Deriving Cell Type-Specific Directed Weighted Signed Regulatory Networks from Single-Cell RNA Sequencing Data

Larisa Morales-Soto [1] **Juan P. Bernal-Tamayo** [1] **Robert Lehmann** [1] **Subash Balsamy** [1]
Xabier Martinez-de-Morentin [2] **Amaia Vilas-Zornoza** [3] **Patxi San-Martin** [3] **David Lara** [3] **Felipe Prosper** [4]
David Gomez-Cabrero [5 2] **Narsis Kiani** [6] **Jesper Tegner** [1]

Abstract

A quantitative characterization of gene regulatory networks (GRNs) that control cellular identity is key to our ability to reprogram cells, unlock developmental programs, and mitigate diseases. Here, we develop a data-driven technique for cell-type-specific GRN inference that uses RNA expression and velocity to give strength, direction and effect to each regulatory interaction. The method is evaluated in five public data sets from human and mouse, and a generated mouse B-cell differentiation data set. As validation we find that (i) the similarity of the inferred networks captures the similarities among different cell-types; (ii) the inferred weights permit the reconstruction of a potential (Hopfield) landscape, (iii) in which cell velocities agree with local cell-type-specific dynamics. The (iv) quality of the networks degrades smoothly when single-cell data from different cell-types are purposely mixed, thus demonstrating both robustness and cell-specificity of our method. To our knowledge, this is the first cell-type-specific GRN inference method that recovers directed, signed and weighted regulatory circuits directly from single-cell RNA sequencing data.

1. Introduction

Single-cell RNA sequencing technologies have established a new paradigm to study cell heterogeneity. Despite recent progress in the field, it has proved challenging to advance towards mechanistic studies of transcriptional regulation beyond clustering, cell type assignation, or trajectory inference. In spite of the difficulties that came with this technology, several methods for the inference of GRN from single-cell RNA seq data have been developed (Huynh-Thu et al., 2011; Kim, 2015; Moerman et al., 2019; Huynh-Thu & Sanguinetti, 2015; Specht & Li, 2017; Matsumoto et al., 2017). In a recent benchmark work (Pratapa et al., 2020), it was shown that these methods still lack accuracy and impose significant limitations on the data that can be analyzed due to methodological requirements. Instead of relying on time-series measurements or perturbation experiments, we sought to recover regulatory networks by coupling RNA expression and velocity. We approach this in a cell-type specific manner, as the corresponding cells may harbor strong regulatory signals characterizing that particular state, which would otherwise be obscured by cells from other groups.

2. Methods

2.1. GRN inference

For each cell type, we predict a network W as $X^{+}V + \gamma$. Where X^{+} is the pseudo-inverse of the expression matrix, V is the velocity matrix and γ is the degradation rate.

2.2. Calculation of cluster distances

A recurring problem in the field of GRN inference from transcriptomic data is the lack of state-of-the-art references to assess a method's performance (Pratapa et al., 2020). To circumvent this, we compute all the pairwise distances between clusters using either their count matrix or their inferred networks, and then use a set of statistics to assess the similarity between them. In this sense, an accurate set of networks should preserve the relationship between clusters found when using their entire expression data.

Expression space For every pair of expression clusters, the distance between them is defined as $d(C_i, C_j) = ||c_i - c_j||_2$. Where c_i and c_j are the centroids of clusters C_i and C_j respectively. And the centroid is the vector of average expression of all genes across all cells in the cluster. We used this as a reference distance matrix describing the true hierarchical relationship between clusters.

Network space For every pair of clusters, the distance between their corresponding networks is calculated with the Frobenius norm. As the networks of different clusters may comprise distinct sets of genes, before computing the distances, we match the gene sets in both graphs by adding small random numbers (< 0.001) as weights to the rows and columns of the genes missing in each network. We

compute the Frobenius distance between two graphs as $d(N_i, N_j) = ||N_i - N_j||_F$. Where N_i and N_j are the weighted adjacency matrices of cluster i and cluster j, respectively.

2.3. Cellular landscapes from neural networks

Here we compute an analog of Waddington's epigenetic landscape since we have the system equations for each cell-type or state. To this end, we follow the ideas proposed by Hopfield in his original manuscript (Hopfield, 1982), and interpret the inferred GRN as the interaction matrix of a system of neurons in a neural network, or Hopfield Network (HN), where each neuron (gene) can be in either of two states $ON = 1$ or $OFF = -1$, binarized as S. An energy function can therefore be defined (Hopfield, 1982), for the possible states using our inferred cell-specific GRN (interaction matrix W) as $H(S) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N} S_i W_{ij} S_j$.

2.4. Data sets

We used six real data sets from human and mouse. For the method development process we used an in-house generated data set of mouse B-cell development (mBD20; 8,095 cells). We then applied it to three data sets covering dynamic processes and two comprising differentiated cell types: the development of the mouse spinal cord (Delile et al., 2019) (mSC19; 81,933 cells), glutamatergic neurogenesis in the human fetal forebrain (La Manno et al., 2018) (hFB18; 1,720 cells), the generation of hematopoietic stem cells in human embryos (Zeng et al., 2019) (hED19; 4,805 cells), a mouse brain atlas (Zeisel et al., 2018) (mBA18; 97,186), and a compendium of peripheral blood mononuclear cells from the 10x Genomics 5K PBMC data sets (hPB20; 15,094 cells). For all data sets, cell labels were assigned using SingleR (Aran et al., 2019) and RNA velocity was inferred using scVelo (Bergen et al., 2019).

3. Results

3.1. The Inferred networks preserve cluster distances

The main factors to consider when predicting GRN's are the choice of genes and the cell annotation accuracy. Here, we only address the selection of genes, and rely on existing tools for clustering and cell-type labeling. Thus, to determine the group of genes that better captured the cell-type-specific features, we tested the agreement of the resulting network distance matrix with a reference distance matrix (see Methods). The best-performing settings for the data sets hFB18, hPB20 and mSC19 were 250 genes based on previous analysis (data not shown); and 100 genes for the hED19, mBD20 and mBA18 data sets. There was a clear linear relationship between the network distance and expression distance for multiple cluster-cluster pairs (Figure 1 a-d), also shown by their high correlation coefficients (shown in each panel). Thus, we show that cell-type specific GRNs of at most 250 genes are able to retain the information necessary to differentiate between clusters.

The mouse data sets mBA18 and mSC19 had low Mantel correlations (-0.01 and 0.1 respectively) and the linear relationship between distance matrices was not as clear as that of the others. We consider this to be mainly a problem of cell type annotation as we were not able to replicate the clusters reported in the original publications with an automated pipeline. Due to the nature of our performance assessment method, the clustering step is critical, because an unclear grouping of cells would impact the reconstruction of the reference distance matrix, and hence have a low correlation with any network distance matrix. The biological justification is that only cells belonging to the same cell-type would be expected to carry relevant information of the GRN specific for that very cell-type. Therefore, these results suggest that a proper sorting (clustering) of cells according to true cell-types is essential.

Additionally, we used this strategy to compare our method with three GRN inference methods: GENIE3 (Irrthum et al., 2010), GRNBOOST2 (Moerman et al., 2019) and PIDC (Chan et al., 2017) in three data sets (hFB18,hPB20 and mBD20). Across all of them, the correlations between the expression and network distances, of the three methods, were significantly lower (data not shown). Thus, suggesting that our networks are better at capturing cell-type-specific properties underlying the hierarchical structure of the biological processes.

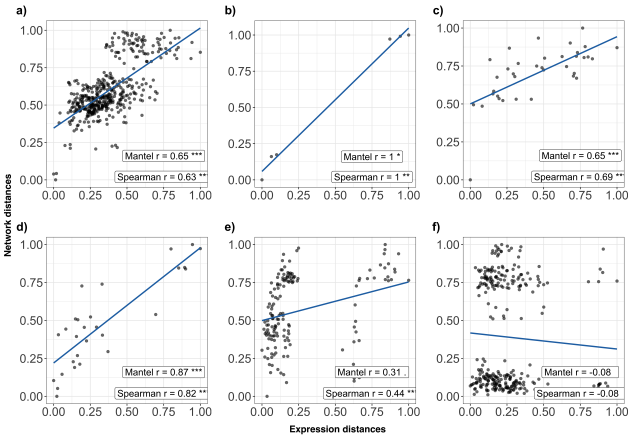

Figure 1. Comparison of cluster distances in expression and network. Each panel shows the normalized cluster distances for different data sets: a) hED19, b) hFB18, c) hPB20, d)mBD20, e)mSC19 and f) mBA18. Each point represents one cluster-cluster pair. Blue lines are linear regression models.

3.2. Derived cellular landscapes recover cell-type specific properties

To determine if these networks were capturing cell-type-specific dynamics, we used the inferred equations to model their corresponding Hopfield landscapes (Figure 2), as the cells' behavior on the landscapes is expected to reflect its phenotypic properties. Following Waddington's ideas, a terminally-differentiated cell type should display a landscape with at least one local minimum, in which the majority of cells should be. On the other hand, for cells that have not yet committed to a certain lineage, the landscape is expected to encompass several hills and basins, representing the multiple developmental pathways. Given the nature of single-cell RNA sequencing, these cells could be anywhere in that landscape, and their velocities could be pointing towards any of those pathways. To test these ideas, we examined the cell-type specific landscapes for each data set, and found topological patterns that agree with Waddington's propositions. For the hED19 data set, we can see that human embryonic stem cells display a very dynamic landscape (Figure 2 a and b), with multiple local minima and a high local maximum, suggesting that stem cells start at high a energy state and then lower their energy as they roll down the slopes into several basins of attraction (Figure 2 a). Moreover, these cells are distributed over a large range of the landscape (Figure 2 b), suggesting that these are cells that have already started their differentiation process. Interestingly, the velocities of these cells are pointing away from the center of the cluster, towards the slopes of the landscape (Figure 2 b), confirming the behavior expected for cells of this type.

Furthermore, we found an interesting pattern for endothelial cells in the hED19 data set (Figure 2 c and d). The topology of their landscape is much less diverse than that of embryonic stem cells (Figure 2 c), as we would expect for a cell type with a more restricted developmental fate. Accordingly, the majority of these cells are distributed over a valley in the middle of the landscape, with some cells placed higher in the slope, and others further down towards the basin (Figure 2 c and d). The individual velocities show a common trend, as they point from the top of the hill to the direction of a valley, possibly suggesting an uphill movement. In this case, we can observe a coordinated trend of cell velocities, as opposed to stem cells in which velocities point to multiple directions. This highlights the developmental commitment and cell heterogeneity of the generation of endothelial cells during human embryonic development. Similar properties were found on the neural-endothelial cluster of cells from the hFB18 data set (Figure 2 e), with the difference that the velocity field suggests a downhill movement of cells.

As we analyzed cell types closer to a terminal differentiation state, we noticed that, in general, their landscapes show a flatter topology compared to transitioning cells, and most of the cells are localized in a basin of attraction (Figure 2 e-h). Here, we only show the landscape of Monocytes from the hPB20 data set (Figure 2 f), Immature B cells from the mBD20 data set (Figure 2 g), and oligodendrocytes from the mBA18 data set (Figure 2 h), but similar results were found for other cell types in all data sets (data not shown).

Monocytes are generated from hematopoietic stem cells in the bone marrow and circulate through the bloodstream to their resident tissues, where they differentiate into multiple cell types. Even though they are in circulation for a short time, they can remain in a steady state even inside the resident tissue (Jakubzick et al., 2017). These characteristics appear to be reflected in their corresponding landscape, as they only display one basin and most of the cells are there (Figure 2 f). Although these cells can still differentiate into more specific types, their landscape's topology is what we would expect from cells that have completed their differentiation, thus highlighting the ability of monocytes to remain in a steady state. Moreover, this also suggests that their developmental fate is mostly driven by environmental signals, rather than predetermined in their transcriptional program.

Immature B cells are the last step of B cell differentiation in the bone marrow. After completing this stage, they migrate to the spleen as transitional B cells and then differentiate into mature B cells once they receive external queues within the tissue (Meffre et al., 2000). Interestingly, their landscape's (Figure 2 g) topology is consistent with their developmental trajectory. It shows two basins connected through a channel, suggesting that even in an intermediate state, immature B cells are stable enough to remain as such for approximately 3.5 days before continuing their maturation (Meffre et al., 2000). The second basin at the end of the channel could represent the mature state, requiring cells to cross an energy barrier (channel) through a transitioning state to reach it. Moreover, this behavior suggests that the external signals required to complete differentiation need not be very strong, as the cells appear to be in a primed state, in which the transcriptional profile is such that distinguishes them, but also allows differentiation to proceed easily.

Oligodendrocytes are fully developed glial cells that form the myelin sheath in the central nervous system (Bradl & Lassmann, 2010). As we would expect, most of the cells are in the local minima and the rest of the landscape is flat. Even though the unclear clustering of cells from the mBA18 data set affected the preservation of distances, our method was still able to extract the regulatory dynamics describing the overall behavior of this cell type Figure 2 h and others in the data set (data not shown). Thus, reaffirming that the low correlation of distance matrices might be due to a poor reconstruction of the reference.

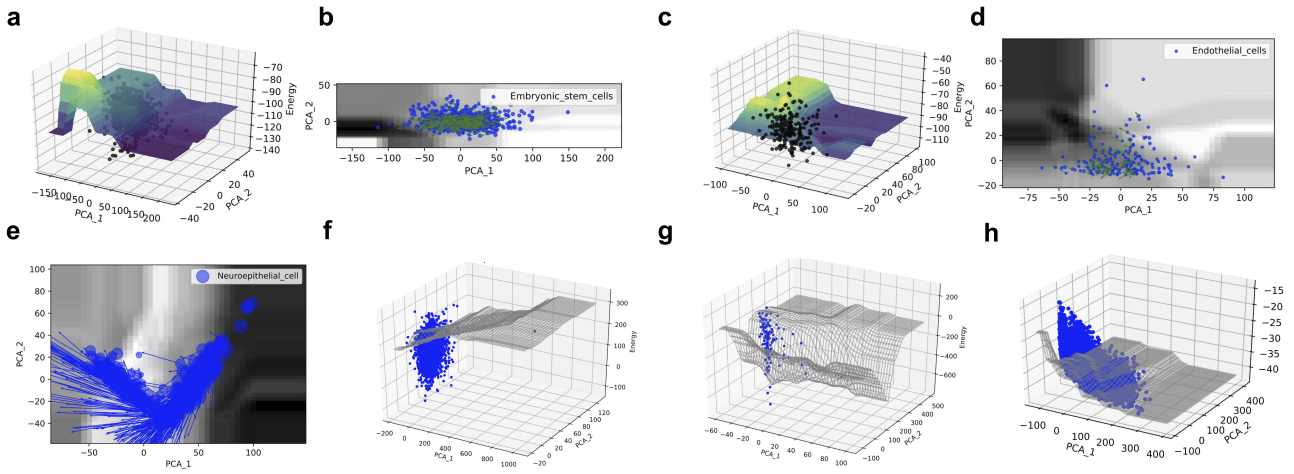

Figure 2. Cell-type-specific Hopfield Landscapes for multiple data sets. **a,b** Embryonic stem cells, **c,d** endothelial cells from hED19 data set. **e** Neuroepithelial cells from hFB18 data set. **f** Monocytes from hPB20 data set. **g** Immature B cells from mBD20 data set. **h** Oligodendrocytes from mBA18 data set. **a** and **c** Landscapes colored according to the energy values, with cells projected on top (black points). **b**, **d** and **e** landscape surface colored by energy values on a gray scale with cells and their velocities projected onto PCA. **f**,**g** and **h** cells projected on top of the landscape contour.

3.3. GRN inference is robust to cell down-sampling and replacement

To determine the extent to which the mislabelling of cells affects our method's ability to recover cell-type specific regulatory signals, we implemented a randomization framework in which we either downsample increasing proportions of cells from each cluster, or randomly replace increasing fractions of cells from each cluster with cells from other clusters. Here, the performance was quantified by the Jaccard index of the top weights in the network inferred from the down-sampled cluster and the one estimated using all the cells in the cluster. In both settings, the performance decreased when decreasing the number of cells retained from the original clusters (Figure 3), with (Figure 3 right panel) or when (Figure 3 left panel) replacing cluster cells with cells from other clusters (selected at random). This results suggest that even if we loose information from a cluster (Cell down-sampling) or mix cells from other clusters (Cell replacement) the main components in the inferred networks are still similar enough (lowest Jaccard index of 0.58) to the network recovered when all the cells are used. Highlighting why we are able to capture meaningful regulatory dynamics in the Hopfield landscapes for clusters from the MBA18 data set despite its sorting difficulties.

To further explore which genes harbor the regulatory dynamics characterizing each cell type, we also modelled their corresponding Hopfield landscapes with networks inferred using a random set of genes (the same number as the best network size for each data set). In all cases, we noticed that the landscapes display an almost flat topology, without any local minima or maxima (data not shown). Therefore, random sets of genes cannot capture cell-type specific features, highlighting the importance of this step when predicting GRNs, and further validating the cell-type-specificity of our method.

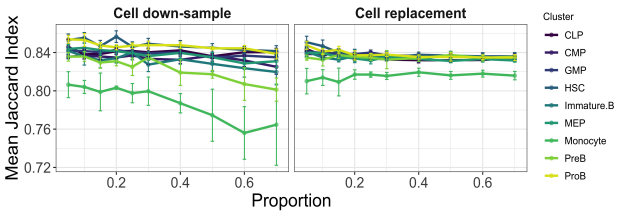

Figure 3. GRN inference robustness measured by the Jaccard Index of top edges in the mBD20 data set. Each point is the mean of 10 random samplings, error bars show the standard error of the mean.

3.4. Conclusions

Here, we developed a data-driven method to predict GRN from single-cell RNA sequencing data, that utilizes the transcriptional dynamics *within* each cell-type. We showed that these networks are robust to cellular perturbations, and can preserve the transcriptionally-derived hierarchical structure of clusters using only a few number of genes. Moreover, the inferred GRNs and the computed landscapes are sufficient to recover cellular dynamics in accordance with the notion of Waddington developmental potential for different cell

types across multiple data sets.

References

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2):163–172, 2019.

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J. Generalizing rna velocity to transient cell states through dynamical modeling. *bioRxiv*, 2019. doi: 10.1101/820936. URL https://www.biorxiv.org/content/early/2019/10/29/820936.

Bradl, M. and Lassmann, H. Oligodendrocytes: biology and pathology. *Acta neuropathologica*, 119(1):37–53, 2010.

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

Delile, J., Rayon, T., Melchionda, M., Edwards, A., Briscoe, J., and Sagner, A. Single cell transcriptomics reveals spatial and temporal dynamics of gene expression in the developing mouse spinal cord. *Development*, 146 (12), 2019. ISSN 0950-1991. doi: 10.1242/dev.173807. URL https://dev.biologists.org/content/146/12/dev173807.

Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. ISSN 0027-8424. doi: 10.1073/pnas.79.8.2554.

Huynh-Thu, V. A. and Sanguinetti, G. Combining tree-based and dynamical systems for the inference of gene regulatory networks. *Bioinformatics*, 31(10):1614–1622, 2015.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., Saeys, Y., and Geurts, P. Inferring gene regulatory networks from expression data using tree-based methods. 2011.

Irrthum, A., Wehenkel, L., Geurts, P., et al. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.

Jakubzick, C. V., Randolph, G. J., and Henson, P. M. Monocyte differentiation and antigen-presenting functions. *Nature Reviews Immunology*, 17(6):349–362, 2017.

Kim, S. ppcor: An r package for a fast calculation to semi-partial correlation coefficients. *Communications for statistical applications and methods*, 22(6):665, 2015.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., et al. Rna velocity of single cells. *Nature*, 560(7719):494–498, 2018.

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

Meffre, E., Casellas, R., and Nussenzweig, M. C. Antibody regulation of b cell development. *Nature immunology*, 1 (5):379–385, 2000.

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

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

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

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., Van Der Zwan, J., Häring, M., Braun, E., Borm, L. E., La Manno, G., et al. Molecular architecture of the mouse nervous system. *Cell*, 174(4):999–1014, 2018.

Zeng, Y., He, J., Bai, Z., Li, Z., Gong, Y., Liu, C., Ni, Y., Du, J., Ma, C., Bian, L., et al. Tracing the first hematopoietic stem cell generation in human embryo by single-cell rna sequencing. *Cell research*, 29(11):881–894, 2019.