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

Zhana Duren[1,#], Xi Chen[1,#], Mahdi Zamanighomi[1,2,#], Wanwen Zeng[1,3], Ansuman T Satpathy[2], Howard Y. Chang[2], Yong Wang[4,5], and Wing Hung Wong[1,2,*]

[1] Department of Statistics, Department of Biomedical Data Science, Bio-X Program
Stanford University, Stanford, CA 94305, USA;
[2] Center for Personal Dynamic Regulomes,
Stanford University, Stanford, CA 94305, USA
[3] MOE Key Laboratory of Bioinformatics,
Bioinformatics Division and Center for Synthetic & Systems Biology,
Department of Automation, Tsinghua University, Beijing 100084, China
[4] CEMS, NCMIS, MDIS, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100080, China.
[5] Center for Excellence in Animal Evolution and Genetics,
Chinese Academy of Sciences, Kunming, 650223, China

# Authors contributed equally to this work

* Corresponding author:

Wing Hung Wong, Email: whwong@stanford.edu

Mailing Address: Department of Statistics, Sequoia Hall, 390 Serra Mall, Stanford University, Stanford, CA 94305-4065

Phone: (1) 6507252915, Fax: (1) 6507258977

Abstract

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-seq and single cell ATAC-seq data.

Key words: coupled clustering, NMF, single cell genomic data

Significance Statements

Biological samples are often heterogeneous mixtures of different types of cells. Suppose we have two single cell data sets, 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 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-seq and single cell ATAC-seq data.

Introduction

Biological samples of interest in clinical or experimental studies are often heterogeneous mixtures, i.e. a sample may consist of many different 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-seq (1) enables transcription profiling, scATAC-seq (2) offers chromatin accessibility data, 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, which aims to resolve the single cells into the constituent subpopulations. Clustering methods for scRNA-seq data were discussed in (4, 5), and clustering of scATAC-seq data were given in (6). Existing methods, however, do not cover the increasing common situation when two or more types of sc-genomics experiments were performed on different subsamples from the same cell population. For example, Figure 1A depicts the situation when one subsample is analyzed by scRNA-seq while another is analyzed by scATAC-seq. Since both types of measurements are informative about the constituent cell types in the heterogeneous population, it is clear that we should aim to 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. In this paper we formulate this "coupled clustering" problem as an optimization problem, and introduce a method, called coupled nonnegative matrix factorizations (coupled NMF), for its solution.

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 factorization $O=W_1H_1$ as follows: the i^{th} column of W_1 gives the mean vector for the i^{th} cluster of units, while the j^{th} column of H_1 gives the assignment weights of the j^{th} 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. To couple two matrix factorizations, we solve the optimization in Figure 1 C, where the cost function contains a term $\lambda_2 tr\left(\widetilde{W_2}^T A W_1\right)$ where $\widetilde{W_2}$ is a submatrix containing a subset of rows of W_2, and 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 of the 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. For the application of interest in the current paper, where gene expression is measured in one sample and chromatin accessibility is measured in the other sample, we will use a diverse panel of cell lines with both expression and accessibility data, to train a prediction model of gene expression from accessibility. See Approach section for details.

Approach

Construction of Data Matrices: From single cell ATAC-seq data, we compute a data matrix O, where O_{ij} denotes the degree of openness (i.e. accessibility) of the *i^th^* region in the *j^th^* cell (6). By region we mean union of predefined regulatory elements (REs) and peaks. From single cell RNA-seq data, we compute the data matrix E where E_{gh} denotes the expression level of the g^{th} gene in the h^{th} cell (7). Details are given under data processing in the Methods and Materials. Note that the single cell ATAC-seq and the single cell RNA-seq data are not measured in the same cell (Fig. 1 A).

Construction of coupling matrix: Our approach to the initialization of A is to look for a subset of genes whose expression is highly predictable from chromatin accessibility of nearby REs. To do this, we take advantage our recent work on modeling paired gene expression and chromatin accessibility data (on bulk samples) across diverse cellular contexts (8). From the PECA model in that work, for each gene g, we can extract a set S_g of REs that regulate that gene. We consider the regression model of target gene (TG) expression (denoted as E_g) on its regulatory elements' (RE) accessibility (denoted as O_i).

$$E_g = \alpha_{g0} + \sum_{i \in S_g} \alpha_{gi} O_i \tag{1}$$

We estimate the parameter α_g by fitting the penalized least square problem (*eq.* 2 below) 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 (Supplementary Table 1).

$$\min_{\alpha_g} \frac{1}{2} \left\| E_g - \alpha_{g0} - \sum_{i \in S_g} \alpha_{gi} O_i \right\|_F^2 + \lambda(\|\alpha_g\|_1 + \|\alpha_g\|_2^2) \tag{2}$$

where λ is determined by 5-fold cross validation. After fitting the model, we select a set of "well predicted" genes (denoted as S) for which the relative prediction error is less than 0.3, that is, $|E_{gs} - \widehat{E_{gs}}| < 0.3 E_{gs}$, for at least 80% of the cell lines in the panel. In this way, we selected 5,281 well predicted genes in mouse and select 6,537 well predicted genes in human. The initial value of the coupling matrix A is obtained from the coefficients α_g associated with these well predicted genes.

Clustering by Coupled nonnegative matrix factorizations: The gene set S and the matrix A allow us to couple expression-based clustering and accessibility-based clustering according to a regulatory model supported by extensive prior data. 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*).

$$\min_{W_1, H_1, W_2, H_2 \geq 0} \frac{1}{2} \|O - W_1 H_1\|_F^2 + \frac{\lambda_1}{2} \|E - W_2 H_2\|_F^2 - \lambda_2 tr\left(\widetilde{W_2}^T A W_1\right) + \mu(\|W_1\|_F^2 + \|W_2\|_F^2) \tag{3}$$

Where $\widetilde{W_2}$ represent a submatrix of W_2 only containing the rows corresponding to the genes in S, the first row of $\widetilde{W_2}$ is [1,1,...1], which corresponding to the first column of matrix A $[\alpha_{10}, \alpha_{20}, \dots \alpha_{n0}]^T$, the intercept of the regression model in *eq. (1).* After solving the optimization, the cluster profile and the cluster assignments for the k^{th} cluster in the accessibility data can be obtained respectively from the k^{th} column of W_1 and the k^{th} row of H_1. Similarly, the expression-based clustering can be obtained from W_2 and H_2 (Fig. 1*D*). In this application, AW_1 gives the cluster-specific predictions of the expression of genes in S based on the cluster-specific accessibilities of REs, and hence the trace term enforces our expectation that the expression of genes in S should be consistent with the predictions based on accessibility of nearby REs. We can refine the coupling iteratively, as follows. 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 single cell RNA-seq clustering. Then we restrict the gene set S in cluster-specific gene and re-cluster the cells by optimizing the objective function in *eq.* (3). We continue until the cluster assignments are not changed by further iterations.

Results

Results on simulation data: The performance of our coupled clustering method is first evaluated in a simulation study where single cell data is simulated by mixing reads that are sampled from two bulk data sets corresponding to two cell types. The bulk data sets used in our simulation study are from two very similar cell types from hematopoietic differentiation process, namely common myeloid progenitor (CMP) and megakaryocyte erythroid progenitor (MEP) (9). We randomly sample 3,000 reads from bulk RNA-seq data and add a Gaussian noise of SNR = 5 to simulate a single cell RNA-seq. Similarly, we randomly select 40,000 reads from bulk ATAC-seq data and add a Gaussian noise of SNR = 5 to simulate a scATAC-seq data. Number of reads in our simulation data are similar to the number of reads in 10 X genomics scRNA-seq and C1 Fluidigm scATAC-seq data. In our simulation, we generated 100 scRNA-seq and 100 scATAC-seq for CMP and MEP respectively.

To simulate a scRNA-seq data set from a mixed population with two cell types, we simply combine the 200 single cell RNA-seq data from two cell lines together and treat it as a single scRNA-seq data set. We then apply k-means and non-negative matrix factorization (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 Frobenius norm. The results of all the 50 runs on scRNA-seq and scATAC-seq data by k-means and NMF are shown in Supplementary Fig. S1. Finally, we perform coupled NMF clustering based on both the scRNA-seq sample (200 cells mixture) and the scATAC-seq sample (200 cells mixture). 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) are presented in Figure 2A). A similar improvement is also seen in the clustering results on the scATAC-seq sample (Fig 2B). It is seen that coupling leads to greatly improved results, reducing the assignment error rate by more than 4 folds over the other two methods (Fig 2C).

Assessment of prediction model before coupling: We are interested in applying coupled NMF to analyze data generated from differentiation of mouse embryonic stem cell, namely scRNA-seq and scATAC-seq at day 4 after retinoic acid (RA) treatment (Methods and Materials). To assess whether the model learned from the diverse panel still have good predictive power in this new biological context, we first generated bulk RNA-seq and ATAC-seq from this context (i.e. RA day-4). Using the model trained on the diverse panel, we predicted the expression at RA day-4 of genes in gene set S based on the accessibility data at RA day-4. Figure 3 presents the observed versus predicted scatter plot, which shows the genes in S can indeed be predicted with high accuracy in this context (R^2=0.86, r=0.93). This gives us confidence in using the model to initiate the coupling.

Results on real single cell data: Next, we test our method on scRNA-seq and scATAC-seq data generated from the RA day-4 cell population. We first perform coupled NMF with K = 2 (i.e. 2 clusters) and then visualize the clustering result on a t-SNE plot computed based on Spearman correlation. There are clearly 2 separated clusters in both the t-SNE plot from scATAC-seq and the t-SNE plot from scRNA-seq (Supplementary Fig. S2). When we increase the number of clusters to 3, we can see 3 well separated clusters in t-SNE plot. However, when K is increased to 4 or 5, the resulting clusters are no longer clearly separated (Supplementary Fig. S2). Thus we conclude that K=3 is sufficient for this data.

For each of the 3 clusters, we identify cluster-specific TFs based on their expression from RNA-seq data, and compare their motif activities across different clusters (Figure 4). Here the motif activity of a TF reflects the degree of enrichment of the TF's motif on accessible REs (see Methods and Materials). Figure 4B shows the motif activities and expressions of some cluster-specific TFs on the t-SNE visualization (Figure 4B, Supplementary Fig. S4). Fig. 4 C shows the heat maps of motif activities and expressions for a subset cluster specific TFs, namely those with expression 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 cluster-1 specific peaks. 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 & 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.

Next, we select cluster-specific genes from RNA-seq data and cluster-specific peaks from ATAC-seq data. We assess whether the cluster-specific peaks are significantly close to the cluster-specific-genes by performing Fisher exact test based on the count of such gene-peak pairs that are within 100kb of distance each other (Supplementary Fig. S5). Figure 4D 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 on the constituent subpopulations. First, we use cluster-specific peaks from single cell ATAC-seq data to annotate the clusters. We collect previously determined enhancers in mouse tissues at 7 developmental stages from 11.5 days post conception until birth (10). Fig. 5 *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 Jaccard index (intersection over union) and NA indicates that enhancer data for that tissue in that stage is 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. We also collect experimentally validated tissue-specific enhancers from VISITA database and overlap them to cluster-specific peaks. Fig. 5 *D* shows the percentage of tissue specific VISITA enhancers overlapped to cluster-specific-peaks. Only those tissues with at least one enhancers overlapping with the cluster-specific peaks are shown. Enhancers from neural 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 vessels enhancers and heart enhancers. These results suggest that cluster 1 and 3 may be related to neural tissues and cluster 2 may be related to heart tissue.

In addition, we analyzed cluster-specific genes from scRNA-seq data. Fig. 5 *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, cluster 1 and 3 are enriched in neuron 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 neural associated clusters, there are interesting differences. Cluster 1 is more enriched in axon guidance and neuron projection guidance, which are relevant for general neuronal functions. On the other hand, cluster 3 are more enriched in brain development and oligodendrocytes differentiation, which are specifically relevant to the central nervous system (CNS). To examine this further, we check the expression of some known markers for motor neurons, which are present in PNS (peripheral nervous system) rather than in CNS. Fig. 5F shows the percentage of cells expressing (i.e. TPM>2) motor neuron markers (Chat, Isl1, Isl2 and En1). Motor neuron markers are more highly expressed in cluster 1 compared to cluster 3. Overall our results suggest that retinoic acid induced stem cell at day 4 is a mixture of cells related to peripheral nervous system, cardiovascular system, and central nervous system. These results are largely consistent with the previous studies (11, 12).

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 Supplementary Fig. S6. It is seen that Klf7, Ebf1, Sox11, and Nhlh1 are playing 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.

Discussions

As far as we know, coupled clustering is a new problem different from other complex clustering tasks such as bi-clustering or multi-view clustering. Bi-clustering (13, 14), also called block clustering or co-clustering, has been used widely used 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 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 multi-view clustering (15), 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 multi-view clustering and coupled clustering is that in the former setting all features are measured on each subject, whereas in the latter 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 in order to initialize the coupling.

Measurements of different types of features generally require different types of reagents and biochemical reactions. To measure multiple types of features in the same cell, it is necessary to carry out all the required reactions within the same cell, which is extremely challenging technically. Although there are current efforts to develop methods for simultaneous measurement of two types of measurement in single cells (for example, RNA+DNA, RNA+accessibility, or RNA+CpG methylation, etc), these methods are still not yet ready for general use. Furthermore, even when we have two types of data (say RNA + accessibility) on single cells, there are always additional features (say CpG methylation) that we may want to incorporate into the analysis. For example, we may have two types of features measured in single cells in the first sample, and a third type of features measured on single cells in another sample. Then we are again facing a coupled clustering problem. Thus our method for coupled clustering is of interest to single cell genomics.

Methods and Materials

Optimization algorithm: We optimize the object function in *eq. (3)* by multiplicative update algorithm.

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

$$w_{ij}^2 \leftarrow w_{ij}^2 \frac{\left(XH_2^T + \frac{\lambda_2}{2\lambda_1} AW_1\right)_{ij}}{(W_2 H_2 H_2^T + 2\mu W_2)_{ij}}$$

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

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

Where w_{ij}^1 represent the element of *ith* row and *jth* column in matrix W_1, 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 t-test to define the cluster-specific genes and cluster-specific peaks, default p-value cut-off 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 correlation $K \times K$ matrix of AW_1 with $\widetilde{W_2}$, which is denoted by R. We use the determinant of correlation matrix R to measure the consistency of true expression values and the predicted values. Higher determinant means higher diagonal of the matrix, which means higher correlation between matched clusters and lower correlation in unmatched clusters.

Parameters 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 alternating least squares (ALS) algorithm with 50 different initializations using a Monte Carlo type approach (16) and get the solutions W_{10}, H_{10}, W_{20}, H_{20}, which are used as initial solution in our optimization problem. We choose parameters $\lambda_1 = \frac{\|O - W_{10} H_{10}\|_F^2}{\|E - W_{20} H_{20}\|_F^2}$, $\lambda_2 = \eta \frac{\frac{1}{2}\|O - W_{10} H_{10}\|_F^2}{tr\left(\widetilde{W_{20}}^T A W_{10}\right)}$. Tuning parameters η and μ 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. The number of clusters K can be determined by a method similar to that in ref (17).

TF motif activity: We use software chromVAR (18) to calculate the TF motif activity on each of the 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 retinoic acid (RA) treatment. After 4 days' induction, we sample cells for bulk RNA-seq and bulk ATACC-seq experiments for use in validating the coupling. To test the couple NMF clustering method, we also generated single cell ATAC-seq and single cell RNA-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 single cell ATAC-seq reads to reference genome mm9 and remove duplicates. We employed MACS2 (19) to do peak calling by merging all the reads from all the single cells. We only consider the narrow peaks which at least present (1 or more reads) on 10 cells. Read counts for each region on each cell are calculated by bedtools (20) with intersect command. Feature of scATAC-seq data is regions including regulatory elements (REs) and narrow peaks from MACS2. REs include promoters and enhancers. We use REs that regulate at least one target gene from PECA network (8).

Single cell RNA-seq raw reads are mapped to mm10 by STAR (21) with ENCODE options. Gene expression transcripts per million (TPM) are calculated by RSEM (22). The transcriptome annotation we use is GENCODE vM16.

Experimental design of retinoic acid-induced mESC differentiation: Mouse ES cell lines R1 were obtained from ATCC. The mESCs were first expanded on an MEF 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 U/mL Penicillin-Streptomycin with the addition of 1,000 U/mL of LIF (ESGRO, Millipore).

mESCs were differentiated using the hanging drop method (23). Trypsinized cells were suspended in differentiation medium (mESC medium without LIF) to a concentration of 50,000 cells/ml. 20 μl drops (~1000 cells) were then placed on the lid of a bacterial plate and the lid was upside down. After 48 h incubation, Embryoid bodies (EBs) formed at the bottom of the drops were collected and placed in the well of a 6-well ultra-low attachment plate with fresh differentiation medium containing 0.5 μM retinoic acid (RA) for up to 4 days, with the medium being changed daily.

Single cell ATAC-seq: We followed the single cell ATAC-seq protocol published by Jason et al. (24) with the following modifications. The EBs were first incubated with StemPro Accutase cell dissociation reagent (Gibco) at 37^{o}C for 10 min, then the EBs were gently pipetted for additional 15 min to obtain single cell suspension. To further remove non-dissociated EBs, the cell suspension was filtered sequentially with a 40 μM cell strainer (BD Falcon) and a 20 μM pluriStrainer (pluriSelect). After washing 3 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 10-17 μM IFC microfluidic chip using ATAC-seq scripts (24). Total 7 IFC chips were included in this study. The library was sequenced on Illumina NextSeq with 75 bp paired-end reads.

Single cell RNA-seq: To prepare single cell RNA-seq library, we followed 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 10-17 μM IFC microfluidic chip using SMART-Seq v4 scripts. Total 5 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 web.stanford.edu/~zduren/CoupledClustering/. Single cell gene expression data and chromatin accessibility data of RA induction have been deposited in the GEO database under accession no. GSEXXXXX.

A**cknowledgments:** This work was supported by grants R01HG007834, R01GM109836, and P50HG007735 from the National Institutes of Health (NIH).

References

1. Tang F*, et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* 6(5):377.
2. Buenrostro JD*, et al.* (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*.
3. Smallwood SA*, et al.* (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods* 11(8):817.
4. Kiselev VY*, et al.* (2017) SC3: consensus clustering of single-cell RNA-seq data. *Nature methods* 14(5):483.
5. Habib N*, et al.* (2017) Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature methods* 14(10):955.
6. Zamanighomi M*, et al.* (2017) Unsupervised clustering and epigenetic classification of single cells. *bioRxiv*:143701.
7. Bacher R & Kendziorski C (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome biology* 17(1):63.
8. Duren Z, Chen X, Jiang R, Wang Y, & Wong WH (2017) Modeling gene regulation from paired expression and chromatin accessibility data. *Proceedings of the National Academy of Sciences*:201704553.
9. Lara-Astiaso D*, et al.* (2014) Chromatin state dynamics during blood formation. *science* 345(6199):943-949.
10. Gorkin D*, et al.* (2017) Systematic mapping of chromatin state landscapes during mouse development. *bioRxiv*:166652.
11. Lin S-C*, et al.* (2010) Endogenous retinoic acid regulates cardiac progenitor differentiation. *Proceedings of the National Academy of Sciences* 107(20):9234-9239.
12. Maden M & Holder N (1992) Retinoic acid and development of the central nervous system. *Bioessays* 14(7):431-438.
13. Hartigan JA (1972) Direct clustering of a data matrix. *Journal of the american statistical association* 67(337):123-129.
14. Cheng Y & Church GM (2000) Biclustering of expression data. *Ismb*, pp 93-103.
15. Bickel S & Scheffer T (2004) Multi-view clustering. *ICDM*, pp 19-26.
16. Berry MW, Browne M, Langville AN, Pauca VP, & Plemmons RJ (2007) Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis* 52(1):155-173.

17. Brunet J-P, Tamayo P, Golub TR, & Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences* 101(12):4164-4169.
18. Schep AN, Wu B, Buenrostro JD, & Greenleaf WJ (2017) chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *nature methods* 14(10):975.
19. Zhang Y*, et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome biology* 9(9):R137.
20. Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841-842.
21. Dobin A*, et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15-21.
22. Li B & Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* 12(1):323.
23. Wang X & Yang P (2008) In vitro differentiation of mouse embryonic stem (mES) cells using the hanging drop method. *JoVE (Journal of Visualized Experiments)* (17):e825-e825.
24. Buenrostro JD*, et al.* (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523(7561):486.

Figure legends

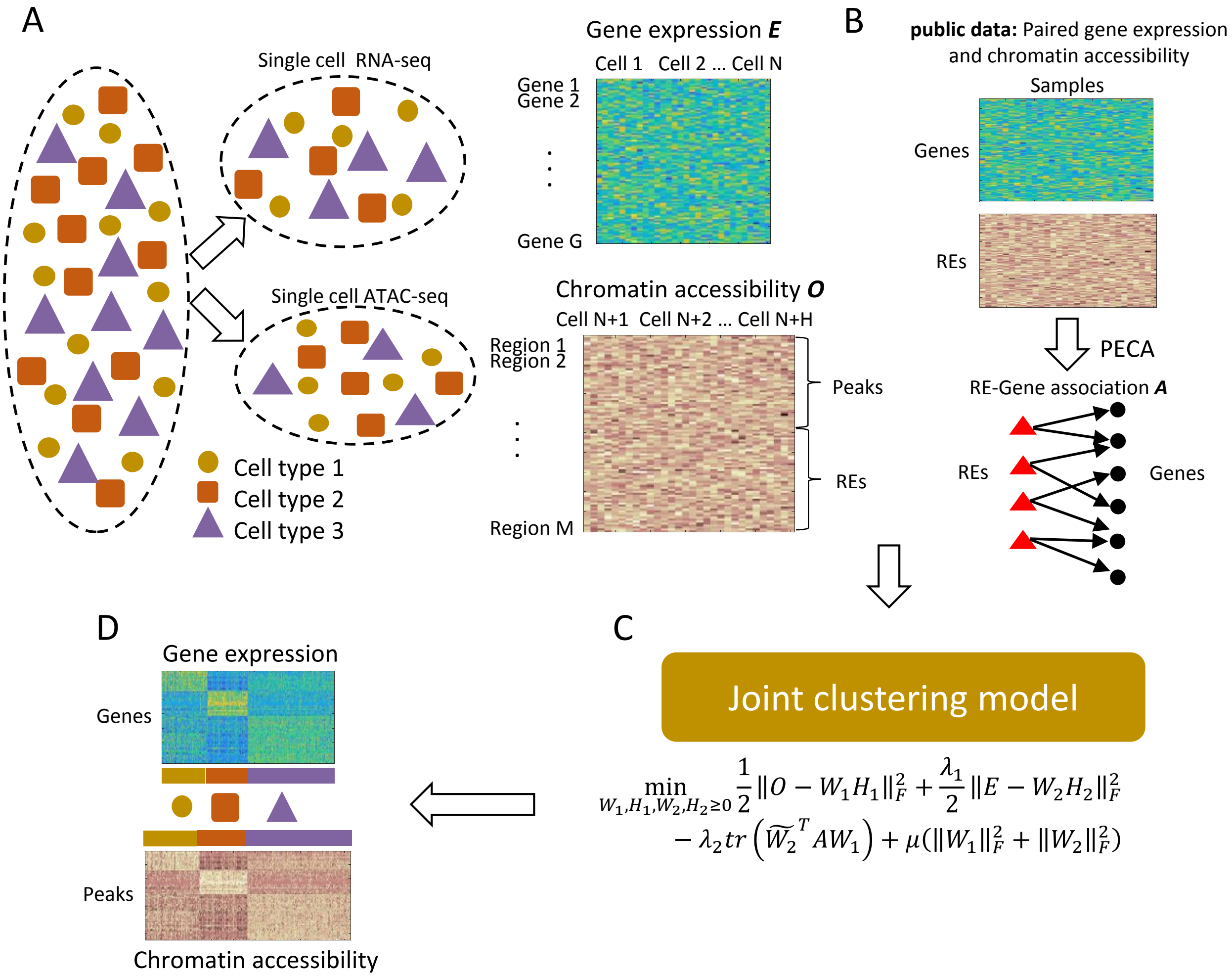
Fig. 1. Overview of the coupled clustering method.

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

Fig. 3. Scatter plot of RA day 4 gene expression (log(FPKM+1)) verses predicted expression values from our model.

Fig. 4. (A) t-SNE plot of single cell RNA-seq (right) and single cell ATAC-seq data (left) from RA day 4. Different colors represent clustering assignment from coupled clustering method. (B) Same t-SNE plots as Fig. 3 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 on cluster level. (D) Overlap of cluster-specific peaks nearby genes with cluster-specific genes. The values in table represent fisher's exact test p-value and fold change.

Fig. 5. (A-C) Similarity of cluster-specific peaks with enhancers of 12 tissues' 7 developmental stages. The number represent 10,000 times Jaccard index and NA represent enhancer data of that tissue in that stage are not available. (D) Percentage of VISITA enhancer that overlapped with cluster-specific peaks. (E) GO enrichment of cluster-specific genes. (F) Comparison of motor neuron markers' expression on cluster 1 and 3. Number represent the percentage of expressed cells (TPM >2) on cluster.



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

Fig. 1

A

K-means (50 replicates)

RNA-seq	CMP	MEP
Cluster 1	**58**	20
Cluster 2	42	**80**

NMF (50 replicates)

RNA-seq	CMP	MEP
Cluster 1	**80**	24
Cluster 2	20	**76**

Coupled clustering

RNA-seq	CMP	MEP
Cluster 1	5	**95**
Cluster 2	**95**	5

B

K-means (50 replicates)

ATAC-seq	CMP	MEP
Cluster 1	30	**51**
Cluster 2	**70**	49

NMF (50 replicates)

ATAC-seq	CMP	MEP
Cluster 1	28	**72**
Cluster 2	**72**	28

Coupled clustering

ATAC-seq	CMP	MEP
Cluster 1	8	**92**
Cluster 2	**92**	8

C

	K-means	NMF	Coupled clustering
RNA-seq err	62	44	10
ATAC-seq err	79	56	16
Error rate	35.25%	25.00%	6.50%

Fig. 2

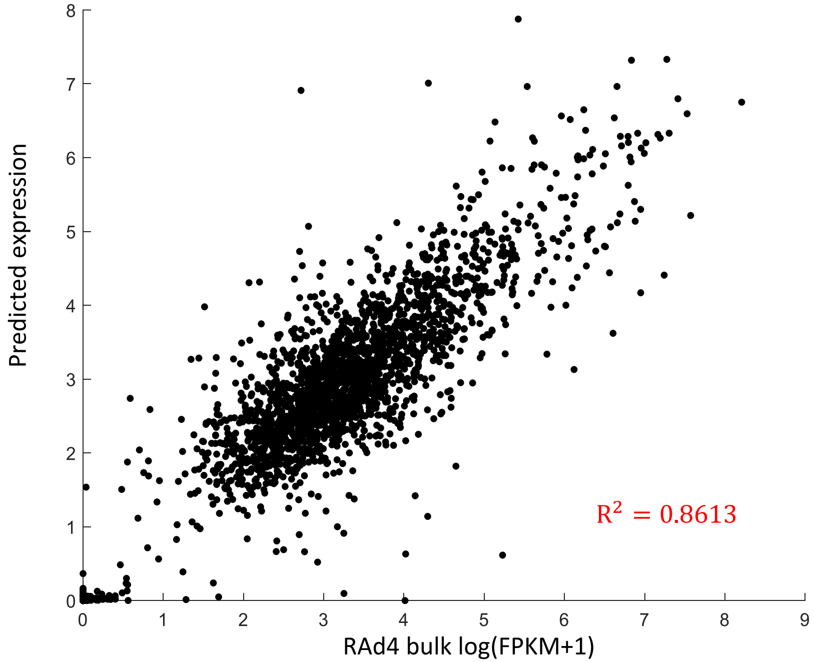
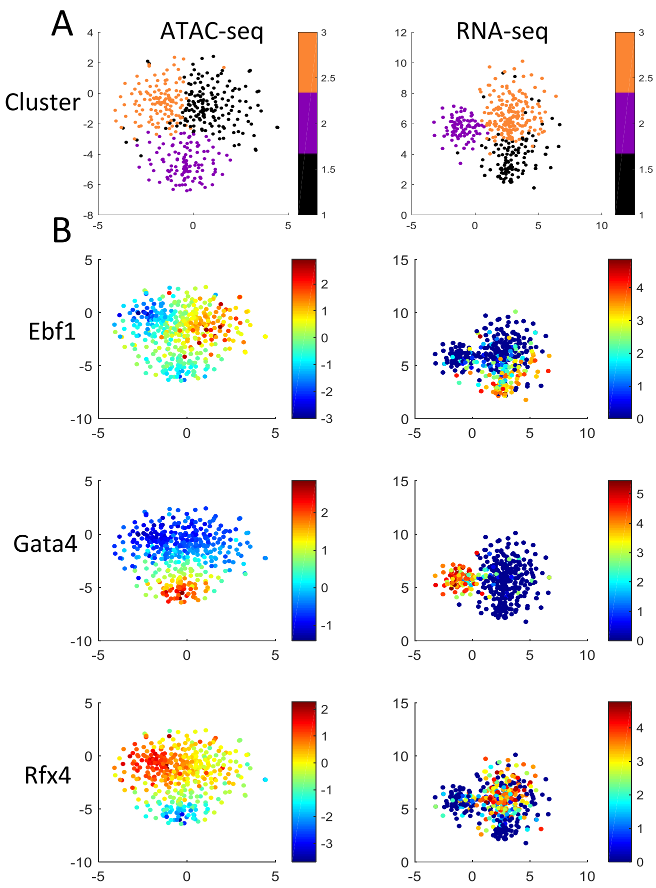
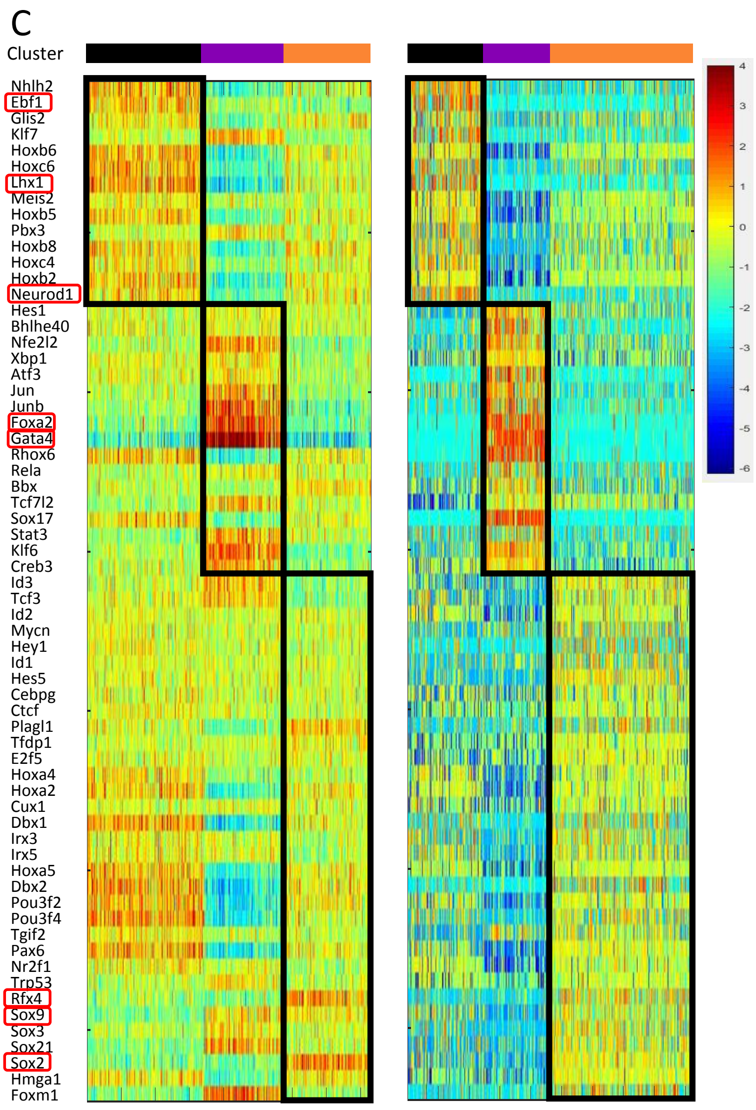

Predicted expression
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
RAd4 bulk log(FPKM+1)
$R^2 = 0.8613$

Fig. 3





D

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

A

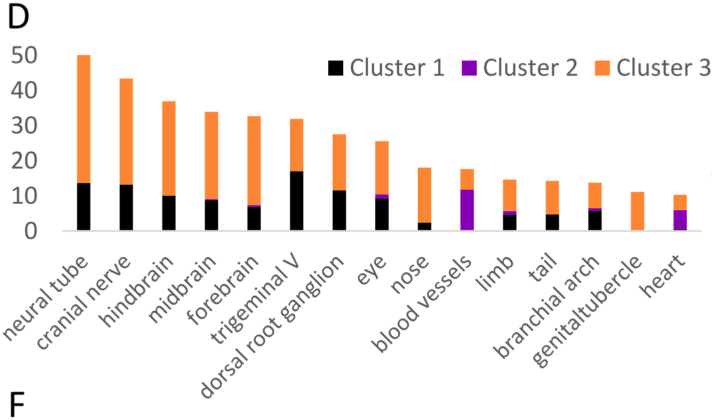
Cluster 1	facial prominence	forebrain	midbrain	hindbrain	neural tube	heart	intestine	kidney	limb	liver	lung	stomach
E11.5	9.29	5.67	13.82	8.87	13.48	7.25	NA	NA	7.62	6.87	NA	NA
E12.5	11.02	24.10	20.59	15.79	9.53	6.55	NA	NA	11.68	14.91	NA	NA
E13.5	8.45	25.25	21.73	16.89	19.76	9.92	NA	NA	5.70	12.54	NA	NA
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

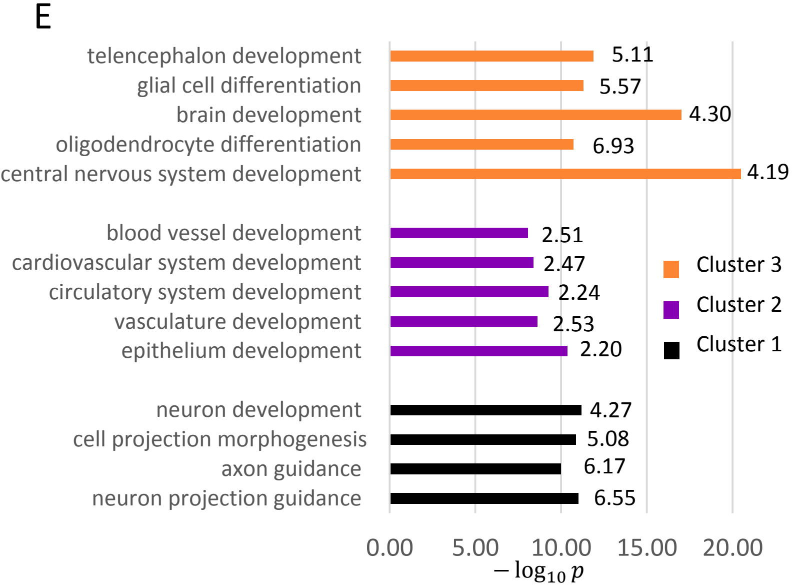
B

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



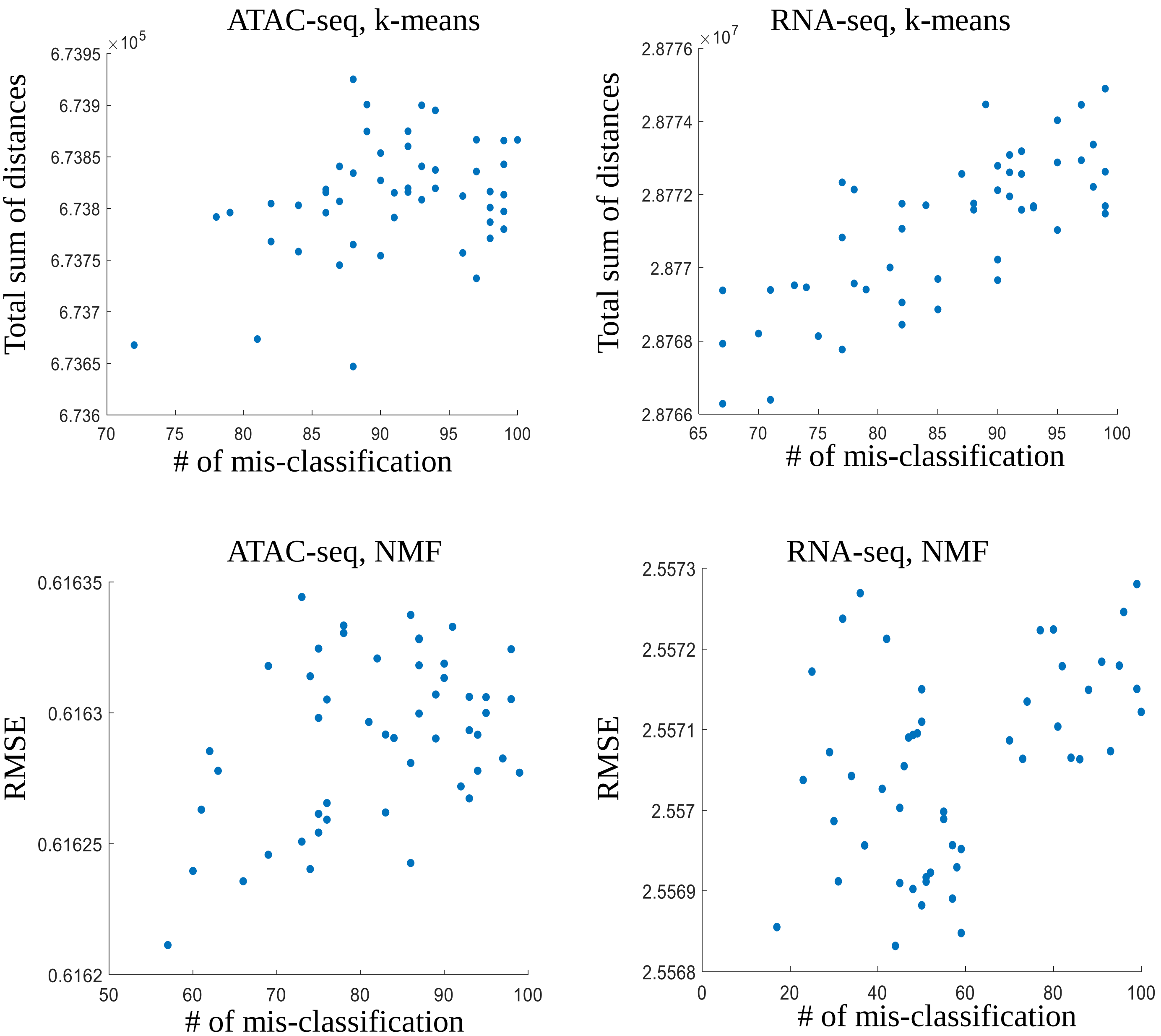


F

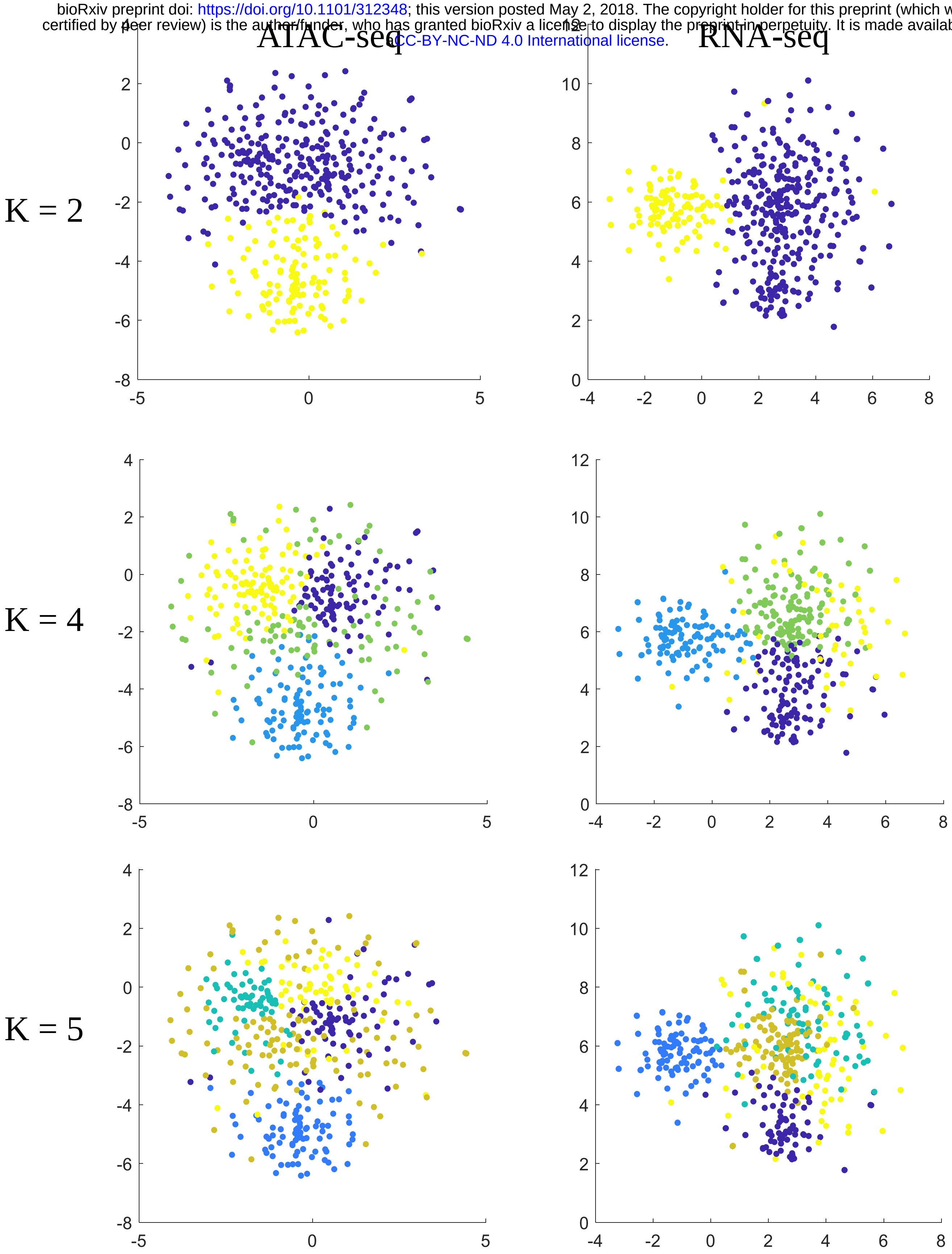
	% of Cluster 1	% of Cluster 3	Fold
Chat	9.68%	0.86%	11.26
Isl1	19.35%	6.47%	2.99
Isl2	13.71%	2.16	6.35
En1	12.90%	0.43%	30.00

Fig. 5

Supporting information



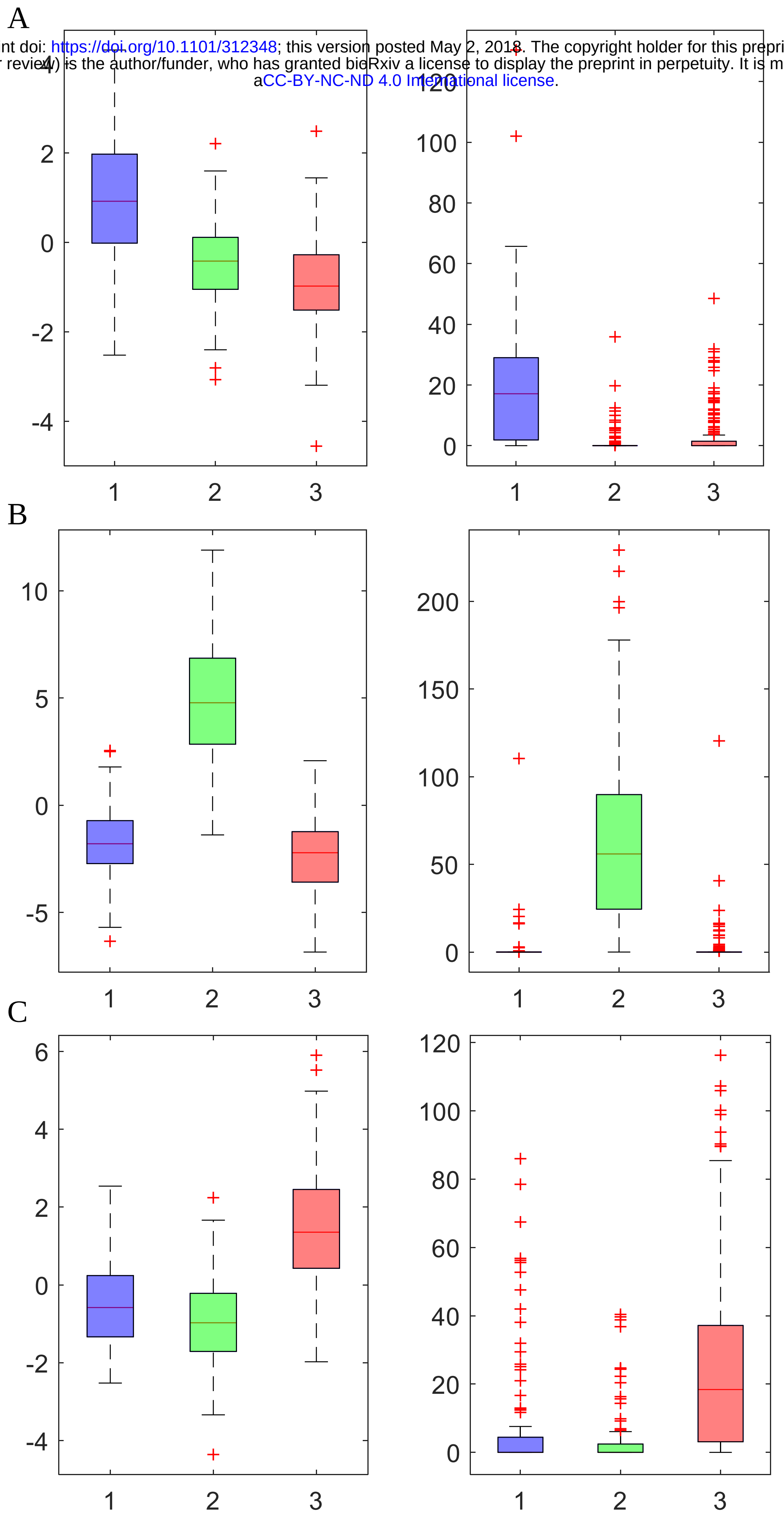
Supplementary Figure S1. Clustering results of k-means and NMF on simulation data. Each dot represents a run with random initial.



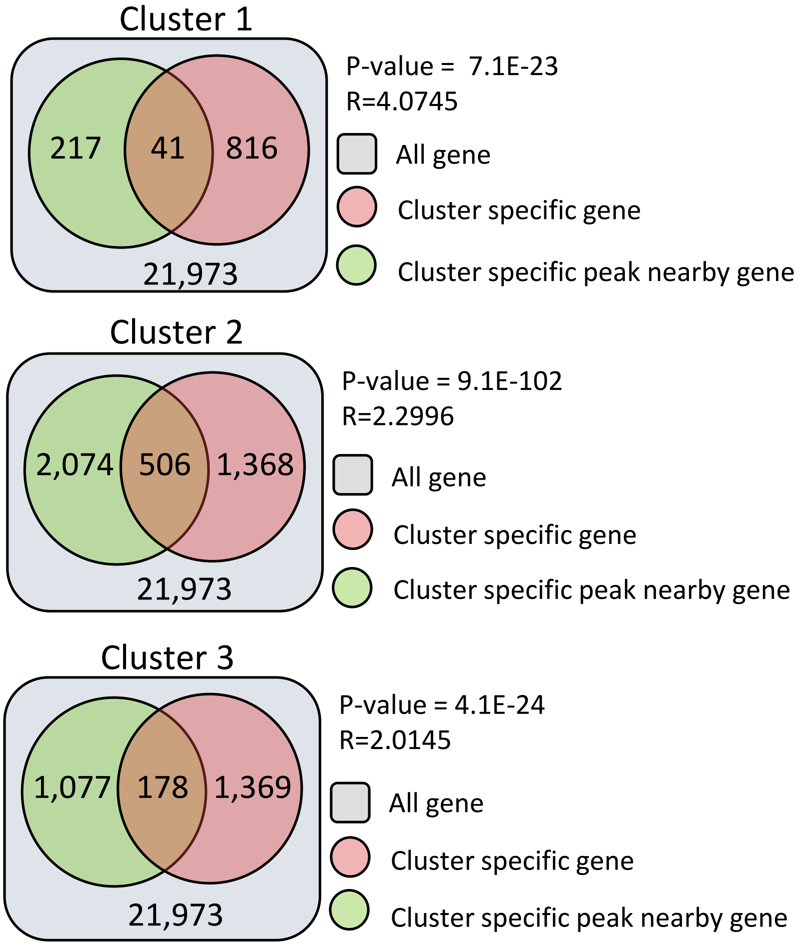
Supplementary Figure S2. t-SNE plot of scATAC-seq and scRNA-seq data on RA day 4 sample. Each dot represents one cell and different colors represent different clusters.

		μ							
		10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}	10^{3}	10^{4}
	10^{-3}	3.31E-09	3.78E-09	6.19E-09	1.97E-08	5.80E-09	-5.21E-07	0	0
	10^{-2}	6.09E-08	5.61E-08	5.52E-08	6.77E-08	7.06E-08	-4.90E-07	0	0
	10^{-1}	2.29E-06	1.83E-06	1.07E-06	6.72E-07	5.62E-07	-2.43E-07	0	0
η	10^{0}	9.04E-05	8.16E-05	4.81E-05	1.56E-05	6.82E-06	2.63E-06	0	0
	10^{1}	0.000804	0.000789	0.0007	0.000496	0.000204	2.18E-05	0	0
	10^{2}	0.003127	0.00312	0.003156	0.002863	0.001867	0.000864	0	0
	10^{3}	0.008771	0.008746	0.008363	0.00793	0.007892	0.009534	0	0
	10^{4}	0.006091	0.007224	0.008244	0.0107	**0.011394**	0.009734	0	0

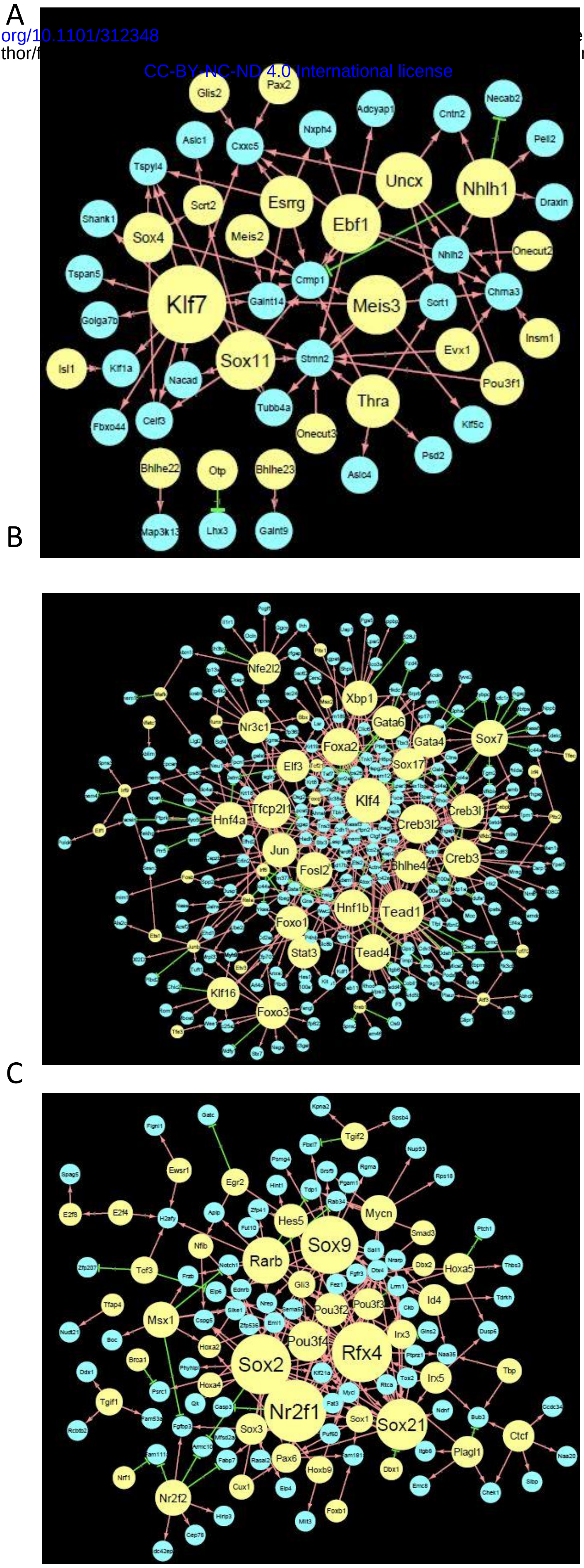
Supplementary Figure S3. Selecting tuning parameters μ and η. The values represent the determinant of correlation matrix R=corr (AW_1,W_2).



Supplementary Figure S4. Boxplot of cluster specific markers', Ebf1, Gata4, and Rfx4, motif activity (left) and gene expression (right).



Supplementary Figure S5. Comparison of cluster specific genes with the cluster specific peak nearby gene on each cluster from RA day 4 data.



Supplementary Figure S6. (A-C) Cluster specific networks of cluster 1, 2, and 3. Yellow color nodes represent TFs, red edges represent activation, and green edges represent repression. Circle size is proportional to the out-degree.

Supplementary Table legends

Supplementary Table 1. List of human and mouse paired gene expression and chromatin accessibility data used in regression model.