A single-cell gene regulatory network inference method for identifying complex regulatory dynamics across cell phenotypes.

Jianhao Peng[1 †], Guillermo Serrano[2 †], Ian M. Traniello[3, 4], Maria E. Calleja-Cervantes[2, 7], Ullas V. Chembazhi[5], Sushant Bangru[5], Teresa Ezponda[7,8, 10], Juan Roberto Rodriguez-Madoz[7,8, 10], Auinash Kalsotra[3,5,6], Felipe Prosper[7,8,9,10], Idoia Ochoa[1,11 *], and Mikel Hernaez[2,3, 8,10 *]

[1]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801.
[2]Computational Biology Program, CIMA University of Navarra, Spain.
[3]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801.
[4]Neuroscience Program, University of Illinois at Urbana-Champaign, Urbana, IL 61801.
[5]Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801.
[6]Cancer Center@Illinois, University of Illinois at Urbana-Champaign, Urbana, IL 61801.
[7]Hemato-Oncology Program, CIMA University of Navarra, Spain.
[8]Centro de Investigación Biomédica en Red de Cáncer, CIBERONC, Spain.
[9]Hematology and Cell Therapy, Clinica Universidad de Navarra, Pamplona, Spain
[10]Navarra Institute for Health Research (IdiSNA), Spain.
[11]Department of Electrical Engineering (TECNUN), University of Navarra, Spain.
[†]These authors contributed equally to this work.
[*]Correspondence: idoia@illinois.edu, mhernaez@unav.es

Abstract

Single-cell RNA-Sequencing has made it possible to infer high-resolution gene regulatory networks (GRNs), providing deep biological insights by revealing regulatory interactions at single-cell resolution. However, current single-cell GRN analysis methods produce only a single GRN per input dataset, potentially missing relationships between cells from different phenotypes. To address this issue, we present SimiC, a single-cell GRN inference method that produces a GRN per phenotype while imposing a similarity constraint that forces a smooth transition between GRNs, allowing for a direct comparison between different states, treatments, or conditions. We show that jointly inferring GRNs can uncover variation in regulatory relationships across phenotypes that would have otherwise been missed. Moreover, SimiC can recapitulate complex regulatory dynamics across a range of systems, both model and non-model alike. Taken together, we establish a new approach to quantitating regulatory architectures between the GRNs of distinct cellular phenotypes, with far-reaching implications for systems biology.

Introduction

Gene regulatory networks (GRNs) infer regulatory interactions between genes, including relationships between transcription factors and their targets. With the advent of next-generation sequencing, GRN inference has become one of the most important steps in determining cellular functions (*1, 2*) and modeling different systemic behaviors (*3, 4*). In addition, GRNs have been found to be reliable surrogates/predictors of behavioral state (*5, 6*).

GRNs 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 broadly classified 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 (*7, 8*).

Until recently, most available gene expression data were derived from "bulk" RNA-Sequencing (RNA-Seq), which does not differentiate among the cellular composition of a tissue sample, and therefore only gives 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. More recently, single-cell RNA-Sequencing (scRNA-Seq) has made it possible to acquire gene expression data for individual cells in samples containing up to millions of cells (*9*). scRNA-Seq data are generally summarized as a matrix containing the expression values of genes for each sequenced cell. Available bioinformatic pipelines then group these cells by similarities in gene expression patterns, thus generating cell clusters that are predicted to represent distinct cell types that share anatomical or functional characterization (*10, 11*). However, the nature of scRNA-Seq data 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 gene expressions across samples are independent and identically distributed (i.i.d.). However, this does not apply to single-cell data, as cells in the same cluster 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 to reduce its intrinsic high sparsity caused by the dropout effect (*12, 13*), and the information on cell types, such as their pseudo-temporal ordering, can be used as side information for GRN inference (*14, 15*). For example, the ordering of cells is used to infer the regulatory relationships among genes via ordinary differential equations (*16*), correlation methods (*17*), information theoretic measures (*18*), or boolean functions (*19*). To control the dropout effect, Papili et al. (*20*) 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, whereas Yuan et al. (*21*) 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 have been used as intermediate steps in the processing of cell type inference for scRNA-Seq data, rather than as the final goal for network inference (*22*).

All the GRN inference methods discussed above produce a unique GRN per input dataset. However, related cells, such as those arising on a cell 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 phenotype (*23*). Similarly, the GRNs of identical cell types under different conditions are expected to vary, and even slight changes can be informative regarding the effects of specific cellular conditions. In this context, directly inferring GRNs independently for each cell phenotype might result in a group of divergent networks that share little in common. It would therefore be useful to add constraints on the inferred GRNs to allow for comparisons of GRN architecture across related cell types in different conditions.

We present SimiC, a GRN inference algorithm for scRNA-Seq data that takes as input single-cell imputed expression data, a list of driver genes, the cell labels (cell phenotypes) and the ordering information, and produces a GRN for each of the different phenotypes. Given the provided ordering between the cell phenotypes, SimiC adds a similarity constraint when jointly inferring the GRNs for each phenotype, ensuring a smooth transition between the corresponding GRNs.

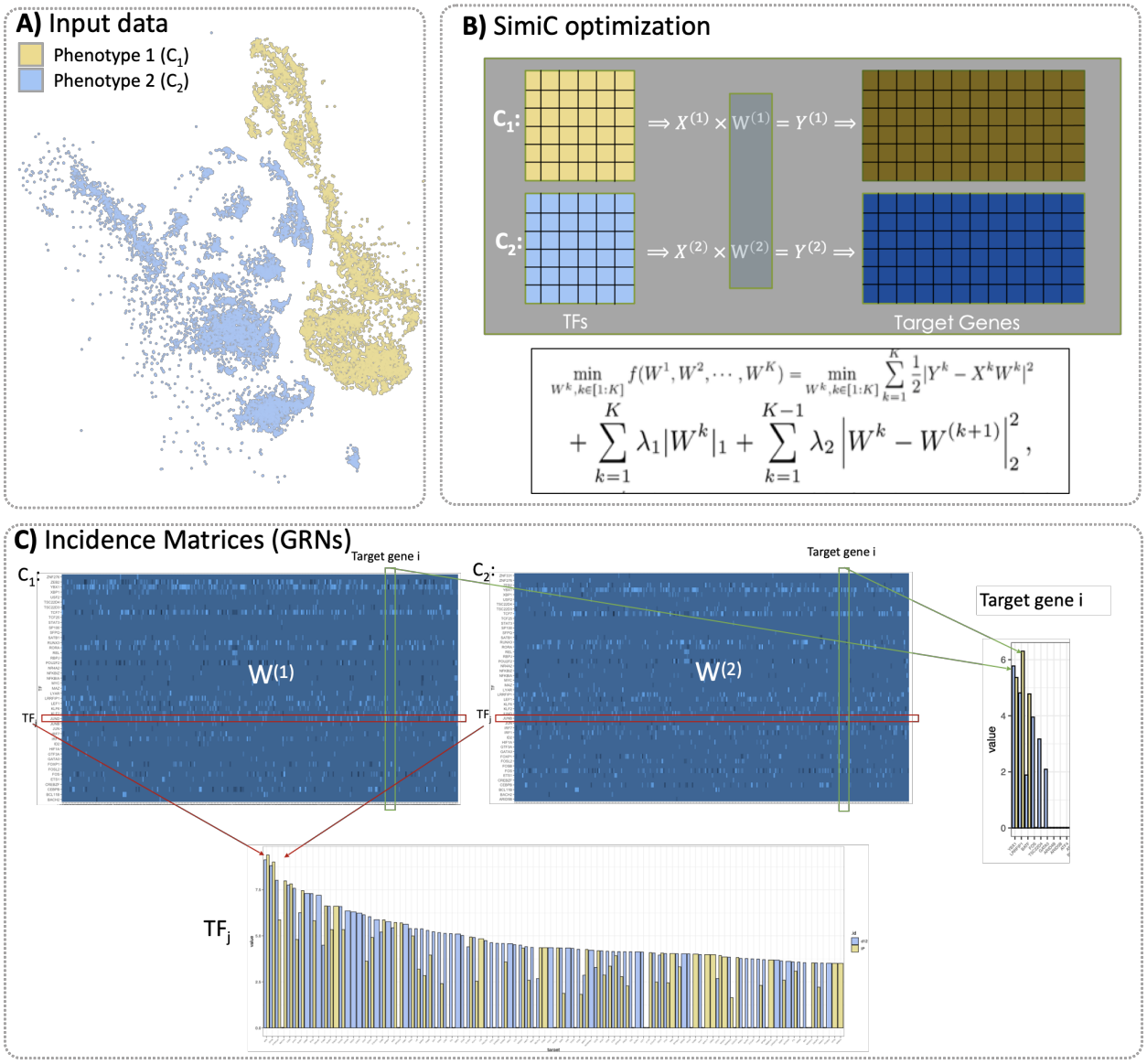
Evaluating the quality and accuracy of GRNs is not a straightforward task, as there are no clear metrics for evaluation (*13, 15*). Besides, most existing GRN inference algorithms only give a static picture of the GRN and provide hypotheses based on such network. SimiC, on the other hand, generates multiple networks at the same time, providing a framework to further evaluate the network dynamics. Specifically, drawing from Aibar et al. (*22*), we propose a *regulon activity score* that captures, per driver gene, the activity of their associated gene set in each individual cell. We further evaluate changes in distribution of the activity scores across cell phenotypes to describe the *regulatory dissimilarity* of each driver gene across cell phenotypes. All these analyses have been incorporated into the SimiC framework, and we show, on both simulated and real datasets, that SimiC provides more robust biological insights into regulatory changes across cell phenotypes than previously proposed methods. In particular, we show on simulated data that SimiC's inferred GRNs obtain higher micro-F1 and Cohen's kappa coefficients, as well as better goodness of fit as measured by the adjusted R^2. We also test SimiC on a variety of real datasets, including model and non-model organisms, and show that SimiC is able to recapitulate complex gene regulatory dynamics and accurately characterize them via the proposed *regulon activity* and *regulatory dissimilarity* scores.

Results

Overview of the method

We consider imputed scRNA-Seq expression data with the corresponding cell labels that indicate the phenotype of each cell (and possibly additional information such as cell type), as well as a linear ordering between cell phenotypes (Figure 1A). In cases with only two cell phenotypes (e.g., case vs control), the ordering is implicit. Note that when the phenotype is not available, the ordering of cells can be inferred, for example, from the pseudotime of cells using pseudotime inference methods.

SimiC expects the scRNA-Seq data to be represented as a matrix of dimension number of genes by number of cells, in which the set of driver genes are specified. We note that in our analyses we use transcription factors



$$\min_{W^k, k\in[1:K]} f(W^1, W^2, \cdots, W^K) = \min_{W^k, k\in[1:K]} \sum_{k=1}^{K} \frac{1}{2}|Y^k - X^k W^k|^2 + \sum_{k=1}^{K} \lambda_1 |W^k|_1 + \sum_{k=1}^{K-1} \lambda_2 \left|W^k - W^{(k+1)}\right|_2^2,$$

Fig. 1: Workflow of SimiC. A: Example of input data to SimiC consisting of the imputed cell expression data of cells with phenotype 1 (C_1) and phenotype 2 (C_2). The data is visualized in the tSNE-reduced space, where the color indicates the cell phenotype. **B:** SimiC objective function based on LASSO (*24*) with an added similarity constraint. The $W^{(i)}$s represent the inferred incident matrices, and the $X^{(i)}$s and $Y^{(i)}$s the expression matrices of the driver and target genes, respectively. The superscripts correspond to the cell phenotypes. λ_1 and λ_2 are hyperparameters to control the weights of the regularization terms. **C:** Incidence matrices inferred by SimiC, one per phenotype. The dimension is number of driver genes, such as transcription factors (TFs), by number of target genes. Weight $W_{i,j}$ represents the influence of the ith TF in the jth target gene. The bar plots show the differences in the inferred weights for a given TF and target gene for the two phenotypes (indicated by color).

(TFs) as the driver genes. SimiC draws from the fused LASSO regression technique (*24*) to infer a GRN for each cell phenotype, while imposing a (learned) level of similarity across GRNs of contiguous cell phenotypes (see Methods). Thus, the output of the optimization problem is a set of incidence matrices representing the GRNs associated to each cell phenotype (Figure 1B). These inferred matrices are of dimension number of TFs by number of target genes, and each entry is a weight representing the influence of the TF in the corresponding target gene (Figure 1C and Figure 2A). In what follows, and following the nomenclature of (*22, 25*), we denote *the regulon of a TF* as its connected target genes. However, unlike Aibar et al. (*22*), we take into account the weights of the edges between the target genes and the TF. We show that factoring the weights allows us to account for both the strength and the direction (promotion or repression) of the regulation between the TFs and their associated target genes.

In order to evaluate the regulatory dissimilarity of the different TFs across the different cell phenotypes, we perform the following analysis on the inferred GRNs. First, we measure how the activity of the different regulons change across cell clusters. We propose to use a new metric to compute it: the *regulon activity score* (Figure 2B, Methods), inspired by the AUC score from (*22, 25*). For each cell, the regulon activity score serves to quantify the relative activity of a given regulon with respect to the expression of all target genes. Intuitively, for a given TF, if the most expressed target genes in a cell correspond to those with the highest weights in its weighted regulon, the TF in question may have a large influence on the expression profile of the cell. This is represented by a large regulon activity score. The regulatory activity of a regulon can also be visualized at single-cell resolution in the tSNE-reduced space, giving more insight into how a specific TF changes its activity at single-cell level (Figure 2C).

For each TF, we also compute the empirical distribution of the regulon activity score for each cell phenotype (Figure 2D). We then compute the distance (via the total variation metric) between the empirical distributions to measure how the activity of the regulons change across different cell phenotypes (Figure 2E, Methods). We refer to this distance as the *regulatory dissimilarity*. Further, subdividing the overall cell population into smaller clusters allows us to deepen into the regulatory dissimilarity across cell phenotypes at cell-cluster level. Thus, the SimiC workflow can generate a heatmap that shows the regulatory dissimilarities for all regulons and all cell clusters. This heatmap allows us to uncover shifts in regulatory activity that are associated with different conditions, environments, or developmental states (Figure 2F).

Finally, when assessing the accuracy of the inferred GRNs, we measure the goodness of fit via the adjusted R^2 coefficient between the true target genes' expression and the one computed as the linear combination of the TFs' expressions inferred by SimiC (via the GRNs). Specifically, we split the data into training and testing sets with a 80/20 ratio. We then infer the GRNs on the training data, and use the inferred networks, together with the TFs' expression on the test data, to predict the target genes' expression. Finally, we compute the adjusted R^2 on the test data using the true target genes' expression and the predicted one. In addition, the sparsity of the regulatory program of each target gene (that is, the set of TFs regulating a given target gene) must be also controlled, as otherwise the LASSO optimization may generate highly dense GRNs, which is undesirable (*7, 26*). Thus, on real datasets, we control the sparsity via hyperparameter tuning to achieve desirable levels of sparsity while maintaining satisfactory goodness of fit (see Methods).

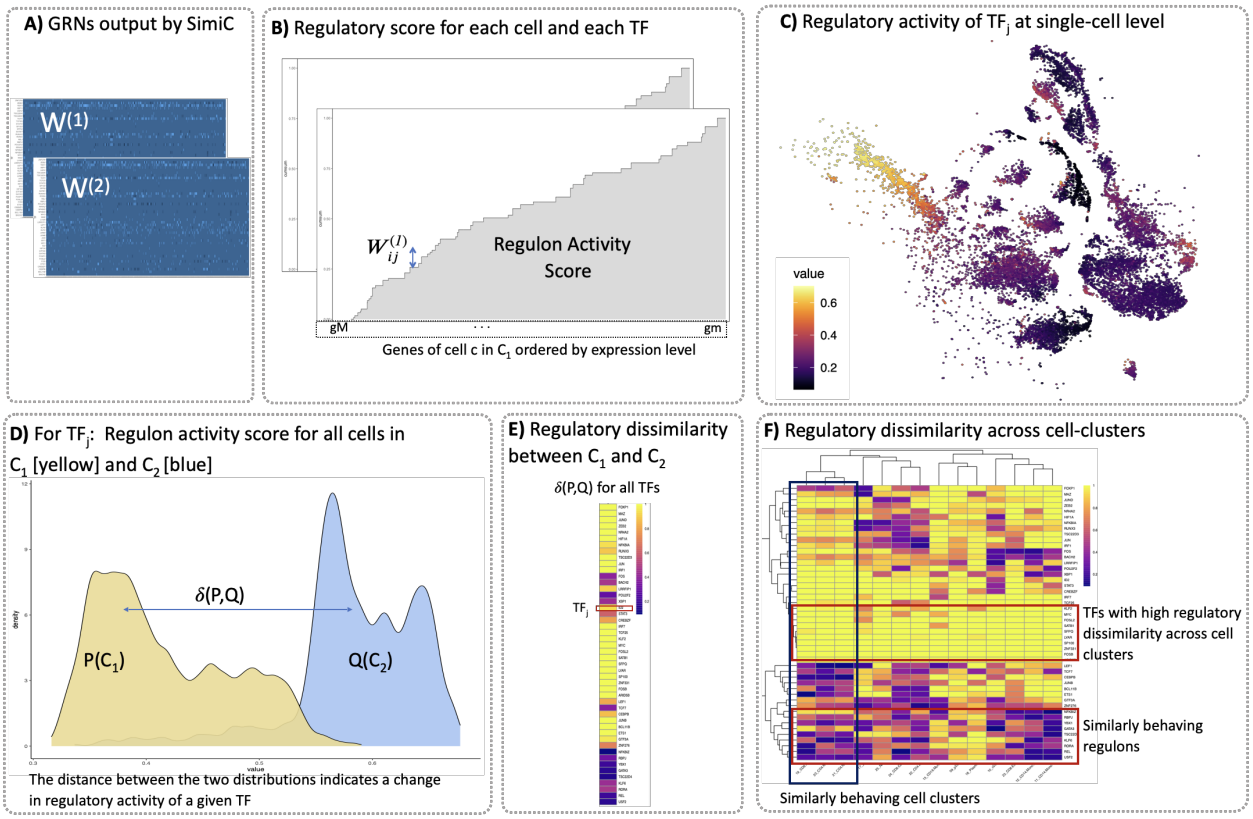


Fig. 2: Proposed metrics to unravel complex gene regulatory dynamics. A: SimiC's inferred incidence matrices $W^{(1)}$ and $W^{(2)}$ for phenotype 1 and 2, respectively, for the input data of Figure 1A. The incidence matrices correspond to the inferred gene regulatory networks (GRNs). **B:** For a given regulon, and hence transcription factor (TF), we compute the regulon activity score per cell as follows. First, for a given cell c, we sort the target genes by their expression values in an increasing order. Then, for each sorted target gene we show the contribution of its corresponding weight in the weighted regulon by increasing the y-axis by that amount. The regulon activity score corresponds to the area under the generated curve. **C:** tSNE plot visualizing the input scRNA-Seq data, where each cell is colored by its regulon activity score for TF$_j$. **D:** For a given TF, using the cells' individual regulon activity score, we can compute the empirical regulon activity score distribution for each cell phenotype. An example of the obtained distributions for the considered two phenotypes and a given TF are depicted, $P(C_1)$ for phenotype 1 (yellow) and $Q(C_2)$ for phenotype 2 (light blue). **E:** Computing the distance between the obtained distributions for each TF provides a metric for the regulatory dissimilarity of each TF between the cells belonging to phenotype 1 (C_1) and those belonging to phenotype 2 (C_2). The heatmap shows an example of the obtained regulatory dissimilarity for several TFs when the total variation distance $\delta(P, Q)$ is used. **F:** The dissimilarity score can be further computed for different cell-clusters across the phenotypes, as shown in the heatmap, which reveals TFs with high dissimilarity scores across cell clusters as well as TFs with similar ones.

SimiC captures network dynamics on synthetic data

We first evaluated the performance of SimiC on synthetic data satisfying the assumption that GRNs changes smoothly between states, and showed that it captured the network dynamics more accurately than the previously proposed inference method SINCERITIES (*20*). To the best of our knowledge, SINCERITIES is the only existing method that produces one GRN per state and connects TFs and target genes with weighted edges, similarly to SimiC, which allows for a fairer comparison. Specifically, we did not consider SCENIC (*22, 25*) for this comparison as it can not be run without transcriptome annotations (such as TF-binding sites), which do not apply in this case. In addition, the GRN inference methods GENIE3 (*27*) or GRNBOOST2 (*28*), which are used internally by SCENIC, do not produce weighted GRNs, which impairs their analysis based on model fitting.

We simulated scRNA-Seq data for 5 different cell states (or phenotypes) with a linear order between them, where the underlying GRNs of consecutive states did not change abruptly (Figure 3A, see Methods). We simulated scRNA-Seq data for $1,000$ cells per state, for a total of $5,000$ cells. We considered 50 TFs and 20 target genes, and GRNs containing 3 types of edges: no regulation (edge weight $= 0$), activation (edge weight $= +1$) and depression (edge weight $= -1$). Finally, we simulated the TF's expression by sampling from a negative binomial distribution (*29, 30*), and computed the targets' expression using the constructed GRNs and the TF's expressions, alongside an added Gaussian noise (Figure 3A, see Methods).

We ran SimiC and SINCERITIES on the generated synthetic data and, to showcase the importance of the state separation in fitting the data, we also ran LASSO on all states combined. Note that in the latter case only one GRN was generated. For each method, we computed the micro-F1 score and the Cohen's kappa coefficient of the inferred networks. The micro-F1 score assesses the quality of multi-label binary problems and measures the F1 score of the aggregated contributions of all classes. The Cohen's kappa coefficient is more robust in multi-class prediction and provides an insight into how well the model performs when compared to a random guess. Since the constructed GRNs only contain edges equal to $0, +1$ or -1, we set a threshold (0.5 by default) such that all weights below -0.5 are converted -1, all weights above 0.5 to $+1$, and the rest to 0.

SimiC obtained the highest micro-F1 score for all states as compared to SINCERITIES and LASSO (Figure 3B). The difference in performance is more pronounced with SINCERITIES. Whereas SimiC obtains micro-F1 scores close to 1, SINCERITIES obtains micro-F1 scores around 0.8. Due to the sparsity of the GRNs, the micro-F1 score is highly biased towards the F1 score of predicting 'no edges' (0 weight) and does not accurately reflect the capability of the method to predict the other two types of edges ($+1$ or -1). The Cohen's kappa coefficient does not suffer from this, as it measures the gain of the model when compared to a random guess based on the support of each type of edges. Whereas SimiC achieved Cohen's kappa coefficients around 0.8, SINCERITIES's scores oscillate around 0 and did not surpass 0.075 (Figure 3C). With respect to LASSO, we observed that it produced Cohen's kappa coefficients below 0.8 in all states. We also observed that the Cohen's kappa coefficient for SimiC was stable with the threshold used to convert the weights into $+1, -1$ and 0, similar to what is observed with LASSO (Supplementary Figure S1A-B). However, this was not the case for SINCERITIES (Supplementary Figure S1C). These results demonstrate

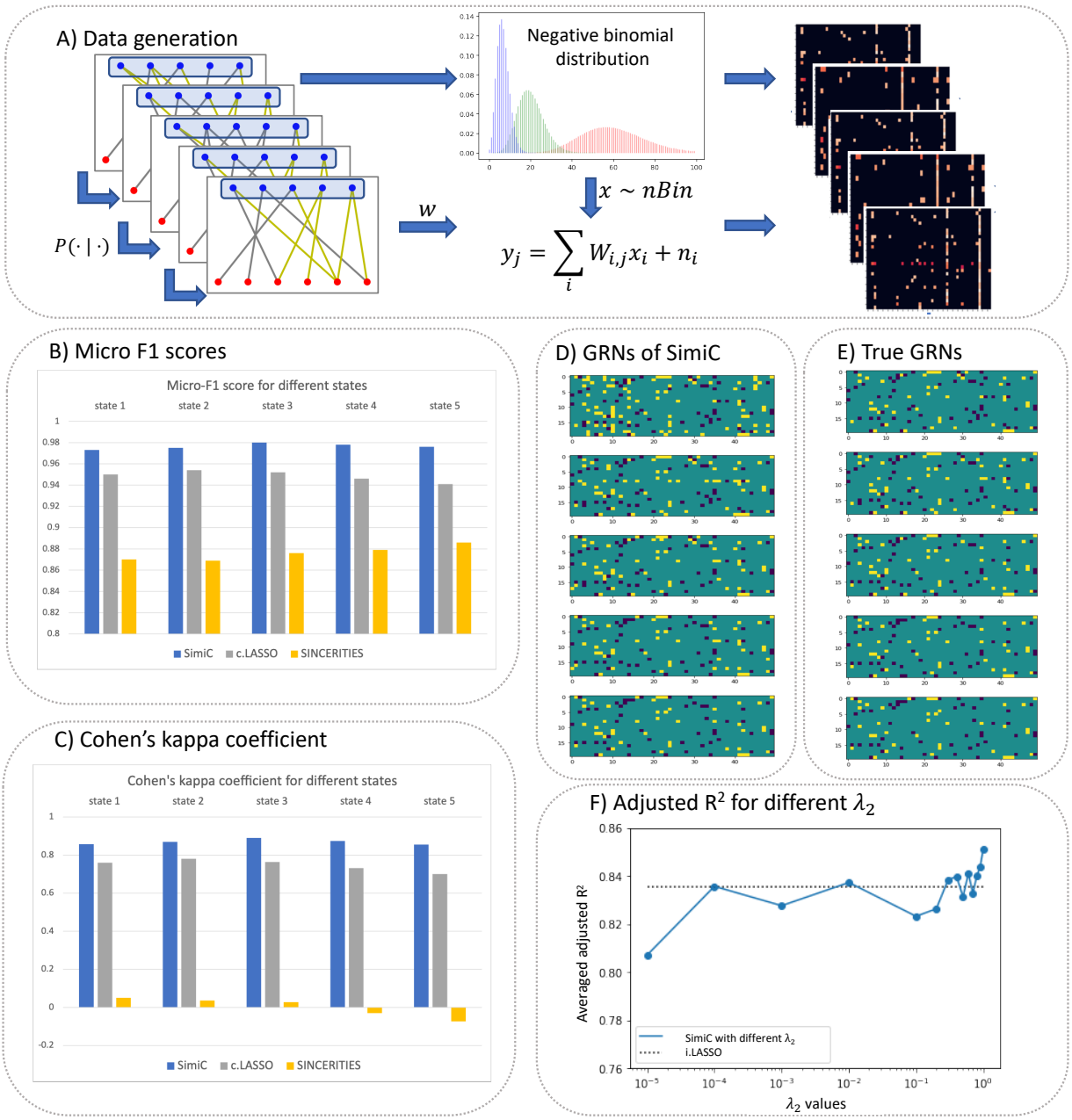


Fig. 3: SimiC accurately infers network dynamics on synthetic data. A: Workflow for generating the synthetic data. First, we generate the regulatory network of each state such that no abrupt changes occur between the networks of two consecutive states. We then generate the TF's expression data by sampling from a negative binomial distribution (*29, 30*). Finally, we generate the target genes' expression data using the generated networks and the expression of the TFs. In particular, the expression of the jth target gene y_j is generated as $\sum_i W_{i,j} x_i + n_i$, where $W_{i,j}$ is extracted from the generated incidence matrix, x_i represents the expression of the ith TF, and n_i is modeled as Gaussian random noise (independent and identically distributed across samples). **B:** Micro-F1 score and **C:** Cohen's kappa coefficient on edge prediction obtained by SimiC, c.LASSO (LASSO run in all states combined) and SINCERITIES. **D:** Obtained incidence matrices for SimiC and **E:** simulated incidence matrices.**F:** Changes of adjusted R^2 with different λ_2.

that when SimiC's assumptions are met, SimiC better captures the states' network dynamics, as it infers more accurate GRNs than SINCERITIES and LASSO. We also observe that, as expected, the GRNs inferred by SimiC do not change abruptly across consecutive states, reproducing the simulated GRNs (Figures 3D and 3E).

To assess the goodness of fit of SimiC we computed the adjusted R^2 coefficient for values of λ_2 ranging from 10^{-5} to 1 (λ_2 represents the weight assigned to the similarity constraint in SimiC's optimization problem, see Methods), while the rest of the hyperparameters are kept constant. We observed that the average adjusted R^2 for SimiC ranged from about 0.8, achieved with the smallest value of λ_2, to about 0.85, achieved with the largest value of λ_2 (Figure 3F). Hence, in this case, putting more weight in the similarity constraint produced more accurate GRNs, which led to more accurate expression for the target genes. Note that this was expected, as the synthetic dataset satisfies SimiC's assumptions in regards to similarity across GRNs of consecutive states. As a comparison, removing the similarity constraint (i.e., setting $\lambda_2 = 0$, which corresponds to LASSO run independently in each state) yielded an adjusted R^2 of about 0.835, which corresponds to a reduction of about 2%. On the other hand, running LASSO on the combined data yielded an adjusted R^2 of approximately 0.49. Overall, these results show that jointly inferring a GRN per state while introducing a similarity constraint better captures the underlying network dynamics when a smooth transition between states exist.

In terms of running time, SimiC employed 4 minutes to infer the GRNs and 3 minutes to compute the regulon activity scores, while SINCERITIES employed 2 minutes to infer the GRNs (see Methods).

SimiC correctly infers regulatory connections between TFs and target genes

To assess the performance of SimiC on real datasets, we first evaluated the correctness of the association between the TFs and their corresponding target genes. Two scRNA-Seq datasets (GEO accession ID GSE139369) were considered for the analysis: i) a scRNA-Seq dataset composed of monocyte cells coming from either bone marrow (BM, 2,681 cells) or peripherial blood (PB, 3,610 cells) (*31*); and ii) a scRNA-Seq dataset composed of CD4+ T-lymphocytes cells also coming from either BM (5,297 cells) or PB (7,563 cells) (*31*). For both datasets, after pre-processing the data (see Methods), we selected the 100 most variant TFs and the 1,000 most variant target genes, where the variation was measured as the mean absolute deviation (MAD).

On the dataset composed of monocyte cells, running SimiC resulted in two GRNs, each representing the regulatory landscape of monocytes in PB or BM. SimiC obtained a median adjusted R^2 on test data of 0.861 (Supplementary Figure S2A), while yielding a median of 7 TFs regulating each target gene (Supplementary Figure S2F). Next, we analyzed in more detail the regulatory relationships inferred by SimiC between TFs and target genes using ChIP-Seq data. For each TF, we sought empirical evidence of TF binding sites up to 5Kbp upstream of its target genes. We extracted the TF binding sites information from the ReMap2020 database (*32*). Out of the 100 considered TFs, 7 were dropped as no target genes were assigned to them (due to potential colinearities with other TFs) and 28 lacked ChIP-Seq experiment data, and hence they were also excluded from the analysis. Out of the remaining 65 TFs, 47 had their set of target genes heavily enriched in the corresponding set of TF binding sites (odds ratio [OR] > 2, p-value < 0.05, see Methods), while the

remaining TFs showed partial ChIP-Seq evidence (Supplementary Table S1, Supplementary Excel File 1). Note that SimiC infers regulons based solely on the gene expression profiles from the input scRNA-Seq data. Therefore, some edges inferred by SimiC may be due to an indirect regulation between the TF and the target gene. In such cases, as the TF is likely not binding in the proximity of the target gene, it may not be possible to validate the regulatory interactions via the performed ChIP-Seq analysis.

When focusing on the differences in the composition of the weighted regulons across tissues (i.e., BM vs PB), we observed, for example, a change in the composition of the regulon of KLF4, a well-known transcription factor involved in essential monocyte development from the monocyte-dendritic cell progenitors (*33*). Thus, as expected by its importance in monocyte development, the regulon of KLF4 had larger weights on the BM than on PB (Supplementary Figure S3A). In addition, KLF4 had, exclusively on BM, connections to target genes *MGST1*, *ICAM3* and *S100A12*, all playing a crucial role during monocyte differentiation and maturation (*34–36*).

We then performed a similar analysis on the CD4+ T-lymphocytes cells coming from either BM or PB (*31*). SimiC also generated two GRNs in this case, which exhibited a median adjusted R^2 of 0.840 on test data (Supplementary Figure S2B) while yielding a median of 6 TFs regulating each target gene (Supplementary Figure S2H). The ChIP-Seq analysis revealed that out of the 63 TFs with ChIP-Seq evidence, 34 had their associated target genes enriched (OR > 2, p-value < 0.05, Supplementary Table S2, Supplementary Excel File 1), while the other 29 presented partial ChIP-Seq evidence (Supplementary Table S2, Supplementary Excel File 1). When delving into specific regulons, we found that *RUNX3*, a TF involved in the activation of naive CD4+ T cells during immune response (*37*), was positively associated with target genes *HLA-F*, *PTPN6*, *GADD45B*, *SFRS9* and *SEMA4D*, all key genes on the stimulation and activation of CD4+ T cells (*38–42*) (Supplementary Figure S3B).

SimiC unravels complex gene regulatory networks and dynamics within CAR T cells

We assessed the capabilities of SimiC to uncover complex gene regulatory dynamics of engineered CD8+ T lymphocytes during immunotherapy of non-Hodgkin lymphoma patients. The development of adoptive cell therapy approaches based on T lymphocytes engineered with chimeric antigen receptors (CAR T cells) has been a breakthrough for cancer immunotherapy strategies (*43, 44*). Functionality of CAR T cells depends on an orchestrated activation of specific signaling pathways after exposure to their cognate antigen, together with an evolution through different T cell subpopulations. Specifically, CAR T cells generated by current technologies are enriched in different types of long-lived memory cells that can quickly expand to large numbers of terminally differentiated effector T cells upon exposure to their antigen. However, after long-term antigen stimulation CAR T cells can become exhausted, leading to a defective or insufficient antitumoral function. Therefore, understanding the changes in gene regulatory dynamics suffered by the infused CAR T cells after a prolonged exposure to tumor cells is of utmost importance for the efficient design and application of this immunotherapy strategy.

In this context, Sheih *et al* (*45*) generated a single-cell transcriptome dataset of CD8+ CAR T cells isolated from the infusion product (IP), as well as from peripheral blood at the expansion peak after treatment (day 7–14; termed D12) of patients with relapsed and refractory non-Hodgkin lymphoma (GEO accession ID

GSE125881). This dataset contains 7,616 cells for the IP phenotype and 8,576 cells for the D12 phenotype, and we selected the most variant 100 TFs and 1,000 target genes for the analysis. We applied SimiC on the scRNA-Seq data to infer two akin GRNs governing the CAR T cell function at IP and D12 timepoints (median adjusted R^2 of 0.834 on test data, Supplementary Figure S2C; median of 5 TFs regulating each target gene, Supplementary Figure S2H).

To understand the potential changes in the overall regulatory dynamics between IP and D12, we computed the *hub and authority* score (*46*) for each gene of each GRN (see Methods). In brief, a hub represents a TF that regulates many other genes, while an authority represents a target gene that is regulated by many different TFs. SimiC was able to capture a shift in the overall authority scores after tumor recognition by CAR T cells, indicating an increment in the regulatory activity after tumor recognition (Figure 4A, Wilcoxon rank test, FDR < 0.05). Specifically, we observed that the authority score of key genes in tumor recognition, such as *GZMB*, *NOSIP* and multiple *HLA* family genes, were significantly incremented in D12. These genes have been widely studied as effectors of the acute response of the immune system (*47, 48*).

In order to delve into the regulatory dynamics of individual TFs, we computed the distribution, across all cells, of the *regulon activity scores* (see Methods). Recall that, for a given TF, this score measures the relative activity of the set of target genes connected to the TF on a given cell. Due to their special relevance at different stages of the immune response, we first focused on the TFs *MYC* and *RUNX3*, genes with an antagonist behavior (*49, 50*). The regulon of the proto-oncogene *MYC* showed its regulatory peak activity at the IP, as expected from its enhanced proliferative capacity (*51*) (Figure 4B, top, green density). In contrast, the regulon of the *RUNX3* gene, which is required to maintain a lytic activity once the cells have met the antigens, showed the regulatory activity peak after prolonged exposure to tumor cells (*37*) (Figure 4B, bottom, purple density). Analyzing the distributions of the regulon activity scores in more detail, we observed an overall lack of activity of the MYC regulon at the expansion peak (D12). This was expected due to the *MYC* role in cellular proliferation, a process that is partially halted on exhausted T cells. On the other hand, different activity peaks could be observed at IP, yielding a multimodal distribution of the regulon activity score. An analogous multimodal distribution could be observed for the RUNX3 regulon at D12, as some cells showed little activity of the RUNX3 regulon, while other cells showed a highly active RUNX3 regulon (Figure 4B, bottom, purple density). As we will show in the next section, the multimodality of these distributions was primarily due to the different levels of exhaustion reached by the CAR T cells at D12.

To examine in more detail the activity of the MYC and RUNX3 regulon at single-cell level, we colored the input single-cell data by their respective regulon activity scores and visualized it in a tSNE plot (Figure 4C). For the MYC regulon, we observed a cluster of MYC-active cells at IP at the bottom left side, and a left-to-right gradient along which the regulon activity decreased progressively. On the other hand, almost all cells at D12 were MYC-innactive, with the exception of a small cluster of cells at the left-most edge. Interestingly, these cells happened to be a specific subtype of CAR T cells which were transitioning from memory CAR T cells to exhausted CAR T cells, as further described in the following section. When focusing on the RUNX3 regulon, we observed a nearly-opposite behaviour to MYC. Cells were mostly RUNX3-innactive at IP, while we observed a clear left-to-right gradient of RUNX3 activity on D12. This gradient was explained by the

progressive exhaustion suffered by the CAR T cells (see next section).

By inspecting the regulons MYC and RUNX3 (for both phenotypes), we observed that their shared targets had opposite weights, such as *GZMB*, which at IP was regulated positively by RUNX3 and negatively by MYC (Figure 4D). Further, at the expansion peak (D12), the negative regulation from MYC was lost, while the positive one from RUNX3 was maintained. Another example of such reversed regulation between MYC and RUNX3 is the target gene *SELL*, which controls the proliferation of memory cells and was positively regulated by MYC and negatively by RUNX3 (*52*). Finally, we note that these shifts in connectivity between the different timepoints were shared across most regulons as the authority scores of target genes were generally increased on D12 (Figure 4A).

SimiC elucidates regulatory activity for different cell subtypes

While GRNs are generally computed on cells coming from similar lineages or phenotypes (*20, 22*), specific regulons of these networks may be differently activated depending on the current state of the cell. For example, differential activation of regulons have been shown to play a key role on differentiating cells (*25, 31*).

In what follows, we show, using the dataset from the previous section, that SimiC is able to capture the underlying regulatory activity within the different CAR T cell subpopulations and unravel the differences in their activity after tumor recognition. In addition, this analysis allows for the identification of clusters of regulons driving such differentiation. Thus, once the GRNs using the whole CAR T cell population were inferred by SimiC (see previous section), and the regulon activity score computed for each cell, we grouped cells by their subtype; namely, CD8+ memory, CD8+ effector and CD8+ exhausted cells. We then computed for each cell subpopulation and each TF the distribution of regulon activity score, which yielded two empirical distributions. Finally, we computed the distance between these distributions and denoted it as the *regulatory dissimilarity score* (see Methods), which measures, for each TF, the differences in overall regulon activity between phenotypes (IP vs D12) (Figure 5A).

The multimodal distributions observed in the regulon activity of RUNX3 and MYC (Figure 4B) are also reflected in the heatmap of Figure 5A, where RUNX3 and MYC (among others) show different levels of regulatory dissimilarity scores for different cell populations. Specifically, a low regulatory dissimilarity score was observed on exahusted CAR T cells for RUNX3 while high dissimiarity can be observed in the same population for MYC.

To understand the different behaviors of these cells, we colored the tSNE projection of the cells by their regulon activity score (Figure 5B). We observed that on D12, the peak of the immune response, RUNX3 was more active on CD8+ effector cells, which are the cells carrying the immune response. Once the cells become exhausted (CD8+ exhausted cells), the activity of RUNX3 was significantly reduced. On the other hand, MYC showed an opposite activity pattern, being most active during IP on the CD8+ memory cells, as it was executing its proliferative and differentiative function. MYC saw its regulatory activity diminished over time, reaching its minimum at D12 on CD8+ exhausted cells.

We were also able to uncover a cluster of TFs showing high regulatory dissimilarity between IP and D12 across all CAR T cell subpopulations (Figure 5A, yellow cluster). TFs yielding such high regulatory dissimilarity scores regardless of the cell subtype are expected to modulate the general activity of the immune

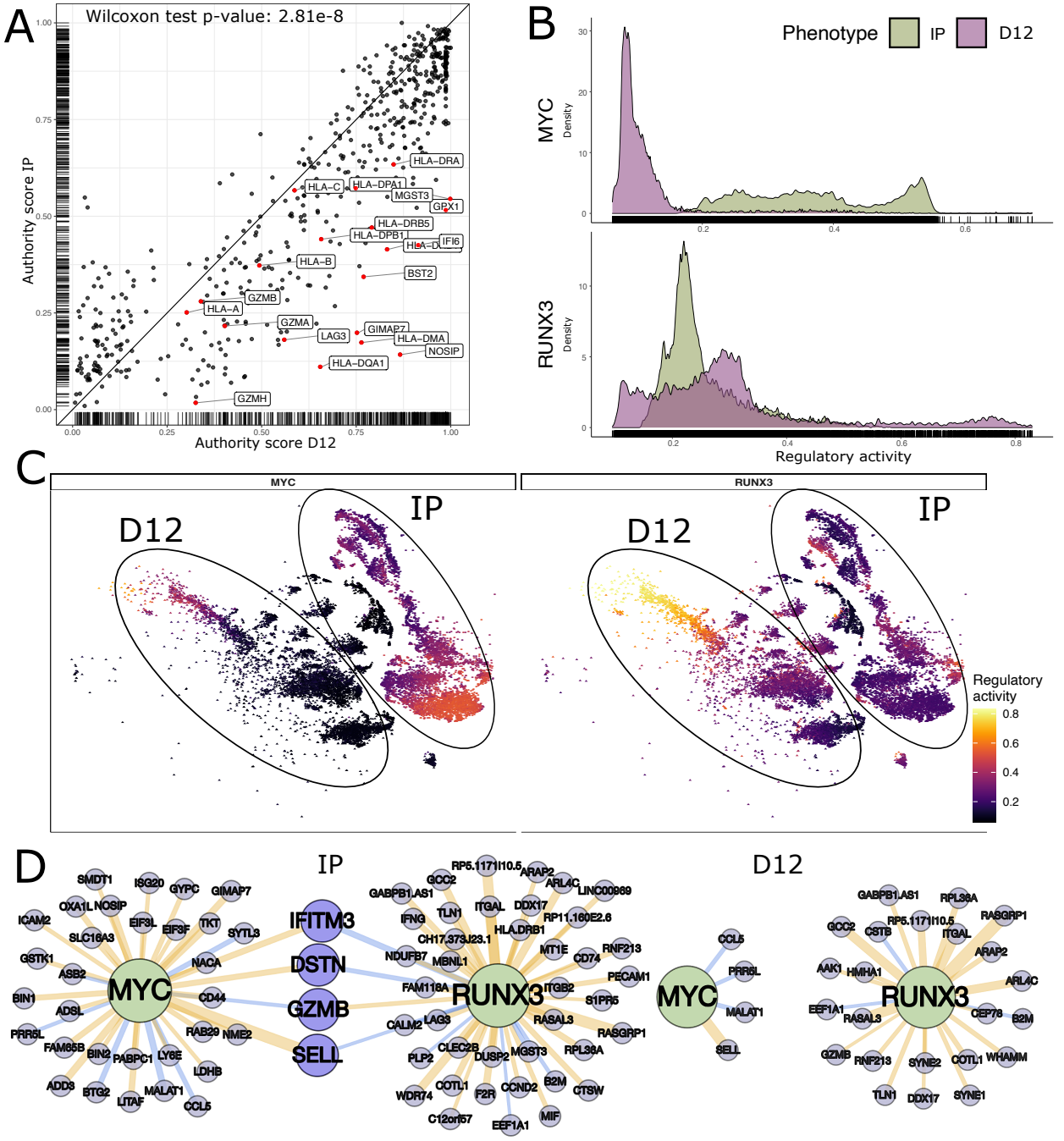


Fig. 4: SimiC captures complex regulatory dynamics within CAR T cells. A: Comparison of the authority scores for each target gene in the the two timepoints IP and D12. We observe an increment of the authority score in D12 as compared to IP (Wilcoxon rank test, FDR < 0.05). **B:** Ridge plot showing the regulon activity scores for genes MYC and RUNX3 on CAR T cells, at the two different timepoints IP and D12. **C:** tSNE plot showing the CAR T cells colored by their regulon activity score for transcription factors MYC (left) and RUNX3 (right). Cells belonging to the two considered timepoints IP and D12 are highlighted. **D:** Graph representing the weighted regulons of the transcription factors MYC and RUNX3 for the considered timepoints IP and D12. The color of the edges linking the TFs with their corresponding target genes indicates whether the regulation is positive (yellow) or negative (blue), and the width of the edge corresponds to its strength.

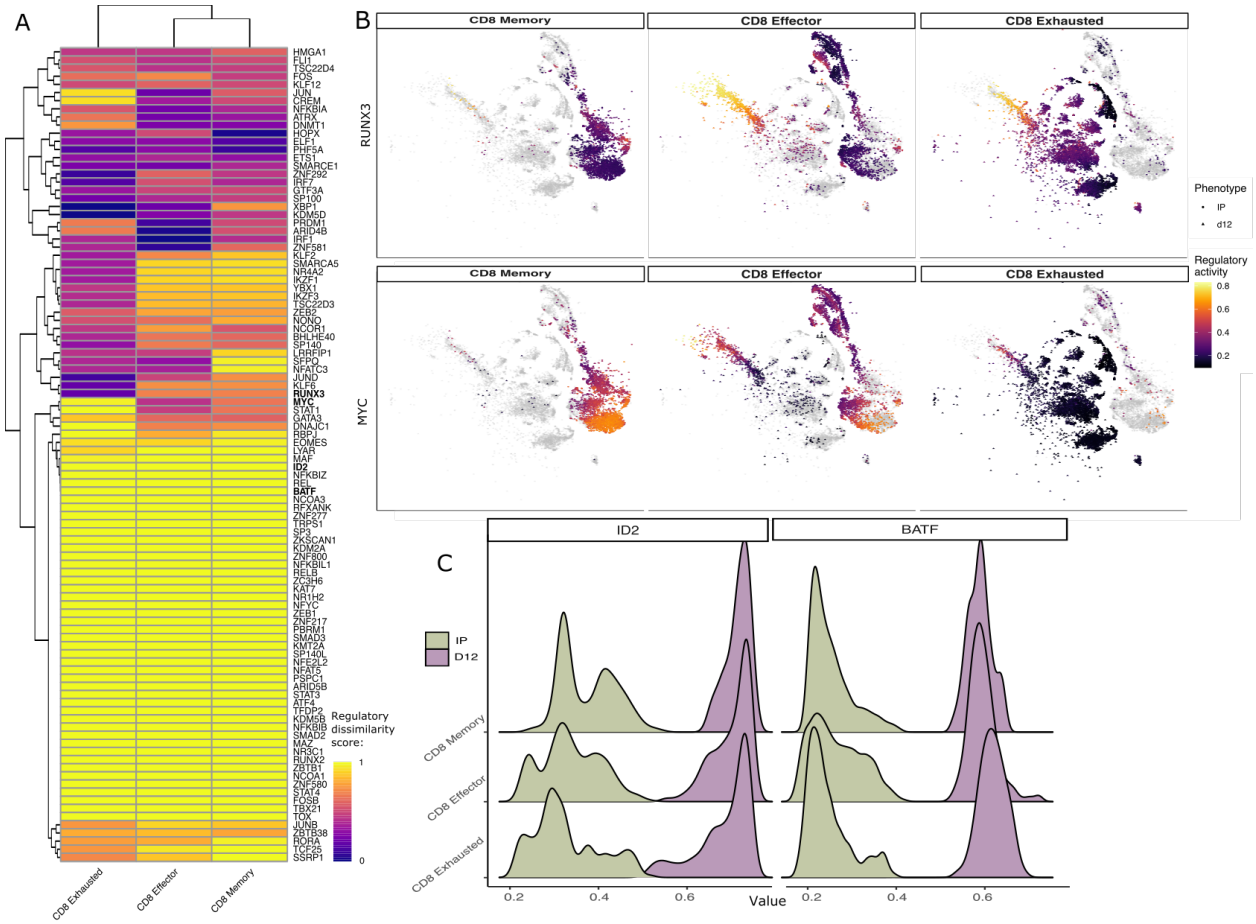


Fig. 5: SimiC's proposed scores capture cell subpopulation specific regulations. A: Heatmap depicting the difference of the regulatory dissimilarity between the two timepoints (IP and D12) of different regulons for CD8+ exhausted, CD8+ effector and CD8+ memory cells, showing cell-dependent variation. **B:** tSNE visualization of the CD8+ subpopulation cells colored by their activity score of the RUNX3 and MYC regulons. The shape of the cell indicates the timepoint (IP or D12). **C:** Ridge plot showing the regulon activity scores for genes ID2 and BATF on the CD8+ subpopulations, for the two timepoints IP and D12.

response after tumor recognition. Example of such TFs are ID2 and BATF genes, key transcriptional regulators essential in the immune response of the CD8+ T cells (*53, 54*). These TFs presented lower values of regulon activity on the IP cells (Figure 5C, green densities, and Supplementary Figure S3C), and a clear increment on their regulon activity after tumor recognition (D12) (Figure 5C, purple densities, and Supplementary Figure S3C).

Finally, to compare the capacity of SimiC to uncover phenotype-specific regulatory activity to that of SCENIC, we performed a similar analysis on the CAR T single-cell dataset using SCENIC. Since SCENIC infers a unique GRN for a given input, we ran SCENIC on the whole dataset, as well as on the IP and D12 cells independently. Note that in the first case only one GRN was inferred for all cells, while in the latter case two GRNs were inferred, each capturing the regulatory dynamics of the IP or D12 cells. While the AUC scores computed by SCENIC provide similar biological insights to that of SimiC's regulatory activity scores, they did not show phenotype- or cell-type-specific clusters in either of the considered scenarios

(Supplementary Figure S4). This is in contrast to SimiC, whose regulon activity scores clustered cells by phenotype or cell type (Figure S4C), manifesting the capacity of SimiC to uncover activity scores related to the phenotype at hand.

Finally, we show that SimiC was able to recapitulate some of the T cell differentiation features related to cell proliferation using the MYC regulon (a proliferation marker). At IP, MYC had its highest regulon activity on the memory cells, due to the proliferation potential of these cells. As the immune activity progressed, the cells became more and more effector-like, halting the proliferation and thus, diminishing the activity of this regulon. On D12, where the effector cells were already transitioning towards the exhaustion state, the activity of the MYC regulon got progressively reduced, as the proliferative capabilities of cells were being lost, reaching its lowest values on the exhausted cells (Figure 4C, left). These trajectories represent the proliferative dynamics of the CAR T cells during the immune response. These subtle, albeit important, dynamics of regulon activities were not recapitulated by SCENIC (Figure S5B-C).

SimiC models GRN-dynamics across several timepoints on liver regeneration

We next show how SimiC was able to model complex regulatory dynamics on datasets with multiple sequential cell phenotypes (i.e., timepoints). To this end, we used an scRNA-Seq dataset (GEO accession ID GSE151309) generated from regenerating mouse livers, which contains timed gene expression profiles of around 12K hepatocytes (*55*). Specifically, 2/3rd partial hepatectomy was performed and cells were sequenced at different timepoints (24h, 48h, and 96h) post-surgery. Healthy adult mouse liver cells were also sequenced, yielding a total of four timepoints, termed adult, PHx24, PHx48 and PHx96. Non-hepatocyte cells were identified and removed from the dataset. The cells were annotated by their functional state as quiescent, proliferative, metabolically hyperactive, or transitioning.

Running SimiC on this dataset generated four akin GRNs, each representing the regulatory landscape at a given timepoint (median adjusted R^2 of 0.885 on test data, Supplementary Figure S2D; median of 7 TFs regulating each target gene, Supplementary Figure S2I). When the number of cell phenotypes is larger than two, we define the regulatory dissimilarity score to be proportional to the area between the largest and the smallest densities at each regulon activity value (see Methods). This score therefore yields a dissimilarity score of 1 for non-overlapping distributions and a score of 0 for completely overlapping distributions. For example, TF YBX1 presented highly overlapping regulon activity densities, yielding a dissimilarity score of 0.41 on proliferating cells; while the densities of the TF CEBPB were mostly non-overlapping across all cell types, notably yielding a dissimilarity score of 1 for proliferating cells (Figures 6A and Figure 6B).

Two important regulatory patterns arise in the modeling, via GRNs, of the regenerating liver (*55*). On the one hand, regulons following an initiation–progression pattern typically show low activity values in adult hepatocytes, while after hepatectomy there is a peak of activity (PHx24 and PHx48); then, the original value is regained again at PHx96. On the other hand, regulons following a termination–rematuration pattern show high activity in adult hepatocytes, while post-hepatectomy there is a consistently less activity (PHx24 and PHx48); and the activity rises again at PHx96.

Two well known representatives of these patterns are the key hepatocyte TFs CEBPB and CEBPA, transcription factors that bind to similar DNA sequences (*56*) and play divergent but key roles in liver

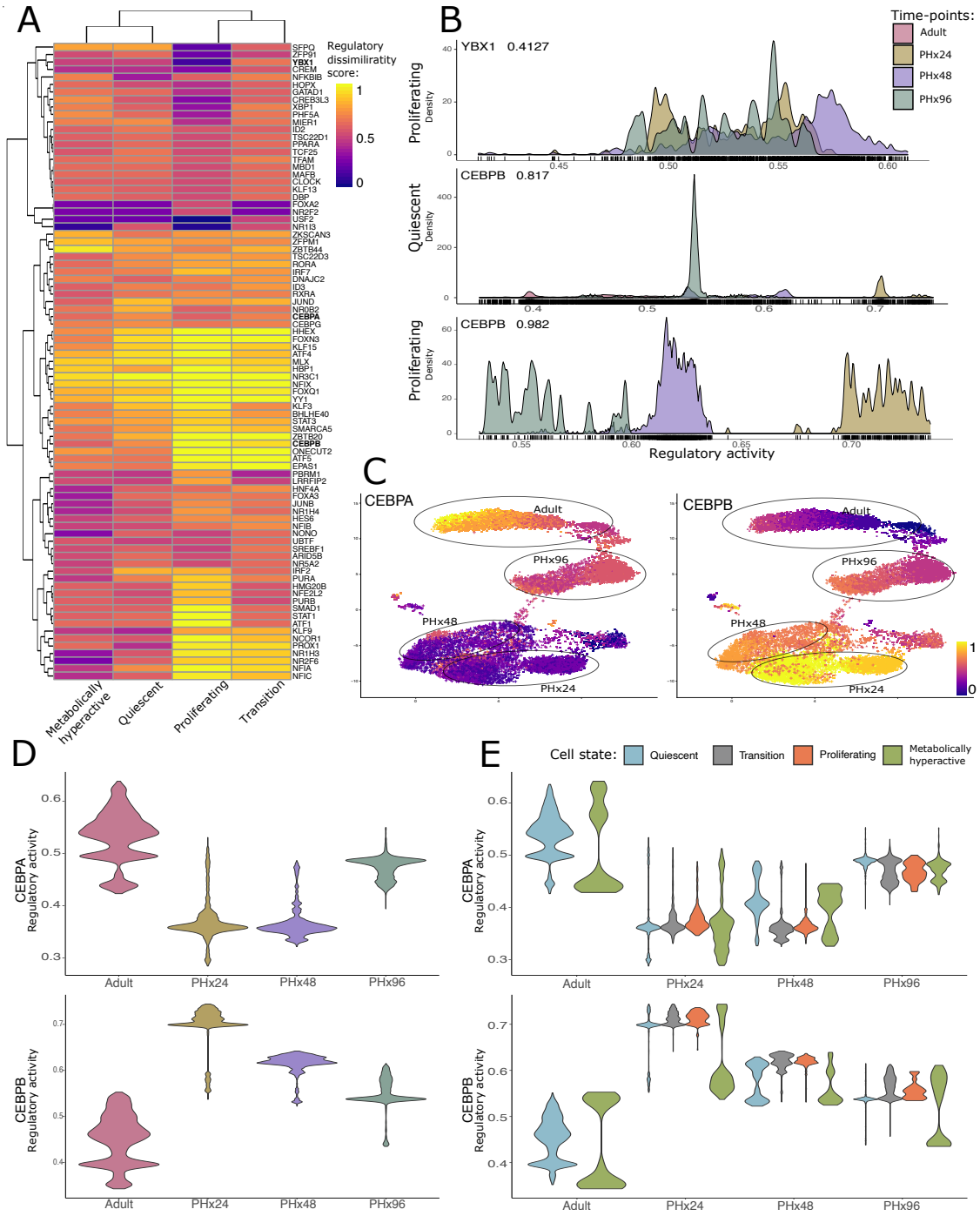


Fig. 6: SimiC's inferred GRNs across timepoints on a regenerating liver capture regulatory dynamics of hepatocytes in different functional states. **A:** Heatmap of the regulatory dissimilarity score of quiescent, proliferative, metabolically hyperactive, and transitioning cell types across the four sequential timepoints for the different regulons, showing cell and time dependent variation. **B:** Density plots showing the regulatory activity for the regulon YBX1 on the proliferating cell population and for the regulon CEBPB on the quiescent and proliferating cell populations, for the four considered timepoints. **C:** tSNE plot showing the hepatocyte cells, colored by their regulatory activity score for the regulons CEBPA and CEBPB. Shapes correspond to the different cell types. Timepoint to which each cell belongs is also specified. **D:** Violin plots showing the distribution of the regulatory dissimilarity scores for regulons CEBPA and CEBPB across timepoints. **E:** Violin plots showing the distribution of the regulatory dissimilarity scores for regulons CEBPA and CEBPB across cellular states of the liver regeneration, for the four considered timepoints.

regeneration (*57*). For instance, it has been shown that CEBPB follows an initiation-progression pattern in liver regeneration, while CEBPA follows a termination-rematuration pattern (*55, 57*).

These activity patterns for CEBPA and CEBPB were clearly represented by their distribution of regulon activity scores across the different timepoints (Figure 6D). Specifically, CEBPB had its activity peak at the first stages of liver regeneration after the hepatectomy (PHx24 and PHx48), and was then progressively being reduced towards the values of the adult cells. Opposite to this drift, the maximum activity of CEBPA was on the adult cells, and dropped to its minimum just when the regeneration of the tissue starts (PHx24), and later recovered its activity when the tissue converges to a mature state.

Next, we split the activity scores at each timepoint by the different cell states (Figure 6E). We observed that almost no transitioning or proliferating cells were present in the adult liver, likely due to its mature state. Interestingly, we observed that on the course of liver regeneration, the activity of regulon CEBPA was similarly distributed on both transitioning and proliferating cells while its activity was regained faster on quiecent cells (Figure 6E, top). On metabolically hyperactive cells, the activity density of CEBPA was bimodal, with some cells regaining CEBPA activity at faster pace than others (Figure 6E, top). In the process of recovering the intermediate metabolic functions, metabolically hyperactive cells activated CEBPA at different paces, as activation of metabolic routes on hepatocytes depend, among other things, in the complex crosstalk between *CEBPA* and other important TFs such as *FOXA1* and *HNF4A* (*58, 59*), which led to bimodalities in the CEBPA regulon activity density. An analogous behaviour was observed on the regulon activity distribution of CEBPB, although following the opposite pattern (Figure 6E, bottom).

Further, regulons with an initiation-progression pattern like CEBPB, YBX1 and FOXA3 (*55, 59*), showed higher activity during proliferation compared to other states (Supplementary Figure S6A). In contrast, regulons following a termination–rematuration pattern such as CEBPA, HNF4A and ONECUT2, which have more mature functionalities (*58*), were more active during the quiescent and metabolically hyperactive states (*55, 58*) (Supplementary Figure S6B). Note that the regulatory dissimilarity score captured the observed differences in distributions across timepoints (Figure 6E and Supplementary Figure S6A-B) in a single score, allowing the clustering of TFs that have similar regulatory dynamics (Figure 6A).

Further, we identified a correlation between the regulon activity and the pseudo-temporal transition of the hepatic cells at single-cell resolution (Figure 6C). Specifically, we observed a clear gradient of regulon activity for CEBPA, which had very low activity at early stages of regeneration (PHx24 and PHx48) and its activity progressively increased while cells progressed towards adult liver cells. The opposite behavior could be observed for the activity of regulon CEBPB, whose activity progressively decreased as the liver regenerated. These results show that SimiC is able to capture the regulatory dynamics of a regenerating liver at single-cell resolution.

Note that a similar analysis was performed by Chembazhi et. al. (*55*) using SCENIC. However, SCENIC cannot jointly compute different GRNs for all timepoints, and therefore, a unique GRN was computed using all data points. In the original analysis, while SCENIC was able to capture similar patterns, the activity differences captured by SCENIC across the different timepoints yielded smooth and unimodal distributions (Figure 4 of Chembazhi et. al. (*55*)). The smooth distributions yielded by SCENIC are a consequence of

computing a unique GRN for all cells and timepoints, which intrinsically averages out the different regulatory patterns across timepoints. On the other hand, the distributions yielded by SimiC across all cells were generally multimodal, which showcases the capability of SimiC to capture the differences in regulatory behavior across cell types. In addition, when clustering the activity scores across all cells, we showed that SimiC produced a more consistent clustering than that produced by SCENIC (Supplementary Figure S7, Figure 4 from Chembazhi et. al. (*55*)). Specifically, we showed that when clustering the activity scores of all cells within a given cell state, the cells belonging to the same timepoint generally clustered together (Supplementary Figure S7). Thus, SimiC was able to uncover the main differences in regulatory dynamics across timepoints for different cell types, offering deeper biological insights than previously available.

Taken together, we showed the importance of being able to jointly optimize, via a similarity constraint, several GRNs from related cell phenotypes (such as different timepoints). The joint optimization allows to better capture regulatory dynamics across timepoints, while maintaining similarities, than when a single GRN is inferred for the whole dataset.

SimiC infers regulatory dynamics on non-model organisms

We examined the performance of SimiC on non-model organism lacking highly curated annotations. Some of the previously proposed methods for scRNA-Seq GRN inference, such as SCENIC (*22*), rely on additional annotations of the transcriptome (such as motifs or TF-binding site coordinates) to perform the GRN inference, and hence they are not suitable for non-model organisms.

We applied SimiC to an existing dataset that provides single-cell brain transcriptomic data generated in western honey bees (*Apis mellifera*, GEO accession ID GSE130785 (*60*)). In the accompanying study, Traniello et. al. (*60*) performed scRNA-Seq on a whole-brain (WB) and mushroom body (MB) sample collected from two female honey bees. The whole-brain sample contained all brain structures, including sensory neuropils like the antennal lobe (AL), optic lobe (OL), and mushroom bodies (MB), a higher-order sensory processing region composing roughly 1/3rd of the bee brain (*61*) (Supplementary Figure S8A). In contrast, for the MB sample, all brain regions except the MB were removed prior to sequencing. Further, cells were clustered using Seurat (*10*), yielding 11 cell clusters (*60*) (Supplementary Figure S8B). Transcription factors were obtained from (*5, 62*). We utilized these data to see if SimiC could identify regulatory activity that is unique to cell types captured in the whole-brain but not in the mushroom body sample. After generating the two GRNs with SimiC (one for MB and one for WB, median adjusted R^2 on test data of 0.834, Supplementary Figure S2E; median of 3 TFs regulating each target gene, Supplementary Figure S2J), we computed the regulatory dissimilarity score between MB and WB cells, for the different cell clusters.

The cell clusters coming from the OL, the primary visual neuropil of the honey bee brain and also one of the largest and most molecularly diverse regions (*63*), clustered together by the dissimilarity score across TFs (right-hand side, Supplementary Figure S8C). These clusters were identified by expression of OL-specific markers like *scr*, *gad1*, *drgx*, *SoxN* (*64–66*), and others (Supplementary Excel File 2), and represent neuronal subtypes exclusive to the insect visual system that would have been completely removed for the mushroom body sample. Furthermore, OL neurons 3, the OL cluster specifically annotated as lobular T4/T5 neurons based on expression of *drgx*, *SoxN*, and *slo* (*65–68*), yielded one of the highest overall dissimilarity scores.

Besides, MB-specific neuronal subtypes, identified by expression of *Mblk-1* (*69, 70*) and *CaMKii* (*71*), also clustered together (left-hand side, Supplementary Figure S8C).

Further, we found that the Drosophila melanogaster orthologue of the immune signaling gene *Deaf1* (LOC412296) had more regulatory activity in the MB compared to WB (Supplementary Figure S8D). In support of our findings, Deaf1 has been specifically identified in the honey bee MB as a regulator of aggression (*62*). Furthermore, we found higher *Deaf1* activity in glia compared to the rest of the clusters (Supplementary Figure S8E). Glia make up only a small fraction of the insect brain but have been hypothesized to play a role in regulating the neurometabolic state underlying aggression (*72*). Therefore, our analysis adds to previous findings (*62*) by implicating a specific cell type–glial cells–in the generation of distinct behavioral states.

Taken together, these results offer important ground-truthing for the comparison of regulatory networks jointly inferred across samples and, more generally, demonstrate the applicability of SimiC to non-model organisms.

Discussion

In this work we introduced SimiC, a gene regulatory network (GRN) inference method for single-cell RNA-Sequencing (scRNA-Seq) data. SimiC expects as input the linear ordering between different cell phenotypes (only if more than two phenotypes are considered). SimiC infers a GRN per phenotype while imposing a similarity constraint that forces a smooth transition between the GRNs. This allows us to compare GRN architectures between distinct phenotypic states, with far-reaching implications for systems biology. In contrast, other GRN inference methods such as SINCERITIES (*20*) or SCENIC (*22*) do not impose any similarity constraints, and are therefore either applied to the whole scRNA-Seq dataset, generating a single GRN for all cell phenotypes, or to each phenotype independently, potentially missing relationships between cells from different phenotypes. We showed that jointly inferring phenotypically distinct GRNs can uncover regulatory relationships (or dissimilarities) that would have otherwise been missed, thus establishing a new approach to quantitating differences or dissimilarities between the GRNs of distinct cellular phenotypes. Another advantage of SimiC with respect to SCENIC is that the inferred regulatory relationships between target and driver genes are accompanied by a weight, indicating the strength and direction of the regulation, providing an extra layer of resolution. For example, factoring in the weights allowed us to recapitulate known activators and repressors of key target genes on CAR T cell therapy. Finally, SimiC works well across a range of systems, both model and non-model alike, and does not require detailed annotations to the transcriptome, whereas other methods rely on prior knowledge of specific features like TF-binding motifs.

In summary, we provided significant evidence to show that SimiC has the potential to reveal complex gene regulatory relationships across different phenotypes or timepoints, for model and non-model organisms. Furthermore, we anticipate that SimiC will be applicable to advancing our understanding of the relationship between brain transcriptomic state and behavioral variation, and GRNs have been shown to be a reliable surrogate of complex behavioral phenotypes (*5, 6*). SimiC builds on these previous studies by allowing for a direct, quantitative comparison of GRN structure between behavioral states. The key of SimiC is its

optimization function that jointly infers the GRNs that model each phenotype. Discovering or validating target-gene relationships can have immediate applications in several fields, including personalized medicine.

MATERIALS AND 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.

Data and Code Availability

- The code generated during this study is available at GitHub https://github.com/jianhao2016/simicLASSO_git.
- Original/source data for synthetic dataset in the paper is available at https://doi.org/10.13012/B2IDB-4996748_V1.

Notation

scRNA-Seq data are expressed as a matrix of dimension number of sequenced cells by 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 MAD(X_i) = 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 a 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 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 independent and identically distributed (i.i.d.) expression profiles, the common approach (*7*) is to minimize the approximation error $\frac{1}{2}|Y - XW|^2$. 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 (*3, 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 (*73*). 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.

Data imputation

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 (*12*). However, 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. We use MAGIC (*12*) for this purpose.

SimiC optimization algorithm

With the imputed scRNA-Seq expression data, the cell state labels for each cell, and the associated ordering, our optimization problem for GRN inference is defined as:

$$\begin{aligned} \min_{W^k, k \in [1:K]} f(W^1, W^2, \cdots, W^K) = \\ \min_{W^k, k \in [1:K]} \sum_{k=1}^{K} \frac{1}{2}|Y^k - X^k W^k|^2 + \\ \sum_{k=1}^{K} \lambda_1 |W^k|_1 + \sum_{k=1}^{K-1} \lambda_2 \left| W^k - W^{(k+1)} \right|_2^2, \end{aligned} \tag{2}$$

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.

Algorithmic 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 in Algorithm 1, 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.

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 $l \in [1, \ldots, K]$

$\widetilde{\nabla}_{W^l} f = \frac{2}{m} {X^l}^T (X^l W^l(t-1) - Y^l) + \lambda_1 \mathit{sign}(W^l(t-1))$

$\nabla_{W^l} f = \widetilde{\nabla}_{W^l} f + 2\lambda_2 \left(2W^l(t-1) - W^{l+1}(t-1) + W^{l-1}(t-1)\right)$

$W^l(t) = W^l(t-1) - \frac{1}{\gamma_l} \nabla_{W^l} f$

end

The hyper-parameter λ_1 controls the sparsity of the W^k matrices, whereas λ_2 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^2 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^2 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.

Activity score

To better understand how the activity changes in the cell population, we propose to use a new metric: the activity score. Similar to (*22*), we compute one activity score per driver gene and per cell. The input to the activity score workflow is the expression profile of the cell and the regulon, and the output is the relative activity of this regulon in the cell. Note that in the objective of Eq. (2), 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 activity 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:

$$\text{Activity Score (driver gene } i) = \frac{\sum_{t=1}^{T} \left(\sum_{n \leq t} \frac{1}{T} \widehat{W}_{i,n} \right)}{\sum_{t=1}^{T} \widehat{W}_{i,t}},$$

where T is the number of target genes in the regulon i. Note that the activity score 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 activity 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 activity score gets closer to 0. The activity 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 a higher precision measure of a driver's activity within the cell population. The activity scores 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 activity score 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.

Regulatory Dissimilarity computation

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

$$\delta(P, Q) = \frac{1}{2} \sum_{\omega \in \Omega} |P(\omega) - Q(\omega)| \tag{3}$$

Let $\{P_1, \ldots, P_K\}$ denote the set of activity score 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 versions of total variation for multiple distributions:

$$\delta_{\text{minmax}}(\{P_1, \ldots, P_K\}) = \frac{1}{K} \sum_{\omega \in \Omega} \left(\max_{P_i} P_i(\omega) - \min_{P_j} P_j(\omega) \right) \tag{4}$$

It can be easily verified that this metric 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 proposed 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 $K = 2$, this metric 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 select those 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 the activity score 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.

Filtering of regulons

In order to remove TFs with very little weight assigned to target genes (due to numerical imprecisions on the optimization algorithm), we apply a filtering based on the bayesian information criterion (BIC). Specifically, SimiC only keeps the smaller set of TFs needed to model at least 90% of its expression variance.

Thus, for each target gene, we order the TFs by the their associated weight and compute the squared sum of all the TF's weights. We then select the TFs with larger weight until the squared sum of their weight is greater than the 90% of all the weights squared. Mathematically, for target gene j, we define the score S_{ij} as:

$$S_{ij} = \frac{\sum_{k=1}^{i} W_{kj}^2}{\sum_{l=1}^{T} W_{lj}^2}.$$

We then find the smallest i that satisfies $S_{ij} \leq 0.9$, and select the first ith TFs (ordered by decreasing weight).

Generation of synthetic GRNs and the corresponding scRNA-Seq data

We generated a synthetic dataset with 5 states. The GRN for the first state is generated randomly, with the weight of an edge being equal to $+1$ with probability 0.1, to -1 with probability 0.1, and to 0 with probability 0.8. To generate the GRN of the next state, we start with the GRN of the previous state and randomly modify each of the edges with a conditional probability that depends on the previous state's network. Specifically, for

a nonzero value edge, we change its sign with probability 0.3 and we set the weight to 0 with probability 0.7. For zero-value edges, we keep the same value with probability 0.5 and change it to $+/-1$ with an equal probability of 0.25. Note that this scheme keeps consecutive incidence matrices at similar sparsity levels. We do this procedure for each state, sequentially. Finally, target gene's expression is generated by sampling from a negative binomial distribution with $n = 5$ and $p = 0.5$.

Computation environment

For all performed experiments, the SimiC framework, including the computation of the regulon activity scores, was run on a Linux server with a 80 cores Intel Xeon CPU and 512GB of memory. On the synthetic dataset, the whole SimiC framework took 7 minutes to run using 16 cores and 5GB of RAM. More specifically, the optimization problem was solved in 4 minutes, and the regulon activity scores were computed in 3 minutes. Note that the running time of SimiC grows roughly linearly with number of cells, number of driver genes and number of target genes. For comparison, SINCERITIES took around 2 minutes to infer the GRNs for the same dataset.

For the CAR T cell dataset, that contains a total of 16K cells, the inference of the GRNs (with 100 TFs and 1,000 target genes) took 32 minutes. On the other hand, SCENIC took around 9 minutes.

Authority and hub scores

To better understand how the genes are modulated, we represent them as a weighted network. This network is created with the R library *iGraph*, where the nodes are genes and the edges are the weights computed by SimiC. These weights represent the strength of the modulation, while the sign represents the polarity. To represent the changes over the different phenotypes, we compute the Kleinberg's authority and hub centrality scores (*46*) for each node on each phenotype, and the statistical comparison was assessed by a Wilcoxon signed-rank test.

ChIP-Seq analysis

In order to asses the goodness of the TF-target connectivity, we validate these connections with empirical binding data. To perform this analysis, we used the database provided by "Integrative ChIP-Seq analysis of regulatory regions (ReMap2020)" (*32*), which contains transcriptional regulators peaks derived from curated experiments coming from Human samples. The peaks were annotated using the R library *ChIPpeakAnno*.

The odds ratio and hypergeometric test were calculated for each TF, by locating on the database all TF's binding sites and annotating them to target genes being located at most 5Kbps upstream the binding site.

Acknowledgments

Funding:

Gobierno de Navarra, Departamento de Salud and Departamento de Industria, Proyecto Estrategicos, DESCARTHES and AGATA (MEC, JRR).

National Institutes of Health (NIH) R01HL126845, R01AA010154 (UVC, SB, AK).

Planning Grant Award from the Cancer Center @ Illinois (UVC, SB, AK).

NIH Tissue microenvironment training program, T32-EB019944 (SB).

Scott Dissertation Completion Fellowship, UIUC (SB).
Instituto de Salud Carlos III (ISCIII) PI20/01308 (TE, FP).
CIBERONC CB16/12/00489, co-financed with FEDER funds (TE, FP).
Investigador AECC award, "Asociación Española Contra El Cáncer", INVES19059EZPO (TE).
The Accelerator award CRUK/AIRC/AECC joint funder-partnership, EDITOR (FP).
Gipuzkoa Fellows from the Basque Government, Spain (IO).
H2020 Marie S. Curie IF Action, European Commission, Grant Agreement No. 898356 (MH).

Author contributions: Conceptualization: JP, GS, IO, MH, Methodology: JP, GS, IO, MH, Software: JP, GS, Formal analysis: GS, MH, Investigation: JP, GS Validation: IMT, MEC, UVC, SB, TE, JRR, Supervision: AK, FP, IO, MH, Writing—original draft: JP, GS, IO, MH, Visualization: GS, IMT, MEC, MH, Writing—review & editing: All authors.

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All data are available in the main text or the supplementary materials.

References

1. Basso, K. *et al.* Reverse engineering of regulatory networks in human B cells. *Nature genetics* (2005).

2. Barabasi, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature reviews genetics* (2004).

3. Champion, M. *et al.* Module analysis captures pancancer genetically and epigenetically deregulated cancer driver genes for smoking and antiviral response. *EBioMedicine* (2018).

4. Califano, A. & Alvarez, M. J. The recurrent architecture of tumour initiation, progression and drug sensitivity. *Nat. rev. Cancer* (2017).

5. Chandrasekaran, S. *et al.* Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. *Proceedings of the National Academy of Sciences* (2011).

6. Sinha, S. *et al.* Behavior-related gene regulatory networks: A new level of organization in the brain. *Proceedings of the National Academy of Sciences* **117**, 23270–23279 (2020).

7. Hernaez, M., Blatti, C. & Gevaert, O. Comparison of single and module-based methods for modeling gene regulatory networks. *Bioinformatics* (2020).

8. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. *Nature communications* (2018).

9. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nature Reviews Genetics* **20**, 257–272 (2019).

10. Stuart, T. *et al.* Comprehensive integration of Single-Cell data. *Cell* (2019).

11. Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nature methods* (2017).

12. Van Dijk, D. *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* (2018).

13. Chen, S. & Mar, J. C. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC bioinformatics* (2018).

14. Fiers, M. W. *et al.* Mapping gene regulatory networks from single-cell omics data. *Briefings in functional genomics* (2018).

15. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods* (2020).

16. Matsumoto, H. *et al.* SCODE: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics* (2017).

17. Specht, A. T. & Li, J. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics* (2017).

18. Chan, T. E., Stumpf, M. P. & Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems* (2017).

19. Woodhouse, S., Piterman, N., Wintersteiger, C. M., Göttgens, B. & Fisher, J. SCNS: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data. *BMC systems biology* (2018).

20. Papili Gao, N., Ud-Dean, S. M., Gandrillon, O. & Gunawan, R. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* (2018).

21. Yuan, Y. & Bar-Joseph, Z. Deep learning for inferring gene relationships from single-cell expression data. *PNAS* (2019).

22. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nature methods* (2017).

23. MacNeil, L. T. & Walhout, A. J. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome research* **21**, 645–657 (2011).

24. Hastie, T., Tibshirani, R. & Wainwright, M. *Statistical learning with sparsity: the lasso and generalizations* (CRC press, 2015).

25. Van de Sande, B. *et al.* A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nature Protocols* **15**, 2247–2276 (2020).

26. Logsdon, B. A. *et al.* Sparse expression bases in cancer reveal tumor drivers. *Nucleic acids research* **43**, 1332–1344 (2015).

27. Irrthum, A., Wehenkel, L., Geurts, P. *et al.* Inferring regulatory networks from expression data using tree-based methods. *PloS one* **5**, e12776 (2010).

28. Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* (2019).

29. Grün, D., Kester, L. & Van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nature methods* **11**, 637–640 (2014).

30. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* **4**, e309 (2006).

31. Granja, J. M. *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nature biotechnology* **37**, 1458–1465 (2019).

32. Chèneby, J. *et al.* ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic acids research* **48**, D180–D188 (2020).

33. Ghaleb, A. M. & Yang, V. W. Krüppel-like factor 4 (KLF4): what we currently know. *Gene* **611**, 27–37 (2017).

34. Bräutigam, L. *et al.* MGST1, a GSH transferase/peroxidase essential for development and hematopoietic stem cell differentiation. *Redox biology* **17**, 171–179 (2018).

35. Van Buul, J., Mul, F., Van Der Schoot, C. & Hordijk, P. ICAM-3 activation modulates cell-cell contacts of human bone marrow endothelial cells. *Journal of vascular research* **41**, 28–37 (2004).

36. Rouleau, P. *et al.* The calcium-binding protein S100A12 induces neutrophil adhesion, migration, and release from bone marrow in mouse at concentrations similar to those found in human inflammatory arthritis. *Clinical immunology* **107**, 46–54 (2003).

37. Milner, J. J. *et al.* Runx3 programs CD8+ T cell residency in non-lymphoid tissues and tumours. *Nature* **552**, 253–257 (2017).

38. Carlini, F. *et al.* Association of HLA-A and non-classical HLA class I alleles. *PloS one* **11**, e0163570 (2016).

39. Meininger, I. *et al.* Alternative splicing of MALT1 controls signalling and activation of CD4+ T cells. *Nature communications* **7**, 1–15 (2016).

40. Pike, K. A. & Tremblay, M. L. Protein tyrosine phosphatases: regulators of CD4 T cells in inflammatory bowel disease. *Frontiers in immunology* **9**, 2504 (2018).

41. Ju, S. *et al.* Gadd45b and Gadd45g are important for anti-tumor immune responses. *European journal of immunology* **39**, 3010–3018 (2009).

42. Duran-Struuck, R. *et al.* A novel role for the semaphorin Sema4D in the induction of allo-responses. *Biology of Blood and Marrow Transplantation* **13**, 1294–e1 (2007).

43. June, C. H. & Sadelain, M. Chimeric antigen receptor therapy. *New England Journal of Medicine* **379**, 64–73 (2018).

44. Guedan, S., Ruella, M. & June, C. H. Emerging cellular therapies for cancer. *Annual review of immunology* **37**, 145–171 (2019).

45. Sheih, A. *et al.* Clonal kinetics and single-cell transcriptional profiling of CAR-T cells in patients undergoing CD19 CAR-T immunotherapy. *Nature communications* **11**, 1–13 (2020).

46. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* **46**, 604–632 (1999).

47. Klein, J. & Sato, A. The HLA system. *New England Journal of Medicine* **343**, 702–709 (2000).

48. García-Ortiz, A. & Serrador, J. M. Nitric oxide signaling in T cell-mediated immunity. *Trends in molecular medicine* **24**, 412–427 (2018).

49. Bretones, G., Delgado, M. D. & León, J. Myc and cell cycle control. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **1849**, 506–516 (2015).

50. Puig-Kröger, A. & Corbí, A. RUNX3: a new player in myeloid gene expression and immune response. *Journal of cellular biochemistry* **98**, 744–756 (2006).

51. Haque, M. *et al.* C-Myc regulation by costimulatory signals modulates the generation of CD8+ memory T cells during viral infection. *Open biology* **6**, 150208 (2016).

52. Schuster, K. *et al.* Homeostatic proliferation of naïve CD8+ T cells depends on CD62L/L-selectin-mediated homing to peripheral LN. *European journal of immunology* **39**, 2981–2990 (2009).

53. Cannarile, M. A. *et al.* Transcriptional regulator Id2 mediates CD8+ T cell immunity. *Nature immunology* **7**, 1317–1325 (2006).

54. Kurachi, M. *et al.* The transcription factor BATF operates as an essential differentiation checkpoint in early effector CD8+ T cells. *Nature immunology* **15**, 373–383 (2014).

55. Chembazhi, U. V., Bangru, S., Hernaez, M. & Kalsotra, A. Cellular plasticity balances the metabolic and proliferation dynamics of a regenerating liver. *Genome research* (2021).

56. Osada, S., Yamamoto, H., Nishihara, T. & Imagawa, M. DNA binding specificity of the CCAAT/enhancer-binding protein transcription factor family. *Journal of Biological Chemistry* **271**, 3891–3896 (1996).

57. Jakobsen, J. S. *et al.* 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**, 592–603 (2013).

58. Argemi, J. *et al.* Defective HNF4alpha-dependent gene expression as a driver of hepatocellular failure in alcoholic hepatitis. *Nature communications* **10**, 1–19 (2019).

59. Tomaru, Y. *et al.* Identification of an inter-transcription factor regulatory network in human hepatoma cells by Matrix RNAi. *Nucleic acids research* **37**, 1049–1060 (2009).

60. Traniello, I. M. *et al.* Meta-analysis of honey bee neurogenomic response links deformed wing virus type A to precocious behavioral maturation. *Scientific Reports* (2020).

61. Strausfeld, N. J. Organization of the honey bee mushroom body: representation of the calyx within the vertical and gamma lobes. *Journal of Comparative Neurology* **450**, 4–33 (2002).

62. Shpigler, H. Y. *et al.* Honey bee neurogenomic responses to affiliative and agonistic social interactions. *Genes, Brain and Behavior* (2019).

63. Strausfeld, N. J. *Arthropod brains: evolution, functional elegance, and historical significance* (Belknap Press, 2012).

64. Yoo, S. *et al.* Knock-in mutations of scarecrow, a Drosophila homolog of mammalian Nkx2. 1, reveal a novel function required for development of the optic lobe in Drosophila melanogaster. *Developmental biology* **461**, 145–159 (2020).

65. Davie, K. *et al.* A single-cell transcriptome atlas of the aging Drosophila brain. *Cell* **174**, 982–998 (2018).

66. Contreras, E. G. *et al.* The transcription factor SoxD controls neuronal guidance in the Drosophila visual system. *Scientific Reports* **8**, 1–12 (2018).

67. Pankova, K. & Borst, A. Rna-seq transcriptome analysis of direction-selective T4/T5 neurons in Drosophila. *PloS one* **11**, e0163986 (2016).

68. Schilling, T., Ali, A. H., Leonhardt, A., Borst, A. & Pujol-Martí, J. Transcriptional control of morphological properties of direction-selective T4/T5 neurons in Drosophila. *Development* **146** (2019).

69. Paul, R. K., Takeuchi, H. & Kubo, T. Expression of two ecdysteroid-regulated genes, Broad-Complex and E75, in the brain and ovary of the honeybee (Apis mellifera L.). *Zoological science* **23**, 1085–1092 (2006).

70. Suenami, S., Oya, S., Kohno, H. & Kubo, T. Kenyon cell subtypes/populations in the honeybee mushroom bodies: possible function based on their gene expression profiles, differentiation, possible evolution, and application of genome editing. *Frontiers in psychology* **9**, 1717 (2018).

71. Sun, D., Lye-Barthel, M., Masland, R. H. & Jakobs, T. C. The morphology and spatial arrangement of astrocytes in the optic nerve head of the mouse. *Journal of Comparative Neurology* **516**, 1–19 (2009).

72. Rittschof, C. C. & Schirmeier, S. Insect models of central nervous system energy metabolism and its links to behavior. *Glia* **66**, 1160–1175 (2018).

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