1 Introduction

The advent of next-generation DNA and RNA sequencing technologies has revealed the content of the genome and begun to permit its functional characterization. However, understanding how genes interact with and regulate each other remains an area of active research. While analyses of gene co-expression [1] have elucidated modules of genes associated with disease [2], the specific regulatory interactions underlying these modules remain poorly understood. Inferring the structure of gene regulatory networks from expression data alone remains challenging due to the limited data provided by a single bulk RNA-seq experiment, and approaches integrating other types of data may yield more accurate networks [3].

Single-cell RNA-sequencing (scRNA-seq) can potentially improve the inference of regulatory networks from expression data, as a single scRNA-seq experiment yields hundreds or thousands of data points for a given cell condition [4]. However, the application of network inference methods to scRNA-seq data remains in its infancy, and scRNA-seq data has been shown to display different statistical characteristics from bulk RNA-seq data [5]. More work is required to understand the limitations of existing algorithms when applied to scRNA-seq data, and to evaluate newer algorithms designed for application to scRNA-seq.

Previously, the regulatory interactions predicted by various network inference algorithms have shown limited concordance [6]. In fact, prior comparative analyses of network inference from microarray data have found that no single algorithm out-performs its competitors, and an integrative consensus prediction from an ensemble of algorithms provides the most accurate regulatory predictions [7]. A comprehensive evaluation of a suite of network inference algorithms would therefore provide insight into identifying and designing an optimal strategy to infer regulatory networks from scRNA-seq data.

I propose performing a comparative analysis of a subset of existing network inference algorithms to appraise their relative performance on scRNA-seq data, focusing on methods which only incorporate expression data.

2 Approach

I plan to begin with comparing MERLIN [8], PIDC [9], and SILGMM [10]. If possible, I will extend my analysis to include other methods (NOTE: TBD based on discussions with Sushmita/Sunnie Grace/Viswesh).

Possible scRNA-seq datasets to analyze could include some of the following. (NOTE: final selection of datasets will also be based on our group discussion. I think it may make sense to reflect the species/cell types/scRNA technologies the lab is planning to analyze.)

1. Mouse embryonic stem cell (Fluidigm, n=704 cells) [11]
2. Human lung epithelial cell (Fluidigm, n=80 cells) [12]
3. Human peripheral blood mononuclear cell (10X, n=68,000 cells) [13]
4. Mouse neuron (10X v2, n=9,128 cells) [14]

5. Human breast epithelial cell (10X, n=24,646 cells) [15]

Prior efforts to compare network inference methods by the DREAM consortium have relied on simulated benchmark networks [16]. While simulated data theoretically offers a gold standard, it frequently makes simplifying assumptions that fail to test the limits of an algorithm on real biological data. A simple approach to evaluating the concordance of two algorithms would be to take the Jaccard similarity between their reported edges. A possible approach to evaluating the standalone performance of an algorithm would be to measure its sensitivity towards known, experimentally validated gene interactions, such as those reported in the TRRUST database [17] or provided by Gerstein et al. 2012 [18] or Neph et al. 2012 [19].

3 Significance

This project is of importance due to the opportunity that scRNA-seq presents to potentially reliably infer regulatory networks from expression data. While approaches integrating other types of data, such as protein-protein interactions [20] or protein binding sites experimentally identified with chromatin immunoprecipitation [21], have demonstrated an improvement in recapitulating regulatory network structure for a given cell condition, these methods have been less effective than non-prior-based algorithms at subsequently predicting regulatory interactions in a novel condition [21]. Further, omitting priors reduces the required complexity and cost of a network inference project by eliminating the need to perform complementary experiments to generate the priors. Characterizing the performance of existing network inference algorithms on scRNA-seq data will begin to illuminate the adaptations necessary to reliably predict regulatory interactions from expression data, and inter-algorithm comparisons may provide insight into developing a consensus approach.

The two primary results I anticipate obtaining from this project are an assessment of the algorithms' concordance when deployed on scRNA-seq data, and a measure of sensitivity towards recapitulating known regulatory interactions. I hope to generate some insight into modifications to the implemented algorithms that could improve network inference accuracy, including potentially outlining a framework for jointly integrating their results into a consensus network prediction. If possible, I would also like to evaluate how well the networks inferred by these algorithms in bulk RNA-seq data are replicated in scRNA-seq data, based on publicly available bulk RNA-seq data from equivalent species and cell types.

From this project I expect to gain experience working with single-cell RNA-seq data, familiarity with the theory underlying popular network inference algorithms and practical exposure to deploying them on real data, and experience in comparative analyses of network inference results. In addition, I hope to begin learning the limits of these methods and developing strategies to improve their shortcomings or build a consensus from their results.

References

[1] P. Langfelder and S. Horvath, "Wgcna: an r package for weighted correlation network analysis," *BMC bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[2] I. Voineagu, X. Wang, P. Johnston, J. K. Lowe, Y. Tian, S. Horvath, J. Mill, R. M. Cantor, B. J. Blencowe, and D. H. Geschwind, "Transcriptomic analysis of autistic brain reveals convergent molecular pathology," *Nature*, vol. 474, no. 7351, p. 380, 2011.

[3] D. Chasman, A. F. Siahpirani, and S. Roy, "Network-based approaches for analysis of complex biological systems," *Current opinion in biotechnology*, vol. 39, pp. 157–166, 2016.

[4] C. Trapnell, "Defining cell types and states with single-cell genomics," *Genome research*, vol. 25, no. 10, pp. 1491–1498, 2015.

[5] R. Bacher and C. Kendziorski, "Design and computational analysis of single-cell rna-sequencing experiments," *Genome biology*, vol. 17, no. 1, p. 63, 2016.

[6] R. De Smet and K. Marchal, "Advantages and limitations of current network inference methods," *Nature Reviews Microbiology*, vol. 8, no. 10, p. 717, 2010.

[7] D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, A. Aderhold, R. Bonneau, Y. Chen, *et al.*, "Wisdom of crowds for robust gene network inference," *Nature methods*, vol. 9, no. 8, p. 796, 2012.

[8] S. Roy, S. Lagree, Z. Hou, J. A. Thomson, R. Stewart, and A. P. Gasch, "Integrated module and gene-specific regulatory inference implicates upstream signaling networks," *PLoS computational biology*, vol. 9, no. 10, p. e1003252, 2013.

[9] T. E. Chan, M. P. Stumpf, and A. C. Babtie, "Gene regulatory network inference from single-cell data using multivariate information measures," *Cell systems*, vol. 5, no. 3, pp. 251–267, 2017.

[10] R. Zhang, Z. Ren, and W. Chen, "SILGGM: An extensive r package for efficient statistical inference in large-scale gene networks," *PLoS computational biology*, vol. 14, no. 8, p. e1006369, 2018.

[11] A. A. Kolodziejczyk, J. K. Kim, J. C. Tsang, T. Ilicic, J. Henriksson, K. N. Natarajan, A. C. Tuck, X. Gao, M. Bühler, P. Liu, *et al.*, "Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation," *Cell stem cell*, vol. 17, no. 4, pp. 471–485, 2015.

[12] B. Treutlein, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, and S. R. Quake, "Reconstructing lineage hierarchies of

the distal lung epithelium using single-cell rna-seq," *Nature*, vol. 509, no. 7500, p. 371, 2014.

[13] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature communications*, vol. 8, p. 14049, 2017.

[14] "10X single cell gene expression datasets." https://support.10xgenomics.com/single-cell-gene-expression/datasets. Accessed: 2018-10-09.

[15] Q. H. Nguyen, N. Pervolarakis, K. Blake, D. Ma, R. T. Davis, N. James, A. T. Phung, E. Willey, R. Kumar, E. Jabart, *et al.*, "Profiling human breast epithelial cells using single cell rna sequencing identifies cell diversity," *Nature communications*, vol. 9, no. 1, p. 2028, 2018.

[16] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky, "Revealing strengths and weaknesses of methods for gene network inference," *Proceedings of the national academy of sciences*, 2010.

[17] H. Han, J.-W. Cho, S. Lee, A. Yun, H. Kim, D. Bae, S. Yang, C. Y. Kim, M. Lee, E. Kim, *et al.*, "Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions," *Nucleic acids research*, vol. 46, no. D1, pp. D380–D386, 2017.

[18] M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K.-K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, *et al.*, "Architecture of the human regulatory network derived from encode data," *Nature*, vol. 489, no. 7414, p. 91, 2012.

[19] S. Neph, A. B. Stergachis, A. Reynolds, R. Sandstrom, E. Borenstein, and J. A. Stamatoyannopoulos, "Circuitry and dynamics of human transcription factor regulatory networks," *Cell*, vol. 150, no. 6, pp. 1274–1286, 2012.

[20] N. Novershtern, A. Regev, and N. Friedman, "Physical module networks: an integrative approach for reconstructing transcription regulation," *Bioinformatics*, vol. 27, no. 13, pp. i177–i185, 2011.

[21] A. F. Siahpirani and S. Roy, "A prior-based integrative framework for functional transcriptional regulatory network inference," *Nucleic acids research*, vol. 45, no. 4, pp. e21–e21, 2016.