bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

SimiC: A Single Cell Gene Regulatory Network Inference method with Similarity Constraints

Jianhao Peng[1], *Ullas V. Chembazhi*[2], *Sushant Bangru*[2, 3], *Ian M. Traniello*[4, 5], *Auinash Kalsotra*[2, 3, 5], *Idoia Ochoa*[1, 6,*] *and Mikel Hernaez*[5, 7,*]

[1]*Electrical and Computer Engineering department, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.*
[2]*Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL, USA.*
[3]*Cancer Center@Illinois, University of Illinois at Urbana-Champaign, Urbana, IL, USA.*
[4]*Neuroscience Program, University of Illinois at Urbana-Champaign, Urbana, IL, USA.*
[5]*Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA.*
[6]*TECNUN School of Engineering, University of Navarra, Donostia, Gipuzkoa, Spain.*
[7]*Center for Applied Medical Research (CIMA), University of Navarra, Pamplona, Navarra, Spain.*

Abstract

Motivation: With the use of single-cell RNA sequencing (scRNA-Seq) technologies, it is now possible to acquire gene expression data for each individual cell in samples containing up to millions of cells. These cells can be further grouped into different states along an inferred cell differentiation path, which are potentially characterized by similar, but distinct enough, gene regulatory networks (GRNs). Hence, it would be desirable for scRNA-Seq GRN inference methods to capture the GRN dynamics across cell states. However, current GRN inference methods produce a unique GRN per input dataset (or independent GRNs per cell state), failing to capture these regulatory dynamics.
Results: We propose a novel single-cell GRN inference method, named SimiC, that jointly infers the GRNs corresponding to each state. SimiC models the GRN inference problem as a LASSO optimization problem with an added similarity constraint, on the GRNs associated to contiguous cell states, that captures the inter-cell-state homogeneity. We show on a mouse hepatocyte single-cell data generated after partial hepatectomy that, contrary to previous GRN methods for scRNA-Seq data, SimiC is able to capture the transcription factor (TF) dynamics across liver regeneration, as well as the cell-level behavior for the regulatory program of each TF across cell states. In addition, on a honey bee scRNA-Seq experiment, SimiC is able to capture the increased heterogeneity of cells on whole-brain tissue with respect to a regional analysis tissue, and the TFs associated specifically to each sequenced tissue.
Availability: SimiC is written in Python and includes an R API. It can be downloaded from `https://github.com/jianhao2016/simicLASSO_git`.
Contact: idoia@illinois.edu, mhernaez@illinois.edu
Supplementary information: Supplementary data are available at the code repository.

Introduction

Gene regulatory networks (GRNs) define regulatory relationships among genes, including for example relationships between transcription factors and their target genes. With the advent of next-generation sequencing, GRN inference has become one of the most important steps in determining cellular functions and modeling different systemic behaviors [4, 3]. Although they do not reflect post-transcriptional modifications, GRNs have been successfully used in many applications to elucidate new biological mechanisms and gene-level relationships in cells [7, 6].

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

These networks are usually represented as graphs, where the nodes are genes, and the edges represent a regulatory (or co-expression) relationship between the genes that they connect. These graphs can be classified, among others, as: directed, if the regulatory direction is known; weighted, where the weight of each edge represents the regulatory strength of the connection; or bipartite, where genes are split into disjoint sets and edges only connect genes of distinct sets. In addition, some GRN inference methods follow a module-based approach, where genes are first clustered in modules and then a GRN is inferred per module, in contrast to other methods that build a unique single GRN for the data [16, 30].

Until recently, most available gene expression data were derived from bulk RNA sequencing (RNA-Seq). These sequencing techniques are inherently agnostic to differences among diversity of cell type within a given sample, and can therefore only give an average measure of the gene expressions across all cells. Hence, GRNs inferred from bulk RNA-Seq data represent the transcriptional regulatory landscape of the sequenced tissue rather than of the individual cells.

With the advancement of single-cell RNA Sequencing technologies (scRNA-Seq), it is now possible to acquire gene expression data for individual cells in samples containing up to millions of cells. scRNA-Seq data is generally summarized as a matrix containing the expression values of genes (or transcripts) for each sequenced cell in a given sample. Further, the expression of each cell can be viewed as a snapshot of the cell development process, and similar snapshots can be clustered and ordered into groups which exhibit similar expression patterns. The groups of cells are often referred to as *cell states* and different states are characterized by different cell functionalities. Unfortunately, currently available scRNA-Seq expression data suffers from high sparsity caused by the dropout effect, which complicates downstream analysis [11]. Recent work on scRNA-Seq data has focused mainly on data cleaning (e.g., imputation [40]), cell type clustering (e.g., cell state inferring [20]), and pseudo-temporal ordering of single cells during cell development [21, 42, 38]

The nature of scRNA-Seq data naturally limits the applicability of traditional GRN inference methods to single cell expression data. In particular, GRN inference methods for bulk RNA-Seq data assume that the gene expressions across samples are independent and identically distributed (i.i.d.). However, this no longer applies to single-cell data, as cells in the same state exhibit similar expression patterns. The newly developed computational methods for scRNA-Seq data provide a new landscape for single-cell GRN inference, as the data can be imputed, and the information on cell types and pseudo-temporal ordering can be used as side information for GRN inference [13, 27]. For example, the ordering of cells is used to infer the regulatory relationships among genes via ordinary differential equations [23], correlation methods [34], information theoretic measures [8], or boolean functions [43]. To control the dropout effect, [26] used time-stamped scRNA-Seq data and transformed the expression data into (ordered) distances between expression profiles before inferring the relationships among genes using ridge regression; and [45] converted the expression profiles into images and used additional side information to train a deep convolutional neural network for inference of gene relationships.

Additionally, GRN inference methods for bulk RNA-Seq data [39, 24] have been used as intermediate steps in the processing of cell-type inference, rather than as the final goal for network inference [1].

Note that all of the above-mentioned GRN inference methods produce a unique GRN per input dataset. However, different cell states across a differentiation path will be potentially characterized by similar, but distinct enough, GRNs, as it is natural to assume that there should be a smooth transition between the GRNs associated with each state. Similarly, same cell-types under different phenotypes are expected to have slightly distinct GRNs, where the differences could be associated with the phenotype. In this context, directly inferring GRNs independently for each state might result in a group of divergent networks that share little in common. It would therefore be desirable in this case to add some constraint on the inferred GRNs to ensure that GRNs of cell states that are closely related to each other do not change abruptly.

We propose a novel GRN inference algorithm for scRNA-Seq data, SimiC, that considers both of the above situations. It takes as input the single-cell imputed expression data, the cell labels (states) and the ordering information (if it exists), and produces a bipartite driver-target network for each of the states. If an ordering between the cell states exist, SimiC adds a similarity constraint when jointly inferring the GRNs for each state, ensuring a smooth transition between the corresponding GRNs. All the GRNs produced by the proposed method share the same nodes, given by the driver and the target genes, and differ only in the edges that connect them. Each edge has a weight associated with it that indicates the influence of the driver gene in the corresponding target gene. Thus, each GRN can be represented as an incidence matrix between the driver and the target genes, where each entry of the matrix represents the weight of the edge connecting the corresponding driver and target gene. The change of network activity between cell states can therefore be inferred from the change on the corresponding incidence matrices.

Evaluating the quality and accuracy of GRNs is not a straightforward task, as there are no clear metrics for

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

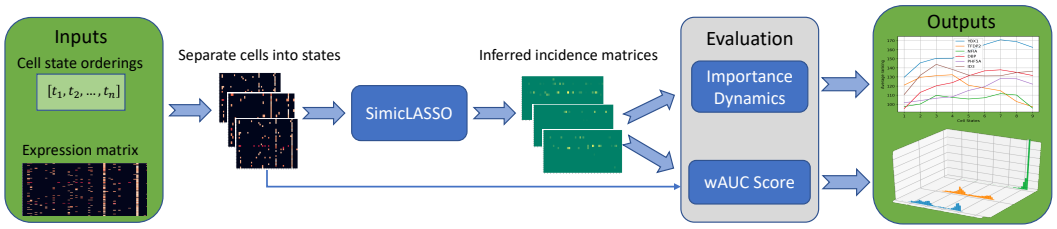


Fig. 1: Workflow of SimiC with the different evaluation metrics.

evaluation [27, 11]. Besides, most existing GRN inference algorithms only give a static picture of the GRN and provide their hypotheses based on such networks. SimiC, on the other hand, generates multiple networks at the same time, providing a framework to further evaluate the network dynamics. We propose several methods to fully evaluate the network dynamics and find the main driver-target gene connections, including: 1) an importance score of the driver genes for each target gene; 2) local regression on the importance of driver genes across states to find the change of roles at different cell states; 3) a novel weighted Area Under the Curve (wAUC) score per driver gene for each cell calculated from its neighbour edges that captures the activity of the driver gene (and associated neighbouring gene set) at each cell state; and 4) changes of distributions of the wAUC scores from one state to another to elucidate the regulatory dynamics of driver genes across cell states. All these analyses have been incorporated into SimiC and provide a more robust and biologically insightful hypothesis of potential driver-target gene regulation pairs.

Methods

In what follows we describe SimiC in more detail, both the method and the evaluation metrics, and provide details on the considered datasets, including any pre-processing steps to generate the input data needed by SimiC.

SimiC: LASSO with similarity constraints

The workflow of the proposed method SimiC is summarized in Fig. 1 (including the evaluation metrics). It takes as input the imputed scRNA-Seq expression data, the cell state (label) of each cell, and the (linear) ordering between cell states, if such information exists. Note that the cell states and their ordering can also be inferred from the pseudotime of cells, if available. The output of SimiC is a set of incidence matrices representing the GRNs associated to each cell state.

Notation: scRNA-Seq data is expressed as a matrix of size the number of sequenced cells by the number of measured genes. We further classify genes as driver or target genes, and use notation X and Y to represent their expression matrices, respectively. If there are c cells, m driver genes and n target genes, $X \in \mathbb{R}^{c\times m}$ and $Y \in \mathbb{R}^{c\times n}$. We use upper case notation $X_i \in \mathbb{R}^c$ ($Y_i \in \mathbb{R}^c$) to represent the expression of driver (target) gene i across cells, and lower case notation $x_i \in \mathbb{R}^n$ ($y_i \in \mathbb{R}^m$) to represent the expression of cell i across target (driver) genes. That is, X_i is the ith column of matrix X (analogously for Y), and x_i is the transpose of the ith row of matrix X (analogously for Y).

Feature selection: In order to apply the algorithm more efficiently and filter out the least informative genes for better robustness, we first perform a feature selection on the gene space. Specifically, we select separately the driver and target genes with the highest median absolute deviation (MAD) from their mean expressions. The MAD value is a measure of the variability across samples that is robust to outliers. Specifically, for a given gene expression profile X_i (or Y_i), its MAD value is given by $\text{MAD}(X_i) = \text{median}(|X_i - \bar{X}_i|)$, where $\bar{X}_i$ is the average value of X_i. The expression of a gene is more scattered if its MAD is larger and vice versa.

GRN inference background: In the general setting of bipartite gene regulatory network inference, given a set of m driver genes and a desire set of n target genes, the goal is to find a weighted bipartite graph between these two sets, where the weights describe the regulation activity of gene pairs. The common assumption is that the expression of the target genes of a cell i, denoted as $y_i \in \mathbb{R}^n$, can be approximated by the linear combination of its driver genes, denoted as $x_i \in \mathbb{R}^m$, under a Gaussian noise assumption, i.e., $y_i = W^T x_i + b + \epsilon_i$. Here the matrix $W \in \mathbb{R}^{m\times n}$ is the incidence matrix between driver and target genes, with the j-th column W_j

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

being the connectivity strength between target gene j and the set of driver genes. For ease of notation, we include the bias term $b \in \mathbb{R}^n$ into W by extending x_i to $[x_i, 1] \in \mathbb{R}^{m+1}$.

Given a group of i.i.d. expression profiles, the common approach is to minimize the approximation error $\frac{1}{2}|Y - XW|^2$ [16]. The solution to such least squares problem usually results in a dense incidence matrix W. Yet, in practice, it is believed that only a subset of driver genes regulate a given target, and hence the connection in the graph should be sparse [16, 7]. In order to have a more robust model and sparse incidence matrices, the most common optimization problem for GRN inference is expressed as:

$$\min_{W \in \mathbb{R}^{m \times n}} f(W) = \min_{W \in \mathbb{R}^{m \times n}} \frac{1}{2}|Y - XW|^2 + \lambda |W|_1 \tag{1}$$

This is a LASSO formulation, where Y is the target expression matrix and X is the driver expression matrix [15]. Such approach is effective when the expression matrix is composed of different samples of bulk sequencing profiles, which can be treated as i.i.d. samples. However, in the case of single-cell RNA sequencing, this assumption is no longer true. In addition, different cell states are expected to be governed by different GRNs, which correspond to different incidence matrices. To account for this, one approach would be to apply LASSO independently on each cell type. However, we would lose the information from the other cell states, which might be useful given that the cells generally come from the same region (or tissue) and a linear ordering may exist between them. In other words, in most scenarios the underlying regulatory networks of the different cell types are expected to share some common functions due to the asynchronous cell progression, which translates into some level of similarity between the corresponding incidence matrices.

SimiC: Here we propose a novel method, SimiC, which takes both the inter cell-state variance and the intra cell-state similarity into account, and jointly infers GRNs for each cell state. With the imputed expression data, the cell state labels for each cell, and the associated ordering, our optimization problem for GRN inference is defined as:

$$\min_{W^k, k \in [1:K]} f(W^1, W^2, \cdots, W^K) = \min_{W^k, k \in [1:K]} \sum_{k=1}^{K} \frac{1}{2}|Y^k - X^k W^k|^2 \tag{2}$$

$$+ \sum_{k=1}^{K} \lambda_1 |W^k|_1 + \sum_{k=1}^{K-1} \lambda_2 \left| W^k - W^{(k+1)} \right|_2^2, \tag{3}$$

where W^k is the incidence matrix of the GRN that we want to infer for cell state k, K is the number of states, Y^k is the target expression matrix of cells in state k, and X^k is the corresponding driver expression matrix. Assuming n target genes and m drivers, and s_k cells under cell state k, the dimensions of Y^k, X^k and W^k are $s_k \times n$, $s_k \times m$, and $m \times n$, respectively. Note that the dimension of X^k and Y^k may change across different states, but the dimension of the incidence matrices is always the same. The reason is that the GRNs for different cell states share the same set of nodes (driver and target genes), and only differ in the edge weights. The first summation in our objective with the ℓ_1 regularization term, $\sum_{k=1}^{K} \lambda_1 |W^k|_1$, serves the same purpose as in LASSO, i.e., it controls the sparsity of the incidence matrices. With or without the cell states ordering, minimizing the first part (i.e., setting λ_2 to zero) is equivalent to solving LASSO for every state independently.

The second regularization term, $\sum_{k=1}^{K-1} \lambda_2 \left| W^k - W^{(k+1)} \right|_2^2$, is the *similarity constraint*. As mentioned in the previous section, we would like to smooth the GRNs transition process (we assume in the formulation that the cell states $[1:K]$ are linearly ordered). With the order of cell states given, it is reasonable to assume that two consecutive states should share common edges. This translates into minimizing the pairwise difference of the corresponding GRNs, so as to maintain the common graph structure among them. Note that adding a second order regularization term will tend to make the incidence matrices denser. The trade-off between adding the sparsity constraint and the similarity constraint is controlled through the values of λ_1 and λ_2. For example, in cases in which the cell states are well separated, the smoothness assumption of the GRNs is weaker and hence λ_2 should be smaller, and vice versa.

Algorithm implementation: Note that our objective function is convex on $[W^1, \ldots, W^K]$, but not smooth due to the existence of the ℓ_1 norm regularization term. To solve the optimization problem, we use a random block coordinate descent (RCD) algorithm, summarized below (if numerical instability is encountered, a smooth version of ℓ_1 norm can be used to approximate the gradient).

where $W(t)$ indicates the incidence matrix W at iteration t and γ_k is chosen to be the largest eigenvalue of ${X^k}^T X^k$ for each cell state. The hyper-parameter λ_1 controls the sparsity of the W^k matrices, whereas λ_2

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Algorithm 1 RCD for SimiC

Input: X^k, Y^k, γ_k, for $k \in [1, \ldots, K]$, λ_1, λ_2, number of iterations T
Initialization: $W^k(0) = \mathbf{0}, \forall k \in [1, \ldots, K]$
for t in $[1, \ldots, T]$ **do**
  Randomly pick a cell state $l \in [1, \ldots, K]$
  $\widetilde{\nabla}_{W^l} f = \frac{2}{m} X^{l^T}(X^l W^l(t-1) - Y^l) + \lambda_1 \text{sgn}(W^l(t-1))$
  $\nabla_{W^l} f = \widetilde{\nabla}_{W^l} f + 2\lambda_2 \left(2W^l(t-1) - W^{l+1}(t-1) + W^{l-1}(t-1)\right)$
  $W^l(t) = W^l(t-1) - \frac{1}{\gamma_l} \nabla_{W^l} f$
end for

controls the inter-matrix dependencies and thus can be tuned based on the underlying structure of the cell states. We choose them from the polynomial set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. Specifically, λ_1 and λ_2 are chosen using 5-fold cross validation, where the dataset is randomly split into training and validation sets in a proportion of $80:20$. We evaluate the approximation performance by the average adjusted R-square value on the left out validation sets. This process is done in each pair of combination of λ_1 and λ_2 from the pre-selected polynomial set and the ones resulting in the highest average adjusted R-square are chosen as the final values of λ_1 and λ_2. The output of Algorithm 1 is a group of incidence matrices $[W^1, \ldots, W^K]$, each corresponding to the GRN of one state. The W matrices all have the same dimension $m \times n$ and the same column/row index for the corresponding driver-target pairs, i.e., entry $W_{i,j}$ is the weight between the i-th driver and the j-th target genes.

Evaluation metrics

Unlike other methods, SimiC produces a group of GRNs, one per cell state, in which a similarity constraint is imposed across progressive cell states. To further analyze the underlying structure and dynamics of the GRNs, we propose here several evaluation metrics that capture the importance of the driver genes in a given state, as well as their change of role across different states. As part of the analysis, we also look at the importance of the driver genes at the cell level by taking into account the graph structure and the cell's expression profile.

Importance dynamics of driver genes across cell states

For simplicity, we first consider the case in which we only infer one incidence matrix $W \in \mathbb{R}^{m \times n}$, with m and n denoting the number of driver genes and target genes, respectively. Hence the i-th column of W represents the regulatory importance of the different driver genes over target gene i, defined by the absolute value of the values in W_i. To filter out low weights, we perform a z-test on each W_i and select those with p-value smaller than 0.05. Then, for each target gene (column of W), we rank the significant edges by their absolute z-score in ascending order, and assign each one an importance score according to their position in the ranking (note that the importance score for the discarded edges is zero). That is, the more significant edges (higher ranked) will have a larger importance score. This process produces an importance matrix $M \in \mathbb{R}^{m \times n}$.

Since SimiC produces K incidence matrices $[W^1, \ldots, W^K]$, we infer the corresponding K importance matrices $[M^1, \ldots, M^K]$, with $M^k_{i,j}$ representing the importance score of driver gene i for target j at state k. After applying the threshold described above, we can analyze the change of activity of each driver-target pair from state k to $k+1$.

In addition, from the importance matrices $[M^1, \ldots, M^K]$, we also compute the average importance score of each driver gene across all target genes. This corresponds to taking the mean of each row in M^k, yielding the matrix $\bar{M} \in \mathbb{R}^{m \times K}$, where column $\bar{M}_k$ represents the average importance score of each driver gene at cell state k. See Fig. 2 for an illustration. We then use a local linear regression to fit the state-importance curve, where the regression coefficients are weighted by a bell-shape exponential function. We select from this set of curves those exhibiting the largest variances. This gives us an estimation of how the role of a driver gene changes across different states. Hence, this analysis provides a glimpse of the network dynamics, as we can observe the change of importance of each driver gene across states. Note that the scoring features we use are robust to noise and there is no need for normalization, as the scale is the same for all driver-target pairs and different cell states.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

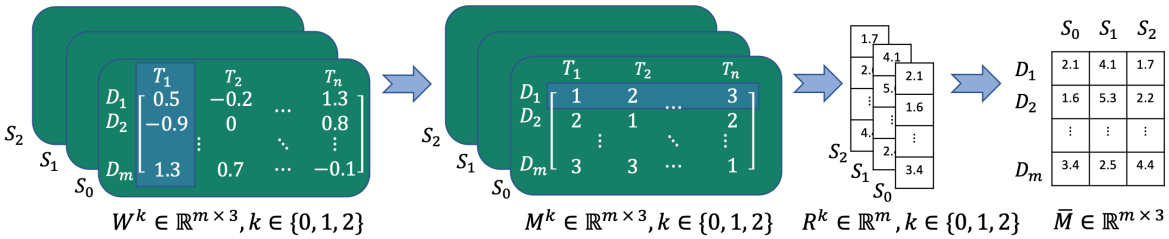


Fig. 2: Workflow of driver importance. For illustration, we use a toy example with 3 states $\{S_0, S_1, S_2\}$. W^k is the incidence matrix for state k; $D_i,\ i \in \{1,\ldots,m\}$, are the driver genes; $T_j,\ j \in \{1,\ldots,n\}$, are the target genes; M^k is the importance matrix after the z-test for state k; R^k is the average importance vector for state k; and by concatenating all R^ks together we have the importance dynamics matrix $\bar{M}$.

wAUC score

The importance dynamics used in the previous section look at the group effect of driver genes while transitioning from one state to another. To better understand how the activity changes in the cell population, we propose to use a new metric: the weighted Area Under the Curve (wAUC) that is based on the AUC score from [1]. In what follows, and following the nomenclature of [1], we denote as "the regulon of a driver gene" to its connected targets. However, contrary to [1], we take into account the weights of the edges between the target genes and the driver gene and coin it *weighted regulon*.

Hence, for each cell, the wAUC score serves to quantify the relative activity of a given weighted regulon with respect to the expression of all target genes in the cell at hand. Intuitively, for a given driver gene, if the overall expression of its weighted regulon in a cell is high and the weights between it and the target genes are large, then the driver gene in question may have a large influence in the expression profile of the cell. This is what the wAUC score tries to capture.

Similar to [1], we compute one wAUC score per driver gene and per cell. The input to the wAUC score workflow is the expression profile of the cell and the weighted regulon, and the output is the relative activity of this gene set in the cell. Note that in the objective of Eq. (3), the weights are directly comparable across drivers for a given target gene, but not across targets for a given driver since the expression for different targets may vary. Hence, we first normalize the incidence matrices W for each driver gene (i.e., each row) by dividing each element by the norm of the corresponding target expression. Next, for each cell, we order the normalized weights by the expression of the target genes in the cell. We denote the resulting normalized and ordered weight matrix by $\widehat{W}$. We then compute the cumulative sum of the ordered weights. The wAUC score of the considered driver gene in the cell is then defined by the normalized area under the cumulative sum curve, which can be expressed as:

$$\text{wAUC (driver gene } i) = \frac{\sum_{t=1}^{T}\left(\sum_{n\leqslant t}\frac{1}{T}\hat{W}_{i,n}\right)}{\sum_{t=1}^{T}\hat{W}_{i,t}},$$

where T is the number of target genes in the weighted regulon i (i.e., the number of non-zero edges of the considered TF). Note that the wAUC is a score between 0 and 1. When the larger weights are ranked higher (i.e., the corresponding target genes are highly expressed in the cell), the numerator is larger and the wAUC score will be closer to 1, which shows that the driver gene under consideration is more active in that particular cell. On the contrary, when the larger weights are ranked lower, the numerator gets smaller and the wAUC gets closer to 0. The wAUC score is hence a comprehensive measure that takes into account both the gene expression and the weighted regulon structure information. More importantly, it can be computed for every cell and every driver gene, which gives us a higher precision measure of a driver's activity within the cell population. The wAUCs are stored in a matrix $A \in \mathbb{R}^{c\times m}$, where c is the number of cells and m is the number of driver genes. We compute the wAUC matrix A^k for each cell state $k \in [1:K]$ based only on the expression of the cells in that state and the corresponding incidence matrix W^k.

Given a driver gene i of interest and the wAUC matrix for state k, A^k, we can compute the distribution of wAUCs for all cells in state k. We can repeat this process for all K states, generating a set of K distributions

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

for a given driver gene. Then, by analyzing the changes in distribution across two states, we can infer whether the driver gene in question plays a role in the state transition. On one hand, when the two distributions are "separated" (e.g., they have a non-overlapping support), the activity of the weighted regulon has shifted between states. This suggests that the driver gene may be highly correlated with the state transition. On the other hand, when the two distributions are similar, the regulon activity remains unchanged in both states. This suggests that the driver gene is less likely to have an influence in the state transition. To formally define the variation of the weighted regulon activity across multiple states, we use a combination of four different metrics based on total variation (TV). The TV of two probability measures P and Q on a countable sample space Ω is defined as:

$$\delta(P,Q) = \frac{1}{2}\sum_{\omega\in\Omega} |P(\omega) - Q(\omega)|$$

Let $\{P_1, \ldots, P_K\}$ denote the set of wAUC distributions. We compute the TV of all pairwise distributions in the group of distributions and put them in a list l_{TV}. We then consider the following 4 versions of total variation for multiple distributions:

1. Average TV: the average of all TVs in l_{TV}.
2. Median TV: the median of all TVs in l_{TV}.
3. Max TV: the maximum of all TVs in l_{TV}.
4. Minmax TV: computed as

$$\delta_{\text{minmax}}(\{P_1, \ldots, P_K\}) = \frac{1}{K}\sum_{\omega\in\Omega}\left(\max_{P_i} P_i(\omega) - \min_{P_j} P_j(\omega)\right)$$

It can be easily verified that all these metrics have values between 0 and 1, with values closer to 0 when the group of distributions is more similar to each other, and closer to 1 when the group of distributions is more divergent. The average TV is an overall measure of how divergent the group of distributions is. The median TV is more robust to outliers, especially when there are only a few pair of distributions that are close (or separated). The max TV only looks at the furthest pair of distributions so it captures how spread the group is. The minmax TV focus on the outline of all distributions jointly, and it takes a higher value when the distributions are more disjoined than when they overlap with each other. Note that when $K = 2$, all of the above metrics degenerate back to the original definition of total variation.

In an attempt to summarize all this information, we order all the driver genes by these 4 metrics, and select the top and last k driver genes from each metric as candidates. We further perform the Kolmogorov-Smirnov test on each distinct pair of distributions, and selected the TFs with a FDR adjusted p-value smaller than 0.05. The intersection of these candidates will correspond to the most variant and invariant driver genes based on wAUC distributions across different states. These sets capture the variation of driver genes from several perspectives and bring more biological insights into the dynamic process of cell state transition.

Data

Each (raw) scRNA-Seq dataset can be viewed as a matrix with dimension the number of cells that were sequenced by the number of genes that were measured, and each entry corresponds to the expression of a given gene for a particular cell. In this work, we test and validate our proposed method on the two single-cell RNA-Seq datasets described below.

Mouse hepatocyte cells: This dataset was used for the study of liver regeneration at single cell resolution and contains around 18K hepatocyte cells obtained at different time points from a mouse regenerating liver. Specifically, 2/3 partial hepatectomy was performed wherein the left lateral and median liver lobes were excised from 8-12-week-old mice. Mouse livers from adult, P14(postnatal-day 14) as well as from multiple time points (24h, 48h, and 96h) post-surgery, were perfused to isolate single liver cells and sequenced with the 10x chromium pipeline. The different sequencing experiments across time points where merged using Seurat [35], batch effects removed with BEER [46], and the linear ordering (based on pseudotime) was inferred using Monocle2 [38, 28]. Non-hepatocyte cells were identified and removed from the dataset. Specifically, two different sets of cell-state annotations were inferred for hepatocytes along the pseudotime trajectory; one with 9 transition states as inferred by Monocle2, and another simplified empirical annotation with 3 transitioning state. Finally, for this dataset, the considered driver genes were the transcription factors (TFs) obtained from the mouse TF database of [19].

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Honey bee brain cells: For this published dataset (GEO accession ID GSE130785, [37]), we compared the GRN dynamics at two different phenotypes (tissue of origin) rather than across a linear cell path. Specifically, scRNA-Seq on the whole brain (WB) and mushroom bodies (MB) of two individual honey bees was performed as previously described [37]. The honey bee WB is composed of 1,000,000 neurons, with the MB constituting roughly 1/3rd of the total cell count. Transcription factors were obtained from [32, 9] and used as driver genes. We used this dataset to explore differences in GRN activity between a cellularly diverse tissue (WB) and a specific subregion (MB).

Due to the limitations of current single-cell sequencing technologies, the raw scRNA-Seq datasets (i.e., the raw expression matrices) are often extremely sparse [40]. However, since thousands of cells are sequenced in a single experiment, it is possible to impute the missing expression values by using combined information from all the sequenced cells. Thus, SimiC uses imputed data for the inference of GRNs. In particular, we use MAGIC [41] to impute the considered datasets.

Results and Discussion

The proposed method SimiC and the wAUC score analysis are provided both as a Python package and as an R API. In addition, we provide a Jupyter notebook to run the method as well as the whole analysis pipeline. All experiments were run on a Linux machine with 80 cores and 504 GB of memory, Python version 3.6.5 and the requirements specified in the GitHub repository. In all experiments, we choose the top 200 TFs with largest MAD values from the given TF set, and the top $1,000$ target genes with largest variance. Finally, the enumeration of the states shown in the results is always sorted by the provided linear ordering.

Transcription factor dynamics of hepatocypte cells in liver regeneration

Importance of TFs: We start by analyzing the importance dynamics of the considered TFs across the 9 states inferred by Monocle on the mouse hepatocyte cells (see Supplementary Data 1 on the code repository). To that end, we apply SimiC to obtain the corresponding incidence matrices. Recall that for each state k we have a TF importance matrix $M^k \in \mathbb{R}^{m \times n}$. Fig. 3 shows in red, for each target gene (column in the matrix), the significant TFs after the z-test based on their regulatory weight inferred by SimiC. To illustrate the importance dynamics of TFs across cell states, we fit the values of a given TF (i.e., a row of the matrix $\overline{M}$) via a local regression function as defined in the previous section. For ease of illustration, we select only some of the TFs highlighted in Fig. 3 that serve as representatives of the different observed behaviors (Fig. 4).

Among the list of TFs found in the importance features, we found a well-known driver gene for liver regeneration: ID3 [36]. ID3 gene encodes protein inhibitor of DNA binding 3 and has been implicated in liver regeneration due to its involvement in progression to S-Phase from G1 phase of cell cycle and thereby promoting liver regeneration. As shown in Fig. 4, its importance score increases rapidly in early stages, and stays high across all regeneration states. This is reflected also in Fig. 3, where we observe a variation on importance across the different cell states for TF ID3. Similarly, TF YBX1, which is known to be upregulated during liver regeneration [10], increases its importance as the liver regenerates.

Other TFs, such as CEBPA, CEBPB and JUND, exhibit a constant importance across all cell states (see Fig. 4), and no large variation for them is found between cell states. This suggests that they might play an important role across all states of liver regeneration. JUND, for example, has been shown to play a protective role, reducing I/R injury to the liver by suppressing acute transcriptional activation of AP-1 [22]. Similarly, CEBPA and CEBPB are known to play active roles in both initiation and termination phases of liver regeneration [17, 18, 2].

wAUC score analysis: Next we analyze the cell dynamics by looking at the wAUC scores. For better illustration, we further group the 9 states inferred by Monocle into 3 meta states by combining states with fewer cells and similar expression profiles. In particular, we map states $1, 2, 3$ into a new state 0, state 4 is renamed as state 1, and states $5, 6, 7, 8, 9$ are mapped into a new state 2.

For each TF and each cell state we compute a distribution of wAUC scores across its cells, yielding a total of 3 different wAUC histograms for each TF (see for example Fig. 5). Using the total variance (TV) metrics, we can further order the TFs by their histogram discrepancy. A shift of histograms indicates a change in the relevance of the weighted regulon across states. For example, a shift towards higher values of the wAUC histogram from State 0 to 1 indicates that the weighted regulon (and, hence, the associated TF) is overall more active in the latter state than the former. Fig. 5 shows the distributions obtained for driver genes CEBPA,

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

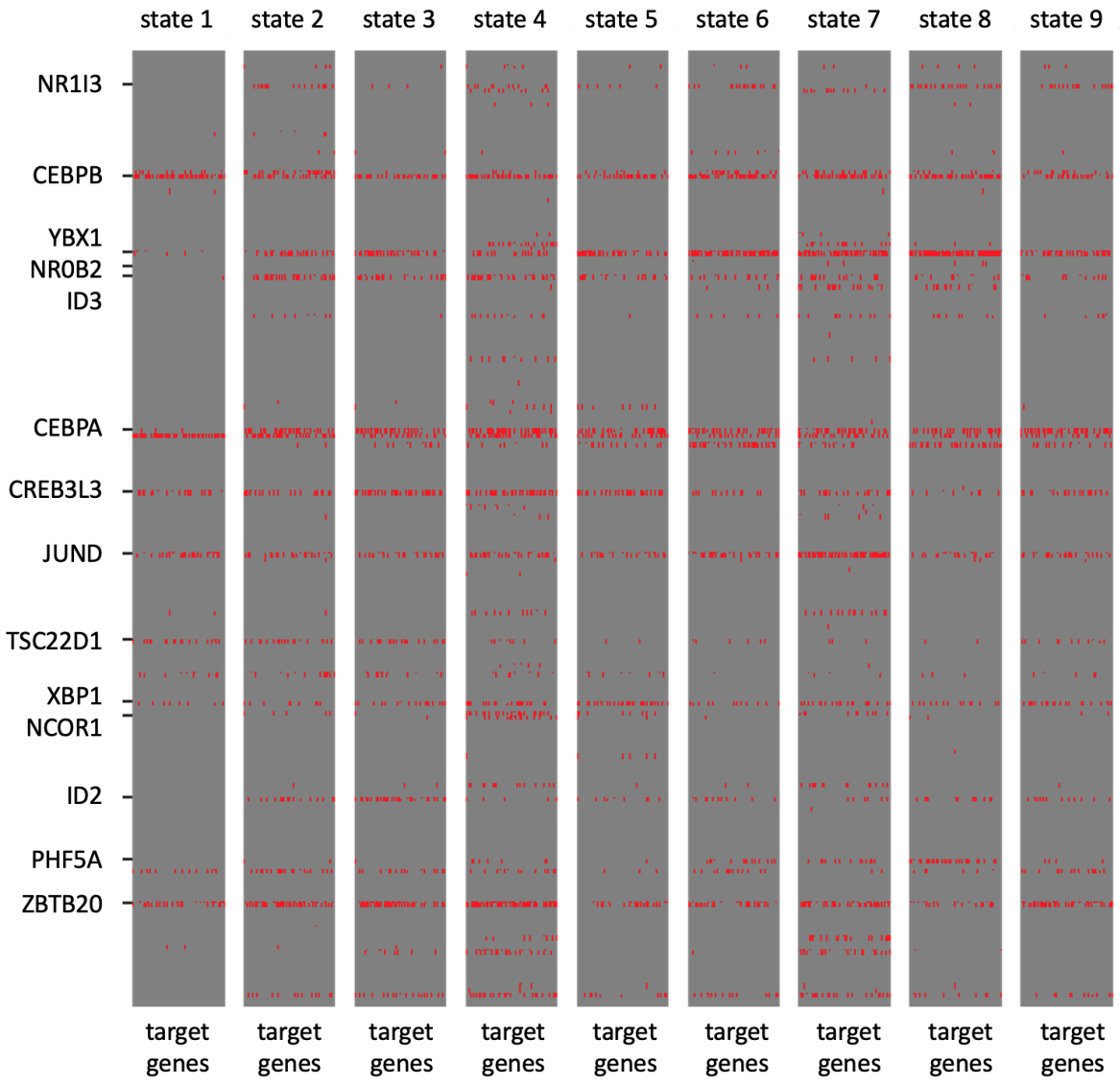


Fig. 3: TF importance matrices M for the 9 states of the Hepatocyte data. The highlighted TFs correspond to the top 5 TFs in each state (i.e., those with the highest score). We also include NCOR1 and NR0B2 as they are mentioned in the discussion.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

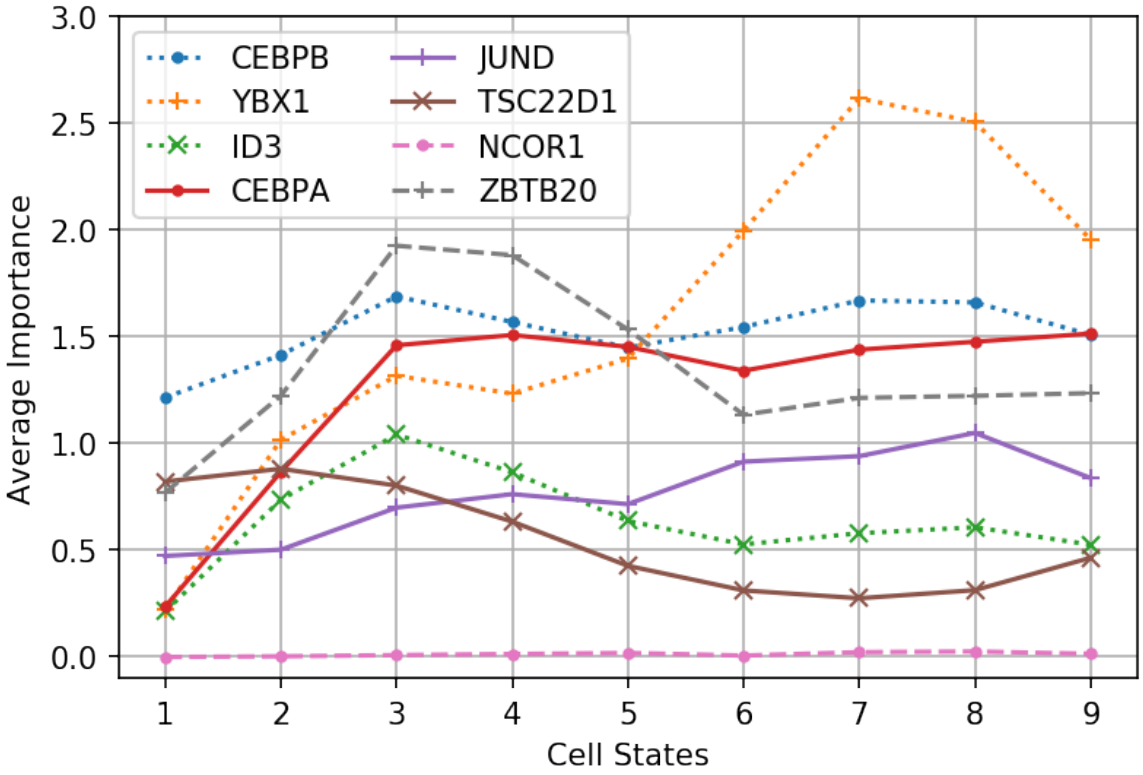


Fig. 4: Importance dynamics across states of some of the TFs highlighted in Fig. 3. See Supplementary for additional plots.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

ID3, NCOR1, and YBX1. We selected these 4 as their wAUC distributions are a good representation of the different behaviours we observe. In addition, they serve to highlight the complexity of the GRNs and the importance of analyzing them from different perspectives to further understand the dynamics involved. In particular, the wAUC score provides additional information to further understand the change of driver activity across states. For example, TF CEBPA has one of the smallest TV values and could have been undetected, even though it plays a key role in all cell states (Fig. 4). On the contrary, TF NCOR1 exhibits a clear shift on the support of the wAUC distributions, which suggests that it becomes more active in cells from state 2. However, it goes undetected if we only focus on the TF importance dynamics across states (Fig. 4). The reason for these different behaviors is that the importance score looks only at the incidence matrices, specifically, at how the relative weight for a TF (with respect to the other TFs) varies across different target genes. However, the wAUC score normalizes the weights for a given TF, extracts the important target genes, and then looks at the expression of those targets cell by cell. As such, the wAUC score captures better the cell dynamics across states, whereas the importance score highlights those TFs that are connected to a large portion of the target genes in the obtained GRNs with a large weight.

We can further investigate the results by looking at the UMAP plot of the imputed expression matrix, and color code each cell by its corresponding weighted regulon wAUC value. An example of the color coded UMAP plots for TFs ID3, NCOR1, and YBX1, as well as the true cell state labels are shown in Fig. 6. TF CEBPA is omitted here since its wAUC histograms do not show much variation across states, resulting in a more uniformly color-coded UMAP plot. As expected, the UMAP plot for TF NCOR1 resembles that of the true label, due to the clear shift obtained in the wAUC distributions. Also, due to the concentration of wAUC values of cells in state 0 for TF YBX1, we observe a uniform color across cells from state 0 in the corresponding UMAP plot. Finally, we refer to the Supplementary Data 2 and 3 on the code repository for the wAUC histograms and UMAP plots for all considered TFs.

Intersection of the set of targets in different states: wAUC score captures the regulon activity by considering both the subgraph structure and the gene expression values. However, we have multiple states, each with a different regulon subgraph structure, used to generate the wAUC score. The previously shown wAUC results are based on all considered targets genes, i.e., the considered regulons consist of the edges from the TF to all target genes (note however that some of these edges may have a weight equal to zero). Different states may therefore have different active links (non-zero edges) between the TFs and the target genes.

We now investigate the wAUC scores and corresponding distributions when considering only the intersection of active target genes in all states for a considered TF. This forces all subgraph topologies to be the same, and allows us to focus on the effect of the weight differences. In addition, we can analyze whether the subgraph topology or the weights play a more important role in the weighted regulon activity. To find the set of active targets for each TF in each state, we perform a two-tailed z-test on the learned incidence matrices $\{W^1, \ldots, W^K\}$. For a given TF and state, we then consider as active target genes those whose associated weights have a p-value smaller than 0.1. As an example, we focus on TF NCOR1, which has been shown to have a role in liver regeneration in mice [25]. We refer to Supplementary Data 4 and 5 on the code repository for all other considered TFs. We first show the Venn diagram of the intersection of active target gene sets for the different states (Fig. 7). From the Venn diagram we observe that the number of active target genes decreases significantly from state to state, and the intersection across all three states is composed of only 10 genes. When looking at the UMAP plots, we see that the result when considering all target genes (Fig. 6c) captures the cell states structure, while maintaining a smaller range of wAUC scores $(0.35, 0.45)$ than when compared to the wAUC scores computed only with the intersected targets (Fig. 7b). These results suggest that in this case the weighted regulon subgraph structure (including the non-intersecting targets) plays a more important role in defining the regulon activity than the change of weights across different states when the considered subgraphs are the same (i.e., when only the intersecting targets are considered).

Comparison with other methods

We first compare SimiC to the state-of-the-art methods SCENIC [1], and SINCERITIES [26]. Note that, to the best of our knowledge, the closest method to SimiC is SCIMITAR [12]. However, it was left out of the analysis as it cannot be run on over 10K cells (an "out of memory" error is obtained). Contrary to SimiC, SCENIC and SINCERITIES do not provide the framework to study the dynamics of the GRNs across cell states. Specifically, while SINCERITIES requires an input very similar to that of our proposed method SimiC, it generates only a unique (and small) GRN for all the cell-states (i.e., the output is a single GRN). On the other hand, SCENIC has an AUC measure similar to that of SimiC, but it is done in an uniform weight

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

setting (i.e., the edges of the corresponding subgraph do not have a weight associated to them). In addition, SCENIC only takes a single cell-state (input expression matrix) as input, lacking the capability of exploiting inter-cell-state information.

For the comparison, since neither of these methods contemplates the TF importance measure implemented in SimiC, we focus mainly on the AUC/wAUC analysis. For a fair comparison, we apply SCENIC independently to each state and then generate the corresponding AUC matrices. For SINCERITIES, an uniform AUC matrix is computed using the generated GRN. Finally, the different methods generate GRNs for a different set of driver genes (TFs). For example, in SCENIC, genes such as NCOR1 and ID3 are filtered out. Hence, for the comparison with SCENIC and SINCERITIES, we focus on the driver gene JUND mentioned before, which is present in all of them. For the AUC histograms and UMAP plots of all other TFs we refer the reder to Supplementary Data 6 on the code repository.

SimiC vs SCENIC: Fig. 8a shows the AUC/wAUC histograms and the corresponding UMAP plots for SCENIC and SimiC. In SCENIC, the histograms for states 0 and 1 are similar, with both histograms concentrated in the same values and having a clear peak. In SimiC, on the contrary, there is a clear distinction between the peak obtained in state 0 and the other states. The higher peak value in state 0 also shows that the TF JUND is more active in early stages and promotes hepatocyte proliferation, which is confirmed in the existing literature [31, 5]. Such separation can also be easily seen in the wAUC colored UMAP plots, as SCENIC has only a few outliers highlighted in the cells from state 2, whereas SimiC shows a different pattern and captures state 0 nicely.

SimiC vs SINCERITIES: Fig. 8b shows the AUC histogram and the corresponding UMAP plot obtained with SINCERITIES for TF JUND. As mentioned before, SINCERITIES utilizes the state ordering information (similar to SimiC), but it only outputs a single GRN. Hence, no GRN dynamics can be inferred from the obtained single histogram. We observe that the AUC histogram resembles a Gaussian distribution with a much larger range (support) than SCENIC and SimiC (Fig. 8a). In addition, cell states can not be differentiated when plotted in the UMAP plot when colored by their AUC scores; and highly active cells (those with high AUC values) are spread out across different states. Thus, SINCERITIES fails to capture the JUND regulon activity in the different cell states. One reason may be the construction of a single GRN instead of one per cell state.

SimiC vs LASSO: We now show the importance of the similarity constraint imposed by SimiC. In particular, we analyze the wAUC scores using the incidence matrices obtained with (SimiC) and without (LASSO) the similarity constraint. Note that in LASSO, the GRN of each state is inferred independently (i.e., $\lambda_2 = 0$ in Eq. (3)). We first focus on gene NCOR1, which exhibited a shift in the wAUC histograms obtained with SimiC (Fig. 5c). The wAUC histograms for LASSO, as well as the corresponding UMAP plot, are shown in Fig. 8c. As it can be observed, the shift on distributions is lost when LASSO is used (i.e., when there is no similarity constraint and the GRNs are inferred independently for each state). As such, the UMAP plot color coded with the wAUC values obtained with LASSO fails to capture the different states. On the contrary, the UMAP plot from SimiC separates state 2 from the other two, showing a higher regulon activity for this state (Fig. 6c).

Another example is shown for TF NR0B2, which is also found to play a important role in liver regeneration [44]. Fig. 8d shows the wAUC histograms and corresponding UMAP plots for SimiC and LASSO. From the wAUC histograms, we can observe that both methods capture a similar behaviour for states 0 and 1, but a different one for state 2. In particular, SimiC shows a decreased in activity for this state, whereas LASSO shows an increased of the regulon activity. Yet, in the UMAP plot, SimiC captures state 2 more accurately than LASSO.

Running time and memory consumption: We use the results obtained with the hepatocyte dataset to illustrate the complexity of each of the considered methods. The full pipeline of SimiC took approximately 38 hours, including GRNs inference (40 minutes) and the computation of the wAUC scores (37 hours). The full pipeline of SCENIC took instead approximately 45 hours of CPU time, and SINCERITIES employed 1 hour for the network inference part (i.e., without the computation of the AUC matrices). Finally, note that SCENIC also provides an optimized Python package which utilizes all the cores and runs in about 1 hour. Our code could be optimized in a similar fashion, boosting the running time. The memory consumption of all methods was under 32GB in all performed experiments.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Transcription factor dynamics of the honey bee whole brain and a brain subregion.

To further validate the proposed method, we analysed the transcriptional dynamics between two brain tissues samples from the honey bee (*Apis mellifera*). Specifically, for this data SimiC infers GRNs across two brain samples rather than on an ordered set of cell states. We aim to find which TFs are specific to (or more relevant in) the honey bee mushroom body (MB) with respect to the whole brain (WB), using published scRNA-Seq data [37], and analyze their associated regulatory activity. Additionally, since the honey bee is not a model organism, any GRN inference method that requires side information, such as [1, 45] would not work on these data, as side information on the honey bee is very limited.

As before, we start with the TF importance dynamics between MB and WB. As shown in Fig. 9a, several TFs, such as LOC413474 and LOC409483 have a high importance score for WB, but not MB. LOC413474 has been previously identified as driver genes of GRNs in WB tissue [29, 9, 14], and LOC409483 was shown to be expressed throughout the entire brain, including MB, in another insect species [33]. Moreover, [32] showed that LOC100576186 and LOC412916 are differentially expressed in MB following caregiving, which may be driving the difference in importance observed between MB and WB (Fig. 9a).

We next focus on the cell dynamics for the above-mentioned TFs. We focus on two particular cases which show two different TF dynamics across cells. First, the larger support of the wAUC histograms of WB in comparison to MB aligns to the larger cell heterogeneity found on WB tissue (Fig. 9c and 9d, see also Supplementary Data 8 available in the code repository for the wAUC histogram and UMAP plot for all TFs). In addition, we can observe that the UMAP plots, when colored based on the wAUC scores, show a progressive gradient across the cells. This implies that the wAUC score captures the cells progressions within one tissue (Fig. 9). Finally, note that the cells in the UMAP plot are not separable, however, for some TFs, such as LOC409483 (Fig. 9d, other examples available in Supplementary Data 8), one could perfectly separate both MB and WB cells by appropriately thresholding the wAUC score based on the histograms.

Conclusion

We have introduced SimiC, a novel single-cell gene regulatory network inference (GRN) method. Contrary to state-of-the-art methods, SimiC jointly infers a GRN for each cell state across a differentiation path by imposing a similarity constraint across GRNs of contiguous states. Hence, SimiC is able to model the transition regulatory dynamics via the inferred GRNs. To help analyze and interpret the inferred GRNs, as well as provide insights into the dynamics of the driver and target pairs, SimiC includes a set of evaluation metrics. In particular, SimiC identifies, per cell state, the driver genes that play an important role in regulating the target genes. In addition, for a given driver, SimiC computes a histogram (one per state) that captures its activity in the cells. We show on a hepatocyte and a honey bee single-cell datasets that SimiC is able to capture the state dynamics, identifying in both cases known important TFs. Comparisons with existing methods SCENIC and SINCERITIES highlight the importance of the added similarity constraint. Finally, SimiC does not use any species-specific side information, and hence it can be applied to single-cell data of any species.

Funding

M.H is supported by the EC under the H2020 Marie Curie Individual Fellowship Action (Grant Agreement N° 898356). A.K. is supported by grants from the US National Institute of Health (R01HL126845, R01AA010154), Muscular Dystrophy Association (MDA514335), Planning Grant Award from the Cancer Center at Illinois, and Beckman Fellowship from the Center for Advanced Study at the University of Illinois Urbana-Champaign. U.V.C. is supported by the Herbert E. Carter fellowship in Biochemistry, UIUC. S.B is supported by the NIH Tissue microenvironment training program (T32-EB019944).

References

[1] S. Aibar, C. B. González-Blas, T. Moerman, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, J. van den Oord, et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[2] S. Bangru and A. Kalsotra. Cellular and molecular basis of liver regeneration. *Seminars in Cell Developmental Biology*, 100(4):74–87, 2020.

[3] A.-L. Barabasi and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

[4] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. Reverse engineering of regulatory networks in human b cells. *Nature genetics*, 37(4):382–390, 2005.

[5] A. Behrens, M. Sibilia, J.-P. David, U. Möhle-Steinlein, F. Tronche, G. Schütz, and E. F. Wagner. Impaired postnatal hepatocyte proliferation and liver regeneration in mice lacking c-jun in the liver. *The EMBO journal*, 21(7):1782–1790, 2002.

[6] A. Califano and M. J. Alvarez. The recurrent architecture of tumour initiation, progression and drug sensitivity. *Nature reviews Cancer*, 17(2):116, 2017.

[7] M. Champion, K. Brennan, T. Croonenborghs, A. J. Gentles, N. Pochet, and O. Gevaert. Module analysis captures pancancer genetically and epigenetically deregulated cancer driver genes for smoking and antiviral response. *EBioMedicine*, 27:156–166, 2018.

[8] T. E. Chan, M. P. Stumpf, and A. C. Babtie. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems*, 5(3):251–267, 2017.

[9] S. Chandrasekaran, S. A. Ament, J. A. Eddy, S. L. Rodriguez-Zas, B. R. Schatz, N. D. Price, and G. E. Robinson. Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. *Proceedings of the National Academy of Sciences*, 108(44):18020–18025, 2011.

[10] H.-M. Chao, H.-X. Huang, P.-H. Chang, K.-C. Tseng, A. Miyajima, and E. Chern. Y-box binding protein-1 promotes hepatocellular carcinoma-initiating cell progression and tumorigenesis via wnt/β-catenin pathway. *Oncotarget*, 8(2):2604, 2017.

[11] S. Chen and J. C. Mar. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC bioinformatics*, 2018.

[12] P. Cordero and J. M. Stuart. Tracing co-regulatory network dynamics in noisy, single-cell transcriptome trajectories. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*, pages 576–587. World Scientific, 2017.

[13] M. W. Fiers, L. Minnoye, S. Aibar, C. Bravo González-Blas, Z. Kalender Atak, and S. Aerts. Mapping gene regulatory networks from single-cell omics data. *Briefings in functional genomics*, 17(4):246–254, 2018.

[14] A. R. Hamilton, I. M. Traniello, A. M. Ray, A. S. Caldwell, S. A. Wickline, and G. E. Robinson. Division of labor in honey bees is associated with transcriptional regulatory plasticity in the brain. *Journal of Experimental Biology*, 222(14):jeb200196, 2019.

[15] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[16] M. Hernaez, C. Blatti, and O. Gevaert. Comparison of single and module-based methods for modeling gene regulatory networks. *Bioinformatics*, 36(2):558–567, 2020.

[17] J. S. Jakobsen, J. Waage, N. Rapin, H. C. Bisgaard, F. S. Larsen, and B. T. Porse. Temporal mapping of cebpa and cebpb binding during liver regeneration reveals dynamic occupancy and specific regulatory codes for homeostatic and cell cycle gene batteries. *Genome research*, 23(4):592–603, 2013.

[18] J. Jin, I.-H. Hong, K. Lewis, P. Iakova, M. Breaux, Y. Jiang, E. Sullivan, N. Jawanmardi, L. Timchenko, and N. A. Timchenko. Cooperation of c/ebp family proteins and chromatin remodeling proteins is essential for termination of liver regeneration. *Hepatology*, 61(1):315–325, 2015.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[19] M. Kanamori, H. Konno, N. Osato, J. Kawai, Y. Hayashizaki, and H. Suzuki. A genome-wide and nonredundant mouse transcription factor database. *Biochemical and biophysical research communications*, 322(3):787–793, 2004.

[20] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, et al. Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483–486, 2017.

[21] S. Linnarsson and S. A. Teichmann. Single-cell genomics: coming of age, 2016.

[22] J. J. Marden, Y. Zhang, F. D. Oakley, W. Zhou, M. Luo, H. P. Jia, P. B. McCray, M. Yaniv, J. B. Weitzman, and J. F. Engelhardt. Jund protects the liver from ischemia/reperfusion injury by dampening ap-1 transcriptional activation. *Journal of Biological Chemistry*, 283(11):6687–6695, 2008.

[23] H. Matsumoto, H. Kiryu, C. Furusawa, M. S. Ko, S. B. Ko, N. Gouda, T. Hayashi, and I. Nikaido. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, 33(15):2314–2321, 2017.

[24] T. Moerman, S. Aibar Santos, C. Bravo González-Blas, J. Simm, Y. Moreau, J. Aerts, and S. Aerts. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35 (12):2159–2161, 2019.

[25] Q. Ou-Yang, X.-M. Lin, Y.-J. Zhu, B. Zheng, L. Li, Y.-C. Yang, G.-J. Hou, X. Chen, G.-J. Luo, F. Huo, et al. Distinct role of nuclear receptor corepressor 1 regulated de novo fatty acids synthesis in liver regeneration and hepatocarcinogenesis in mice. *Hepatology*, 67(3):1071–1087, 2018.

[26] N. Papili Gao, S. M. Ud-Dean, O. Gandrillon, and R. Gunawan. Sincerities: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, 34(2):258–266, 2018.

[27] A. Pratapa, A. P. Jalihal, J. N. Law, A. Bharadwaj, and T. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, pages 1–8, 2020.

[28] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature methods*, 14(10):979, 2017.

[29] C. C. Rittschof, S. A. Bukhari, L. G. Sloofman, J. M. Troy, D. Caetano-Anollés, A. Cash-Ahmed, M. Kent, X. Lu, Y. O. Sanogo, P. A. Weisner, et al. Neuromolecular responses to social challenge: Common mechanisms across mouse, stickleback fish, and honey bee. *Proceedings of the national Academy of Sciences*, 111(50):17929–17934, 2014.

[30] W. Saelens, R. Cannoodt, and Y. Saeys. A comprehensive evaluation of module detection methods for gene expression data. *Nature communications*, 9(1):1–12, 2018.

[31] I. Schulien, B. Hockenjos, A. Schmitt-Graeff, M. G. Perdekamp, M. Follo, R. Thimme, and P. Hasselblatt. The transcription factor c-jun/ap-1 promotes liver fibrosis during non-alcoholic steatohepatitis by regulating osteopontin expression. *Cell Death & Differentiation*, 26(9):1688–1699, 2019.

[32] H. Y. Shpigler, M. C. Saul, E. E. Murdoch, F. Corona, A. C. Cash-Ahmed, C. H. Seward, S. Chandrasekaran, L. J. Stubbs, and G. E. Robinson. Honey bee neurogenomic responses to affiliative and agonistic social interactions. *Genes, Brain and Behavior*, 18(1):e12509, 2019.

[33] C. K. Sim, S. Perry, S. K. Tharadra, J. S. Lipsick, and A. Ray. Epigenetic regulation of olfactory receptor gene expression by the myb–muvb/dream complex. *Genes & development*, 26(22):2483–2498, 2012.

[34] A. T. Specht and J. Li. Leap: constructing gene co-expression networks for single-cell rna-sequencing data using pseudotime ordering. *Bioinformatics*, 33(5):764–766, 2017.

[35] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 2019.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[36] S. Togo, H. Makino, T. Kobayashi, T. Morita, T. Shimizu, T. Kubota, Y. Ichikawa, T. Ishikawa, Y. Okazaki, Y. Hayashizaki, et al. Mechanism of liver regeneration after partial hepatectomy using mouse cdna microarray. *Journal of hepatology*, 40(3):464–471, 2004.

[37] I. M. Traniello, S. A. Bukhari, J. Kevill, A. C. Ahmed, A. R. Hamilton, N. L. Naeger, D. C. Schroeder, and G. E. Robinson. Meta-analysis of honey bee neurogenomic response links deformed wing virus type a to precocious behavioral maturation. *Scientific Reports*, 10(1):3101, 2020. doi: 10.1038/s41598-020-59808-4. URL `https://doi.org/10.1038/s41598-020-59808-4`.

[38] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381, 2014.

[39] A. I. Vân Anh Huynh-Thu, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9), 2010.

[40] D. van Dijk, J. Nainys, R. Sharma, P. Kathail, A. J. Carr, K. R. Moon, L. Mazutis, G. Wolf, S. Krishnaswamy, and D. Pe'er. Magic: A diffusion-based imputation method reveals gene-gene interactions in single-cell rna-sequencing data. *BioRxiv*, page 111591, 2017.

[41] D. Van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R. Moon, C. L. Chaffer, D. Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 2018.

[42] A. Wagner, A. Regev, and N. Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 34(11):1145, 2016.

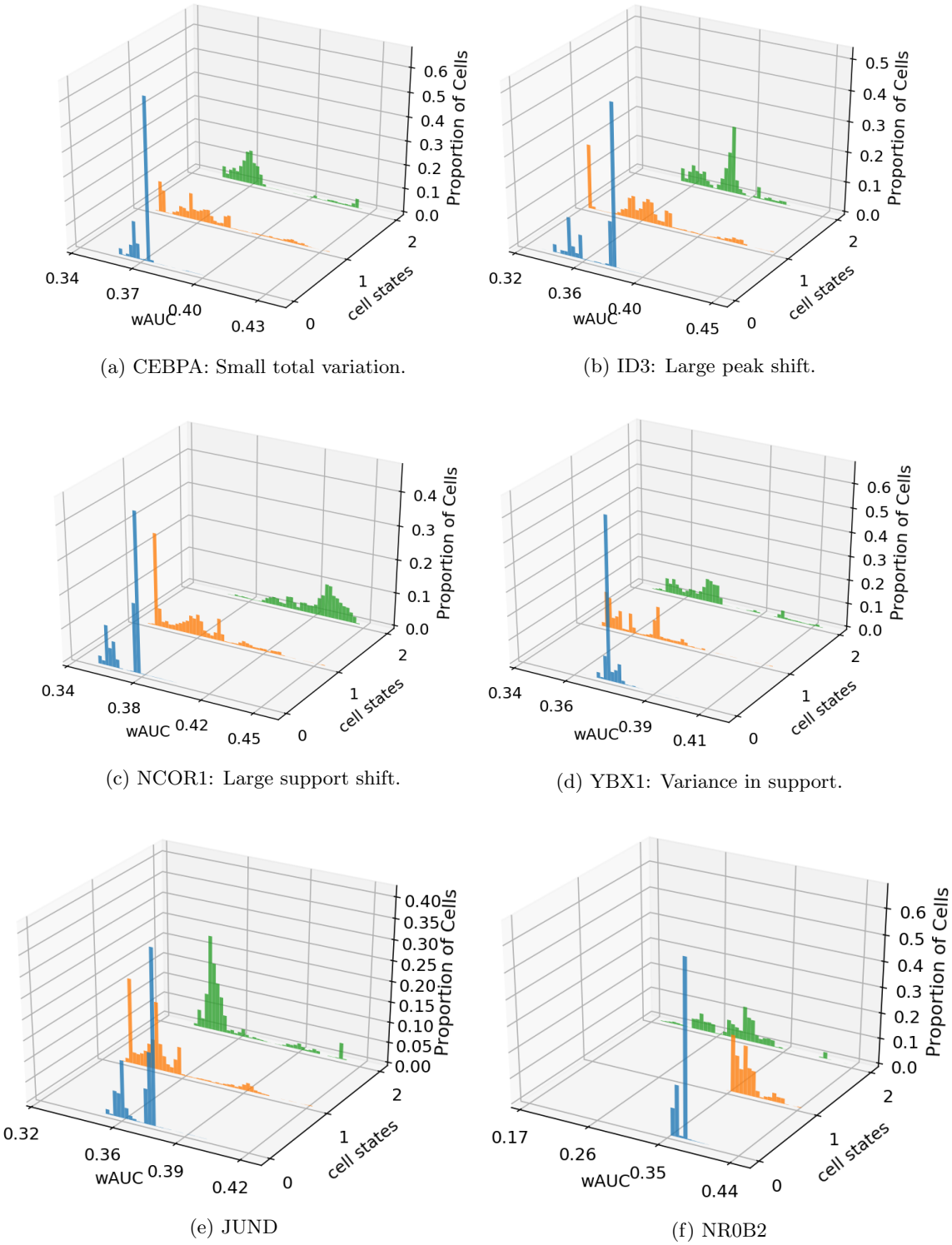
[43] S. Woodhouse, N. Piterman, C. M. Wintersteiger, B. Göttgens, and J. Fisher. Scns: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data. *BMC systems biology*, 12 (1):59, 2018.

[44] C.-S. Xu, C.-F. Chang, J.-Y. Yuan, W.-Q. Li, H.-P. Han, K.-J. Yang, L.-F. Zhao, Y.-C. Li, H.-Y. Zhang, S. Rahman, et al. Expressed genes in regenerating rat liver after partial hepatectomy. *World Journal of Gastroenterology: WJG*, 11(19):2932, 2005.

[45] Y. Yuan and Z. Bar-Joseph. Deep learning for inferring gene relationships from single-cell expression data. *Proceedings of the National Academy of Sciences*, 116(52):27151–27158, 2019.

[46] F. Zhang, Y. Wu, and W. Tian. A novel approach to remove the batch effect of single-cell data. *Cell discovery*, 5(1):1–4, 2019.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



(a) CEBPA: Small total variation.

(b) ID3: Large peak shift.

(c) NCOR1: Large support shift.

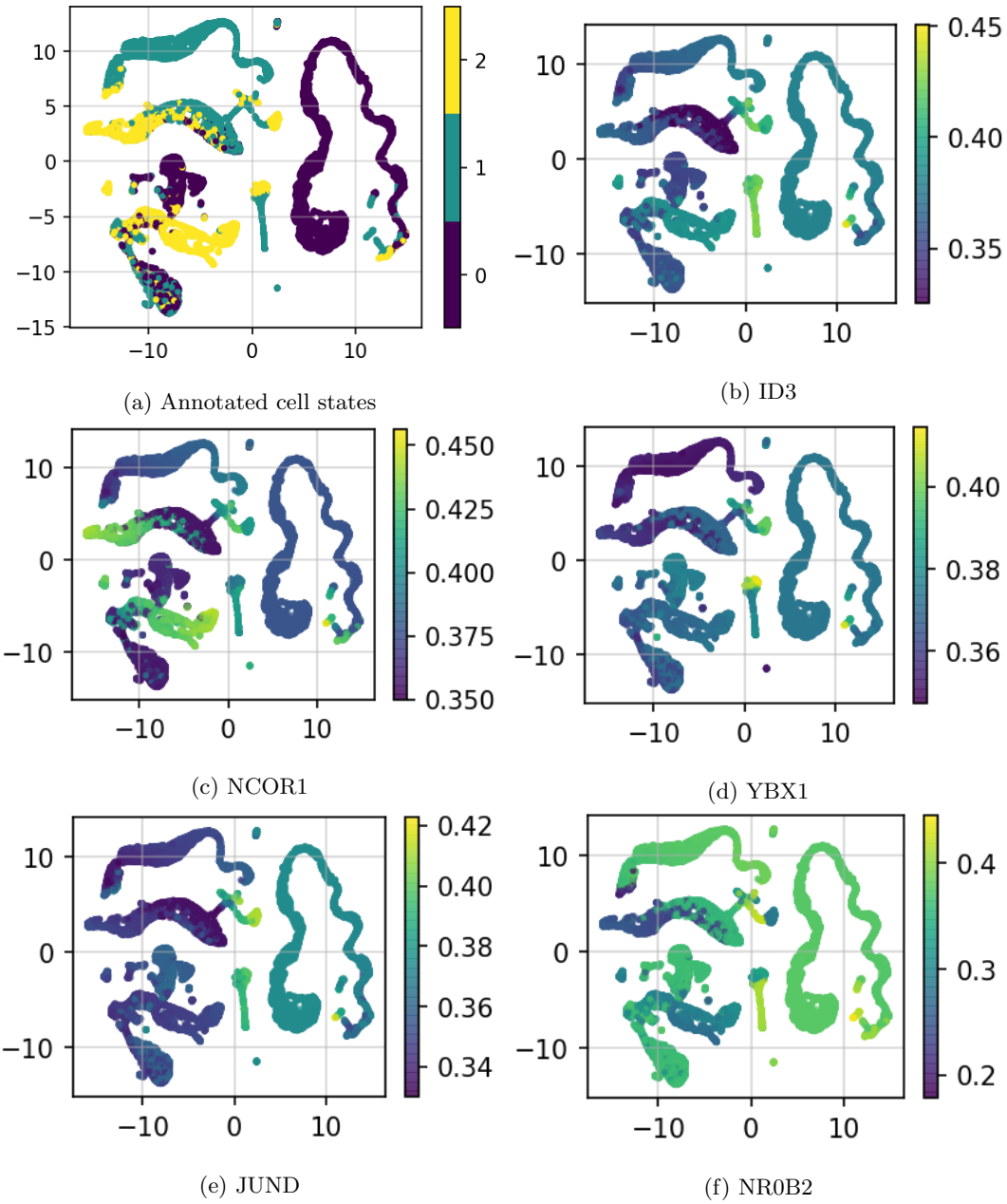
(d) YBX1: Variance in support.

(e) JUND

(f) NR0B2

Fig. 5: Examples of different wAUC histograms obtained with SimiC for different TFs.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



(a) Annotated cell states

(b) ID3

(c) NCOR1

(d) YBX1

(e) JUND

(f) NR0B2

Fig. 6: Examples of UMAP plots color-coded by the wAUC score obtained with SimiC.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

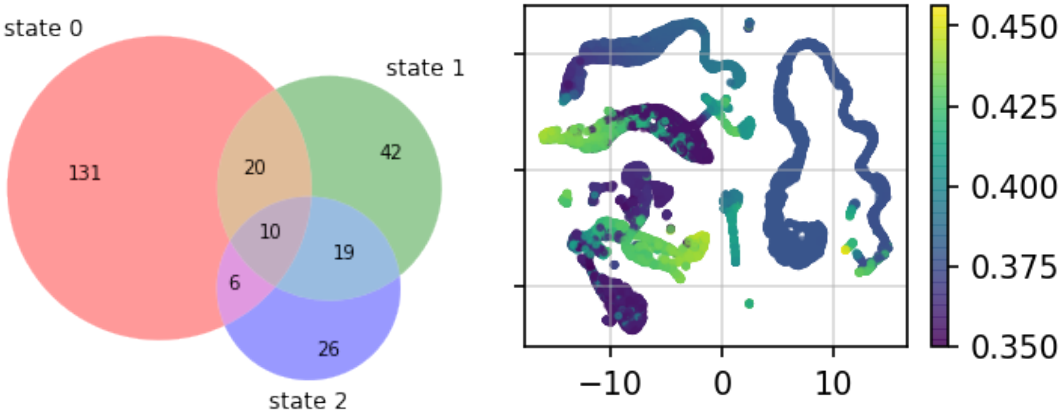
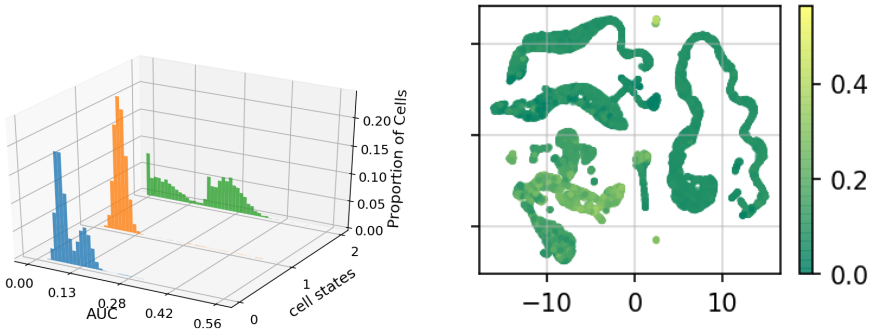
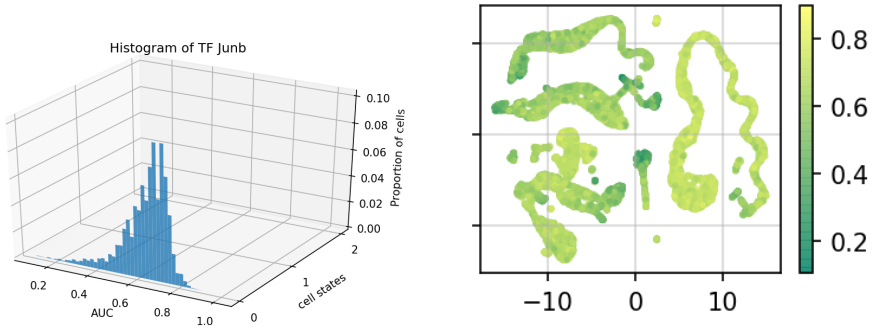


Fig. 7: Venn diagram of active target across states (left) and UMAP plot color coded by the wAUC score when considering only the intersection of active target genes for TF NCOR1.

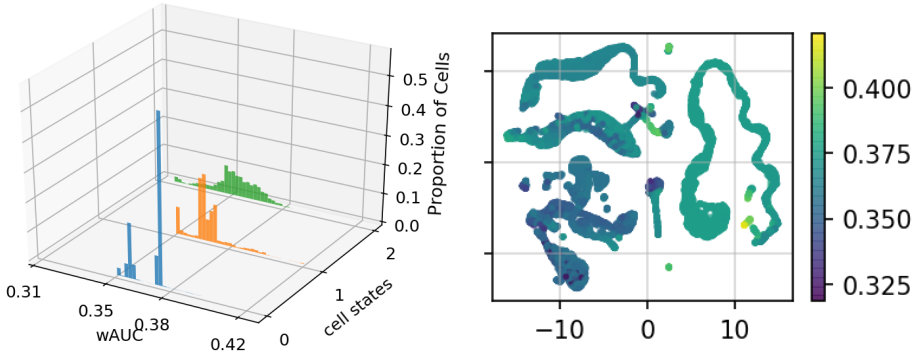
bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



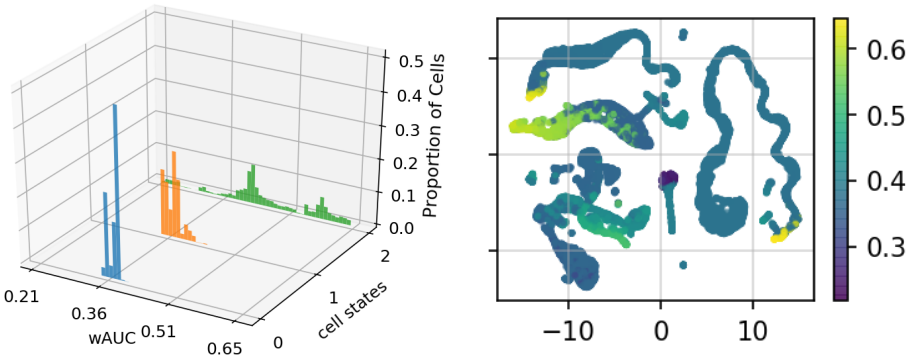
(a) SCENIC: TF JUND



(b) SINCERITIES: TF JUND



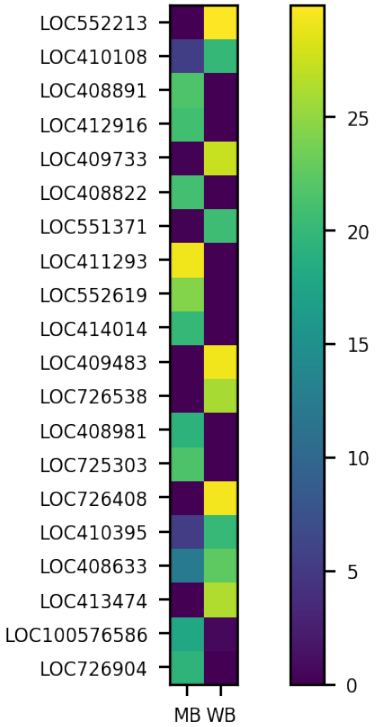
(c) LASSO: TF NCOR1



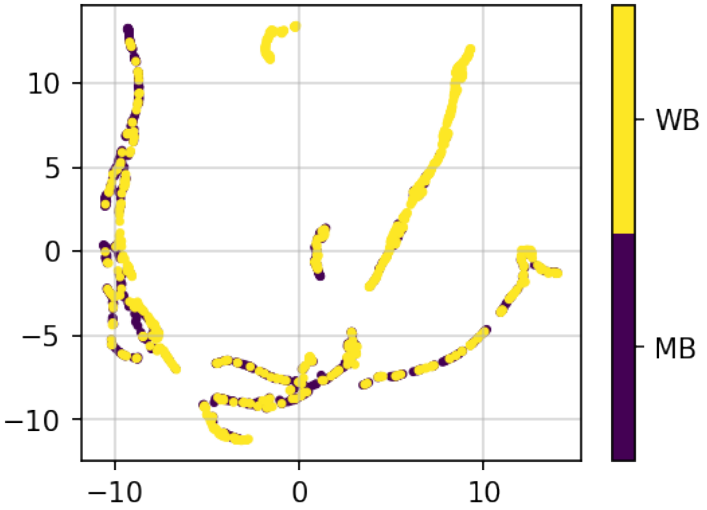
(d) LASSO: TF NR0B2

Fig. 8: wAUC score histograms (left) and the UMAP plot color coded with the wAUC values (right) of different GRN inference methods for different TFs.

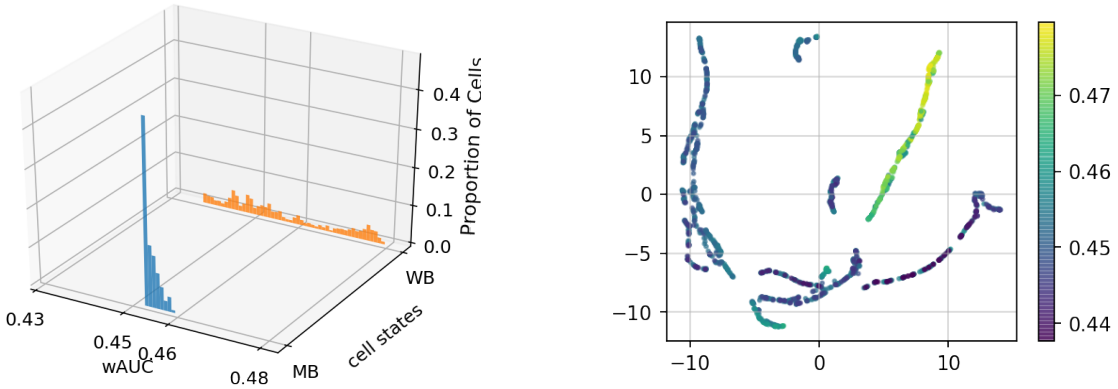
bioRxiv preprint doi: https://doi.org/10.1101/2020.04.03.023002; this version posted April 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



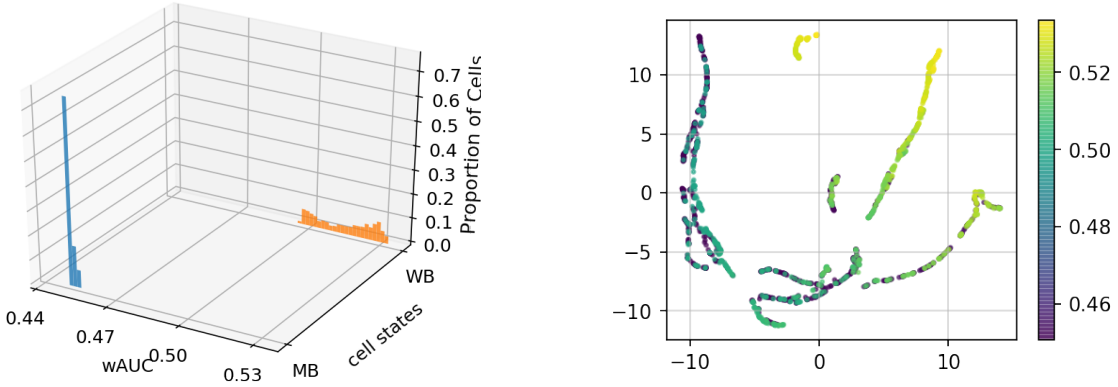
(a) Importance dynamics of TFs across MB and WB.



(b) True label of Bees dataset



(c) UMAP of expression color coded by wAUC of gene Loc413474



(d) UMAP of expression color coded by wAUC of gene Loc409483

Fig. 9: weighted AUC results of intersection of active targets.