and coupled clustering on simulation data of CMP and MEP.

(13) for $K = 2$–5 (*SI Appendix*, Fig. S5). The results show that clustering results are stable when $K = 2$ or 3, while the results are not stable when K is increased to 4 or 5. Hence, we set $K = 3$ for the remaining of the analysis.

For each of the three clusters, we identify cluster-specific transcription factors (TFs) based on their expression from RNA-seq data and compare their motif activities in scATAC-seq data (Fig. 3). Here the motif activity of a TF in a cell reflects the enrichment level of the TF's motif in accessible REs in a scATAC-seq dataset (*Materials and Methods*). Fig. 3*B* shows the motif activities and expressions of some cluster-specific TFs on the t-SNE plots (Fig. 3*B* and *SI Appendix*, Fig. S6). Fig. 3*C* shows the heat maps of motif activities and expressions for a subset of cluster-specific TFs, namely those with expression transcripts per million (TPM) greater than 10 in at least 40 cells. It is seen that cluster-1–specific TFs (e.g., Ebf1, Lhx1, and Neurod1) have high motif activities in the corresponding cluster of the scATAC-seq sample. Similarly, cluster-2–specific TFs, Gata4, Foxa2, and Jun, have high motif activity in cluster 2 and cluster-3–specific TFs, Rfx4, Sox2, Sox9, Pou3f2, and Pou3f4, have high motif activity in cluster 3. This result shows that our method leads to highly consistent TF expression and TF motif activities within each of the inferred constituent subpopulations.

To further assess the coupled NMF results, we select cluster-specific genes from RNA-seq data and cluster-specific peaks from ATAC-seq data. We test whether the cluster-specific genes from scRNA-seq data are significantly overlapped with the genes with nearby (100 kb) cluster-specific ATAC-seq peaks by performing Fisher's exact test based on the overlap of two sets of genes (*SI Appendix*, Fig. S7). Fig. 3*D* gives the *P* values for all possible pairings of the RNA-seq clusters with ATAC-seq clusters. It is seen that the pairings identified by coupled NMF indeed gave dramatically more significant *P* values and higher fold changes than the other possible pairings.

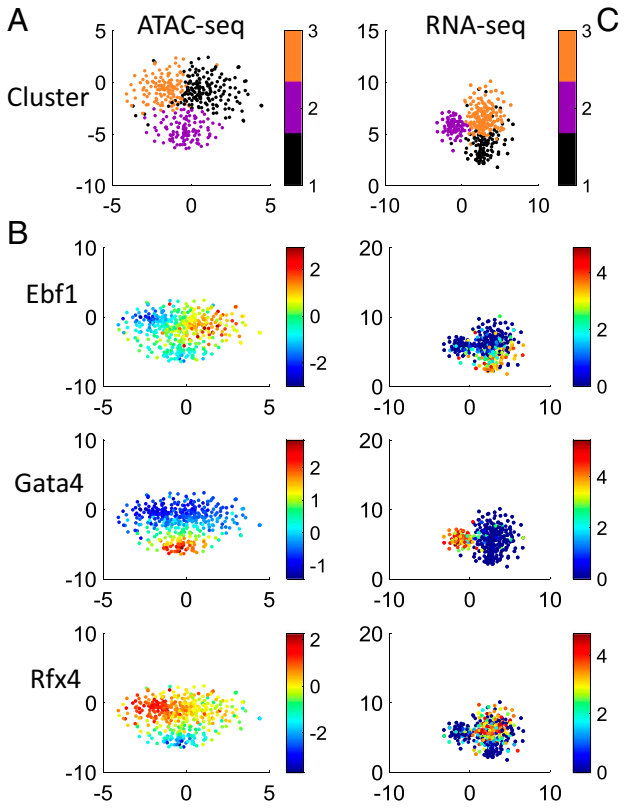
Coupled Clustering of Single Cells Sheds Light on Stem Cell Differentiation. The cluster-specific gene expression profiles and chromatin accessibility profiles provided by our method can provide useful insight into the constituent subpopulations. First, we use cluster-specific peaks from scATAC-seq data to annotate the clusters. We collect previously determined enhancers in mouse tissues at seven developmental stages from 11.5 d postconception until birth (14). Fig. 4 *A*–*C* shows the degree of overlap of our cluster-specific peaks with these developmental enhancers for different tissues and at different developmental stages. The number represents 10,000 times the Jaccard index (intersection over union) and NA indicates that enhancer data for that tissue in that stage are not available. The results show that cluster-1–specific peaks are enriched in forebrain and midbrain enhancers at E12.5 and E13.5. Cluster-2–specific peaks are enriched in heart enhancers at E15.5 and E16.5. Cluster-3–specific peaks are enriched in forebrain enhancers from E12.5 to E16.5 and also in midbrain, hindbrain, and neural tube. In addition, we also collect experimentally validated tissue-specific enhancers from the VISTA database (https://enhancer.lbl.gov/) and overlap them to cluster-specific peaks. Fig. 4*D* shows the percentage of tissue-specific VISTA enhancers overlapped to cluster-specific peaks. Only those tissues with at least one enhancer overlapping with the cluster-specific peaks are shown. Enhancers from nervous system-associated tissue (neural tube, cranial nerve, hind brain, midbrain, forebrain, trigeminal V, dorsal root ganglion, eye, nose) have overlap with cluster-specific peaks from cluster 1 and cluster 3. Cluster-2–specific peaks are enriched in blood vessel enhancers and heart enhancers. These results suggest that clusters 1 and 3 may be related to nervous system tissues and cluster 2 may be related to heart tissue.

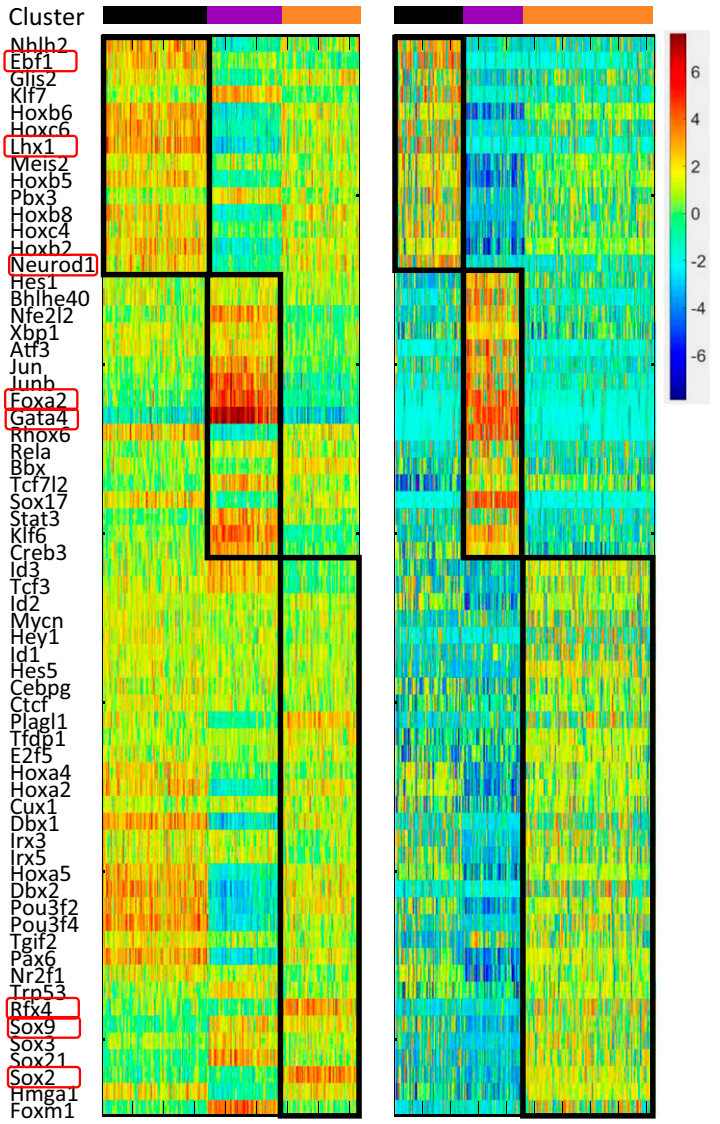
Next, we analyzed cluster-specific genes from scRNA-seq data. Fig. 4*E* presents the most enriched gene ontology (GO) terms, their *P* values, and fold changes in each cluster. The results show that cluster 2 is enriched in blood vessel development and cardiovascular system development, while clusters 1 and 3 are enriched in nervous system-associated terms. The results from scRNA-seq–based annotation are consistent with the results from scATAC-seq–based analysis. Although clusters 1 and 3 are nervous system-associated clusters, there are interesting differences. Cluster 1 is more enriched in axon guidance and neuron projection guidance, while cluster 3 is more enriched in brain development and oligodendrocytes differentiation. It seems cluster 1 is related to neuron-specific development and cluster 3 is more related to general nervous system development. Overall our results suggest that the RA-induced stem cell at day 4 is a mixture of cells related to neuron, cardiovascular system, and nervous system. These results are largely consistent with previous studies (15, 16).

We can construct cluster-specific gene regulatory networks as graphs with directed edges from the cluster-specific peaks to the cluster-specific genes that are within 100 kb distance and directed edges from cluster-specific TFs to cluster-specific peaks containing significant matches to the corresponding motifs. These cluster-specific subnetworks are presented in *SI Appendix*, Fig. S8. It is seen that Klf7, Ebf1, Sox11, and Nhlh1 are playing an important role in the network for cluster 1; Gata4, Gata6, Sox17, Foxa2, Ap1 complex, and Tead family are important in cluster 2; and Rarb, Nr2f1, Rfx4, Sox2, Sox9, Sox21, Pou3f2, Pou3f3, and Pou3f4 are important in cluster 3.

Discussion

In this paper, we proposed a coupled clustering method and applied it to single-cell genomic data. We emphasize that the measurement of multiple data types in the same cell is technically challenging due to the complex cellular reactions. Our method utilizes external information to integrate gene expression and chromatin accessibility that are not measured on the same cell. In the simulation study, we showed that the coupled NMF outperforms clustering results derived from just one data type. Moreover, we showed that our method identifies important peaks and genes that characterize cellular heterogeneity in the





D

	ATAC C1	ATAC C2	ATAC C3
RNA C1	**7.1E-23 (4.07)**	0.3693(0.82)	0.6532(1.10)
RNA C2	0.2974(0.90)	**9.1E-102 (2.30)**	0.5468(1.04)
RNA C3	0.0042(1.39)	0.2538	**4.1E-24 (2.01)**

Fig. 3. (*A*) t-SNE plot of scRNA-seq data (*Right*) and scATAC-seq data (*Left*) from RA day 4. Different colors represent clustering assignment from the coupled-clustering method. (*B*) Same t-SNE plots as in *A*. Different colors represent cluster-specific TFs' (Ebf1, Gata4, and Rfx4) gene expression *Z* score and motif activity *Z* score. (*C*) Comparison of cluster-specific TFs' expression *Z* score with motif activity *Z* score at the cluster level. (*D*) Overlap of cluster-specific peaks nearby genes with cluster-specific genes. The values represent Fisher's exact test *P* value and fold change.

context of the RA-induced stem cell. The proposed method enables a systematic mapping of peaks to genes, informative for downstream analysis such as inferring gene regulatory networks at the single-cell level.

As far as we know, coupled clustering is a problem different from other complex clustering tasks such as biclustering or multiview clustering. Biclustering (17–19), also called block clustering or coclustering, has been used widely to cluster subjects and cluster genes simultaneously based on a p by n data matrix of expression measurements on p genes for n subjects. The same data matrix is used in the clustering in gene space as well as the clustering in subject space. In contrast, two different data matrices are used in coupled clustering of two separate samples. In multiview clustering (20), the set of features measured on each subject can be divided into two independent subsets; for example, one of them may represent gene expression measurements while the other represent accessibility measurements. The important difference between multiview clustering and coupled clustering is that in the former setting all features are measured on each subject, whereas in the latter one only one of the subsets can be measured on any subject. Clearly, coupled clustering is a more challenging task and requires external information such as subject domain knowledge or prior data to initialize the coupling.

Materials and Methods

Construction of Data Matrices. From scATAC-seq data, we compute a data matrix O, where O_{ij} denotes the degree of openness (i.e., read count) of the ith region in the jth cell (6). By region we mean union of predefined REs and peaks. From scRNA-seq data, we compute the data matrix E where E_{gh} denotes the expression level of the gth gene in the hth cell (11). Details are given in *Materials and Methods*, *Data Processing*. Note that the scATAC-seq and the scRNA-seq data are not measured in the same cell (Fig. 1*A*).

Construction of Coupling Matrix. Our approach to the contraction of A is to look for a subset of genes whose expression is highly predictable from chromatin accessibility of REs. To do this, we take advantage of our recent work on modeling paired gene expression and chromatin accessibility data (on bulk samples) across diverse cellular contexts (7). From the paired expression and chromatin accessibility (PECA) model in that work, for each gene g, we can extract a set S_g of REs that

A

Cluster 1	facial prominence	forebrain	midbrain	hindbrain	neural tube	heart	intestine	kidney	limb	liver	lung	stomach
E11.5	9.29	5.67	13.82	8.87	13.48	7.25	NA	NA	7.62	6.87	NA	NA
E12.5	11.02	24.10	20.59	15.79	9.53	6.55	NA	NA	11.68	14.91	NA	NA
E13.5	8.45	25.25	21.73	16.89	19.76	9.92	NA	NA	5.70	12.54	NA	NA
E14.5	8.35	11.75	12.19	14.74	8.70	5.12	4.59	3.99	8.12	13.15	5.95	4.63
E15.5	13.43	17.85	14.90	15.84	16.25	10.93	10.91	8.89	17.02	14.51	8.01	10.02
E16.5	NA	19.03	11.93	12.91	NA	10.92	9.52	7.37	NA	11.57	9.21	7.20
P0	NA	8.00	5.91	6.85	NA	5.92	3.58	5.44	NA	8.86	5.54	8.08

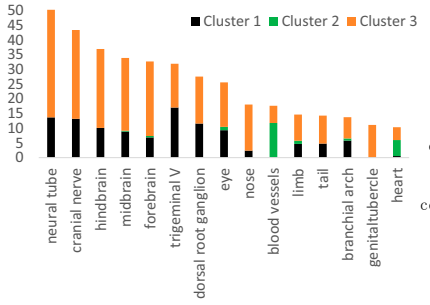
B

Cluster 2	facial prominence	forebrain	midbrain	hindbrain	neural tube	heart	intestine	kidney	limb	liver	lung	stomach
E11.5	5.65	4.01	7.58	2.95	3.41	18.27	NA	NA	4.36	13.29	NA	NA
E12.5	4.88	9.68	9.01	8.08	3.46	12.43	NA	NA	9.03	17.47	NA	NA
E13.5	6.83	4.92	16.07	14.36	14.45	20.23	NA	NA	5.89	16.94	NA	NA
E14.5	6.67	4.12	7.90	6.18	4.18	19.26	11.93	5.17	6.63	16.41	9.31	6.89
E15.5	12.97	11.90	13.40	12.38	11.68	21.25	10.30	21.45	11.15	12.14	17.62	12.89
E16.5	NA	11.47	10.89	11.88	NA	22.52	18.14	14.38	NA	14.87	17.32	17.04
P0	NA	5.58	5.44	7.55	NA	15.93	4.95	6.71	NA	8.71	11.06	18.27

C

Cluster 3	facial prominence	forebrain	midbrain	hindbrain	neural tube	heart	intestine	kidney	limb	liver	lung	stomach
E11.5	8.23	20.05	20.44	14.82	15.18	7.69	NA	NA	2.73	1.78	NA	NA
E12.5	13.87	28.56	25.45	18.64	12.64	9.33	NA	NA	13.37	4.10	NA	NA
E13.5	14.50	30.76	26.94	17.08	24.07	9.21	NA	NA	14.25	4.95	NA	NA
E14.5	16.34	31.51	21.42	18.97	12.02	8.41	4.47	5.27	14.33	4.45	3.62	6.79
E15.5	14.68	31.78	24.58	21.83	19.95	11.95	2.75	9.09	15.04	3.95	7.64	4.93
E16.5	NA	32.39	19.15	25.47	NA	7.89	5.42	9.00	NA	4.09	6.09	2.26
P0	NA	12.52	5.31	6.12	NA	2.37	0.00	1.62	NA	2.40	3.62	0.89

D



E

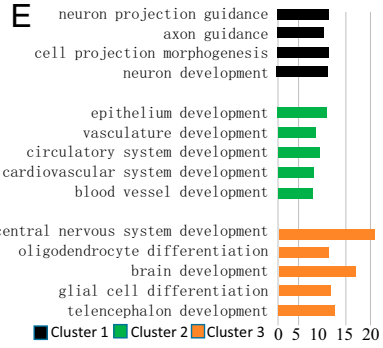


Fig. 4. (*A*–*C*) Similarity of cluster-specific peaks with enhancers of 12 tissues' seven developmental stages. The numbers represent 10,000× Jaccard index and NA indicates enhancer data of that tissue in that stage are not available. (*D*) Percentage of VISTA enhancer that overlapped with cluster-specific peaks. (*E*) GO enrichment of cluster-specific genes.

regulate that gene. We consider the regression model of target gene (TG) expression (denoted as E_g) on its REs' accessibility (denoted as O_i):

$$E_g = \alpha_{g0} + \sum_{i \in S_g} \alpha_{gi} O_i. \quad [2]$$

We estimate the parameter α_g by fitting the penalized least-squares problem (Eq. **3**) based on expression and accessibility data on a diverse panel of cell lines [56 cell lines in the case of mouse and 148 cell lines in the case of human (Dataset S1)],

$$\min_{\alpha_g} \frac{1}{2}\left\| E_g - \alpha_{g0} - \sum_{i \in S_g} \alpha_{gi} O_i \right\|_F^2 + \delta\left(\|\alpha_g\|_1 + \|\alpha_g\|_2^2 \right), \quad [3]$$

where δ is determined by fivefold cross-validation. After fitting the model, we select a set of well-predicted genes for which regression R^2 is greater than 0.8. In this way, we selected 5,966 well-predicted genes in mouse and selected 6,253 well-predicted genes in human. In coupling matrix $A = (\alpha_{gi})$, only the rows corresponding to the selected genes are nonzero.

Optimization Algorithm. We optimize the object function in Eq. **1** by a multiplicative update algorithm,

$$w_{ij}^1 \leftarrow w_{ij}^1 \frac{\left(OH_1^T + \frac{\lambda_2}{2} A^T \widetilde{W_2} \right)_{ij}}{\left(W_1 H_1 H_1^T + 2\mu W_1 \right)_{ij}}$$

$$w_{ij}^2 \leftarrow w_{ij}^2 \frac{\left(XH_2^T + \frac{\lambda_2}{2\lambda_1} A W_1 \right)_{ij}}{\left(W_2 H_2 H_2^T + 2\mu W_2 \right)_{ij}}$$

$$h_{ij}^1 \leftarrow h_{ij}^1 \frac{\left(W_1^T O \right)_{ij}}{\left(W_1^T W_1 H_1 \right)_{ij}}$$

$$h_{ij}^2 \leftarrow h_{ij}^2 \frac{\left(W_2^T E \right)_{ij}}{\left(W_2^T W_2 H_2 \right)_{ij}},$$

where w_{ij}^1 represents the element of the ith row and the jth column in matrix W_1, and the same representation is used in W_2, H_1, and H_2. We stop the iteration when the relative error is less than 0.0001.

Cluster-Specific Features. We apply a t test to define the cluster-specific genes and cluster-specific peaks, and the default P-value cutoff is 0.0001.

Evaluation of the Clustering Results. We evaluate the results in terms of consistency of true expression values and the predicted values. We calculated the $K \times K$ correlation matrix of AW_1 with W_2, which is denoted by R. We use the determinant of correlation matrix R to measure the consistency of true expression values with the predicted values. Higher determinant means higher diagonal of the matrix, which means higher correlation between matched clusters and lower correlation between unmatched clusters. When $K = 1$, the determinant simply reflects the correlation between true gene expression and predicted gene expression. When $K > 1$, the determinant will integrate information from both within-cluster correlations and between-cluster correlations. For example, suppose there are three true clusters in the population but in our clustering result one of the true clusters is randomly split into two subclusters. In this situation, the clustering result has four clusters and all four within-cluster correlations will be high; however, the determinant will be close to zero because the correlation vectors due to the two subclusters will be highly colinear. Thus, the use of the determinant instead of the product of within-cluster correlations will offer protection against overpartitioning.

Parameter Selection. We solve optimization problems $\min_{W_1,\ H_1 \geq 0} \|O - W_1 H_1\|_F^2$, $\min_{W_2,\ H_2 \geq 0} \|E - W_2 H_2\|_F^2$ by the alternating least-squares (ALS) algorithm with 50 different initializations using a Monte Carlo-type approach (21) and get the solutions W_{10}, H_{10}, W_{20}, H_{20}, which are used as initial solutions in our optimization problem. We choose parameter $\mu = \|O - W_{10} H_{10}\|_F^2 / (\|W_{10}\|_F^2 + \|W_{20}\|_F^2)$. Tuning parameters λ_1 and λ_2 are chosen from 0.001, 0.01, 0.1, 1, 10, 100, 1,000, and 10,000. The determinant of correlation matrix R can be used to select the tuning parameters. We choose the tuning parameters which give the highest determinant (*SI Appendix*, Fig. S9). The number of clusters K can be determined by a method similar to that in ref. 13.

TF Motif Activity. We use the software chromVAR (22) to calculate the TF motif activity on each single cell based on its scATAC-seq data.

Single-Cell Sample at RA Day 4. We generated a heterogeneous biological population of cells that arise from the same origin. Specifically, we used the hanging-drop technique to form embryonic bodies (EBs) from mouse embryonic stem cells (mESCs) and induced differentiation by RA treatment. After 4 d of induction, we sample cells for bulk RNA-seq and bulk ATACC-seq experiments for use in validating the coupling. To test the coupled-NMF clustering method, we also generated scATAC-seq and scRNA-seq on the RA day-4 population. After removing low read-count cells (3,000 in RNA-seq and 10,000 in ATAC-seq), we get ATAC-seq data and RNA-seq data on 415 and 463 single cells, respectively.

Data Processing. We align the scATAC-seq reads to reference genome mm9 and remove duplicates. We employed MACS2 (23) to do peak calling by merging all of the reads from all of the single cells. We consider only the narrow peaks which are at least present (one or more reads) on 10 cells. Read counts for each region on each cell are calculated by bedtools (24) with the intersect command. Features defined from scATAC-seq data consist of an openness index on regions including REs and narrow peaks from MACS2. REs include promoters and enhancers. We use REs that regulate at least one TG from the PECA network (7).

scRNA-seq raw reads are mapped to mm10 by STAR (25) with ENCODE options. Gene expression TPM are calculated by RSEM (26). The transcriptome annotation we use is GENCODE vM16.

Simulation of scRNA-Seq and scATAC-Seq. We simulate scRNA-seq data for each single cell from bulk RNA-seq data by following the Splatter pipeline (27). Specifically, it includes three steps: (*i*) adding noise on expression data T, $T = TPM + \varepsilon$, where ε is Gaussian noise with SNR = 5; (*ii*) getting expected read counts per gene $\lambda_i = NT_iL_i/\sum_j T_jL_jP \times 0.5\%$, where N is the total number of read counts in bulk data, L_i and T_i are gene length and its expression for gene i, and P reflects the sequencing depth for each single cell $P \sim Beta(2,4)$; and (*iii*) getting the observed read counts for each gene, $Y_i = D_iX_i$, where read count $X_i \sim Poi(\lambda_i)$, and dropout effect $D_i \sim Ber(1/1+\lambda_i^{-0.1})$. In scATAC-seq simulation, we use the same procedure by replacing the TPM as openness (defined as number of read counts per 1,000 bp per 100 million mapped reads). The distribution of read counts in our simulation data is similar to the distribution of reads counts in 10× genomics scRNA-seq data and C1 Fluidigm scATAC-seq data.

Experimental Design of RA-Induced mESC Differentiation. mESC lines R1 were obtained from ATCC. The mESCs were first expanded on a mouse embryonic fibroblasts feeder layer previously irradiated. Then, subculturing was carried out on 0.1% bovine gelatin-coated tissue culture plates. Cells were propagated in mESC medium consisting of Knockout DMEM supplemented with 15% Knockout Serum Replacement, 100 μM nonessential amino acids, 0.5 mM beta-mercaptoethanol, 2 mM GlutaMax, and 100 units/mL penicillin–streptomycin with the addition of 1,000 units/mL leukemia inhibitory factor (ESGRO; Millipore).

mESCs were differentiated using the hanging-drop method (28). Trypsinized cells were suspended in differentiation medium (mESC medium without LIF) to a concentration of 50,000 cells/mL. Twenty-microliter drops (~1,000 cells) were then placed on the lid of a bacterial plate and the lid was placed upside down. After 48 h incubation, EBs formed at the bottom of the drops were collected and placed in the well of a six-well ultralow attachment plate with fresh differentiation medium containing 0.5 μM RA for up to 4 d, with the medium being changed daily.

scATAC-Seq. We followed the scATAC-seq protocol published by Buenrostro et al. (2) with the following modifications. The EBs were first incubated with StemPro Accutase cell dissociation reagent (Gibco) at 37 °C for 10 min, and then the EBs were gently pipetted for an additional 15 min to obtain a single-cell suspension. To further remove nondissociated EBs, the cell suspension was filtered sequentially with a 40-μM cell strainer (BD Falcon) and a 20-μM pluriStrainer (pluriSelect). After washing three times with C1 DNA Seq Cell Wash Buffer, cells at a concentration of 350–400 cells/μL were loaded on the C1 Single-Cell Auto Prep System (Fluidigm, Inc.). Single cells were captured and processed on a 10- to 17-μM IFC microfluidic chip using ATAC-seq scripts (2). A total of seven IFC chips were included in this study. The library was sequenced on Illumina NextSeq with 75-bp paired-end reads.

scRNA-Seq. To prepare a scRNA-seq library, we followed the SMART-Seq v4 Ultra Low Input RNA Kit for the Fluidigm C1 System (Clontech Laboratories, Inc.). The EBs were first dissociated with Accutase as described previously. Cells at a concentration of 200–250 cells/μL were then loaded on the C1 Single-Cell Auto Prep System (Fluidigm, Inc.). The single cells were captured and processed on a 10- to 17-μM IFC microfluidic chip, using SMART-Seq v4 scripts. A total of five IFC chips were included in this study. After harvest, cDNA concentration for each sample was measured using the Fragment Analyzer Automated CE System (Advanced Analytical Technologies, Inc.) and the cDNA concentration we used for Nextera XT library preparation is ~0.2 ng/μL. The library was sequenced on Illumina HiSeq with 100-bp paired-end reads.

Software and Data. Software and simulation data are available at http://web.stanford.edu/~zduren/CoupledNMF/. Single-cell gene expression data and chromatin accessibility data of RA induction have been deposited in the GEO database under accession nos. GSE115968 and GSE115970.

ACKNOWLEDGMENTS. This work was supported by National Institutes of Health Grants R01HG007834, R01GM109836, and P50HG007735 and the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13000000).

1. Tang F, et al. (2009) mRNA-seq whole-transcriptome analysis of a single cell. *Nat Methods* 6:377–382.
2. Buenrostro JD, et al. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523:486–490.
3. Smallwood SA, et al. (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* 11:817–820.
4. Kiselev VY, et al. (2017) SC3: Consensus clustering of single-cell RNA-seq data. *Nat Methods* 14:483–486.
5. Habib N, et al. (2017) Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods* 14:955–958.
6. Zamanighomi M, et al. (2017) Unsupervised clustering and epigenetic classification of single cells. bioRxiv:10.1101/143701. Preprint, posted December 4, 2017.
7. Duren Z, Chen X, Jiang R, Wang Y, Wong WH (2017) Modeling gene regulation from paired expression and chromatin accessibility data. *Proc Natl Acad Sci USA* 114:E4914–E4923.
8. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.
9. Corces MR, et al. (2016) Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* 48:1193–1203.
10. Lake BB, et al. (2018) Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* 36:70–80.
11. Bacher R, Kendziorski C (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol* 17:63.
12. Lara-Astiaso D, et al. (2014) Immunogenetics. Chromatin state dynamics during blood formation. *Science* 345:943–949.
13. Brunet J-P, Tamayo P, Golub TR, Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci USA* 101:4164–4169.
14. Gorkin D, et al. (2017) Systematic mapping of chromatin state landscapes during mouse development. bioRxiv:10.1101/166652. Preprint, posted August 3, 2017.
15. Lin S-C, et al. (2010) Endogenous retinoic acid regulates cardiac progenitor differentiation. *Proc Natl Acad Sci USA* 107:9234–9239.
16. Maden M, Holder N (1992) Retinoic acid and development of the central nervous system. *BioEssays* 14:431–438.
17. Hartigan JA (1972) Direct clustering of a data matrix. *J Am Stat Assoc* 67:123–129.
18. Cheng Y, Church GM (2000) Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol* 8:93–103.
19. Zhang S, Li Q, Liu J, Zhou XJ (2011) A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics* 27:i401–i409.
20. Bickel S, Scheffer T (2004) Multi-view clustering. *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pp 19–26.
21. Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ (2007) Algorithms and applications for approximate nonnegative matrix factorization. *Comput Stat Data Anal* 52:155–173.
22. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ (2017) chromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* 14:975–978.
23. Zhang Y, et al. (2008) Model-based analysis of ChIP-seq (MACS). *Genome Biol* 9:R137.
24. Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.
25. Dobin A, et al. (2013) Star: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15–21.
26. Li B, Dewey CN (2011) RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* 12:323.
27. Zappia L, Phipson B, Oshlack A (2017) Splatter: Simulation of single-cell RNA sequencing data. *Genome Biol* 18:174.
28. Wang X, Yang P (2008) In vitro differentiation of mouse embryonic stem (mES) cells using the hanging drop method. *J Vis Exp* 17:825, 10.3791/825.