**NIH Proposal Outline**

Six Page Limit – For Activity Codes R03, R13, R21, R36, SC2, SC3

1. **Introduction to Application**

The specific aims of our approach exhibit a separation of concerns

1. **Specific Aims**
2. Development of a novel methodology to infer directed gene regulatory networks from single-cell RNASeq data.
   1. Evaluate the utility of different correlation and dependency metrics in the construction of directed regulatory networks. Possible metrics to consider: Pearson correlation, Spearman corr, dCorr, MIC, and DREMI.
   2. Alternatives to Quadrant approach to directed logic network (gene regulatory network)
   3. from single-cell RNASeq data.
3. Visualize for hypothesis generation tool.
4. Leverage single-cell RNASeq data to develop novel cell state transition tree construction methodologies.
5. **Research Strategy**
   1. **Significance:** Gene dependence and correlation analyses have long been used to investigate the biological processes underpinning samples of interest. Recent work has been done regarding the susceptibility of traditional transcriptomic technologies to Simpson’s Paradox—the confounding of a mixture of signals that suggests a trend With the rise of single-cell RNAseq technologies, transcriptomics can now play a role in answering questions regarding tissue heterogeneity. This methodology promises a glimpse of the ‘rules of the game’ for a given progression across a dynamic cellular biological process.
   2. **Innovation:** To our knowledge no work has attempted to infer regulatory networks from single-cell RNASeq data.
   3. **Approach:** The specific aims of this proposal are presented in the order of intended execution. We believe that the completion of preceding aims will inform the completion of a given aim. In this section we will detail our intended approach to However, it should be noted that these aims still exhibit a separation of concerns, meaning that the success of one aim does not depend on the success of any other specific aim.

The first goal is to develop a novel methodology to infer regulatory gene networks from single-cell RNASeq data. To take full advantage of the resolution of this technology, the dataset(s) should include samples of likely different cellular subtypes. Therefore, tissue whose cells are suspected of going through a dynamic biological process such as proliferation, differentiation, or transition into malignancy may be of interest. Such a dataset would also be well suited for evaluation of aim 2. Furthermore, the number of samples per subtype will impact fthe accuracy of estimated gene dependency metrics. Taking cues from Marinov et al. who looked at various sized pools of single-cells subjected to the SMART-seq protocol and found that populations of 30-100 approach the information content of bulk tissue samples For this we will require a dataset…

Selecting the correct correlation and/or dependency metric(s) is an important sub aim to this work. Not only will different metrics limit the range of gene relationships that can be captured by our regulatory network, but each metric used will alter the interpretation of our resulting network. For these reasons we feel that it is necessary to encompass a broad range of metrics in our procedure, and comparing the regulatory networks that they each produce. We will also take each metric’s computational efficiency into consideration.

The DREMI metric [Eq. 4] is of particular interest to this investigation because it has been shown to expose functional relationships between variables whose joint probability is dominated by a seemingly independent signature (Dana Pe’er).

For instance there exist cell populations that are dominated by a particular cellular state or subtype (we’ll just say ‘state’ here for simplicity). That cellular state may only express genes X and Y in subsets of their total respective ranges when considering the more rare cellular states present in the sample. These rare signatures for X and Y may offer discerning information regarding the functional dependence of genes X and Y, unfortunately many traditional metrics are not well-equipped to pick up this rare state signature in the joint distribution that is dominated by signal in a smaller range.

DREMI de-convolution is achieved by estimating via a non-parametric diffusion kernel [Eq. 1] applied over bins of X values. Data is resampled according to this estimated conditional probability. And finally, mutual information [Eq. 2] is calculated for the down-sampled data. Essentially the authors standardize the joint probabilities density in bins of X. They show that this is equivalent to mutual information [Eq. 3] where every sample is weighted by [Eq. 4].

|  |  |  |
| --- | --- | --- |
|  |  | () |
|  |  | () |
|  |  | () |
|  |  | () |

These dependency metrics give reasonable measures for the strength of functional relationships between gene X and gene Y. The next step is to infer directionality of relationships in our network. In a previous work from our lab,

* + 1. **Aim 1**
       1. *Remove dominating signal.*
          1. *scLVM*
       2. *Compute correlation (DREMI, dCorr, MIC…) pairwise for genes/transcripts*
       3. *Quadrant approach to directed logic network (gene regulatory network)*
       4. *Conditional random fields?*
    2. Run monocle to:
       1. cluster cells
       2. Get tree of pseudotemporal ordering of cellular subtypes
    3. For each cluster (cellular subtype)
       1. Compute correlation (DREMI, dCorr, MIC…) pairwise for genes/transcripts
       2. Quadrant approach to directed logic network (gene regulatory network)
    4. For each pair of adjacent cell subtypes in tree:
       1. Compute correlation (DREMI, dCorr, MIC…) pairwise for genes/transcripts
       2. Quadrant approach to directed logic network (gene regulatory network)
    5. (Visualize dynamic directed gene inference network)
    6. Homology graph

1. **Preliminary Studies for New Applications:** normal text