Boolean Gene Implication Network Construction and Visualization from Single-cell RNA-seq Data

Specific Aims.

Gene expression based causality network construction has long been a goal of transcriptomic endeavors. The realization, of which, promises to unravel the convoluted genetic landscape of many biological contexts. Many publications in the past have contributed to this particular effort [1, 2, 3, 4, 5, 6, 7]. But, one unifying problem that all these studies share is their reliance on tissue sample transcriptomic data—a drawback that precludes reasoning and hypothesis generation at the cellular level. This is erroneous in practice because the majority of transcriptomic questions posed by investigators target molecular and cellular levels of biology. However, single-cell resolution has recently risen in popularity in the form of single-cell RNA sequencing (RNA-Seq) technology [8]. Proper utilization of this exciting new technology may help overcome the interpretation issues of past works and help identify functional relationships between genes that would be otherwise obscured by averaging effects of tissue sample transcriptomics.

Functional relationships between genes can also be obscured by drastic differences in the proportions of represented cell types or states within a dataset. Intuition for this comes from the consideration of rare cell types, which—although small in number—may harbor important, discerning information regarding the shape of a functional relationship between two genes. A number of works in the past have implemented density-dependent normalization techniques to counter this type of disproportional representation [9, 10, 11]. We propose the novel application of such techniques in tandem with the cellular resolution of single-cell RNA-seq data to construct Boolean implication networks between genes. Furthermore, Boolean implication networks are robust to noise owing to the requisite fuzzification of the data distributions under investigation. This quality is of particular interest in the context of Single-cell RNA-seq because the small amount of starting genetic material is relatively susceptible to bias. Our novel methodology promises to bring clarity of analysis never before seen in the study of gene expression implication.

Aim 1. To construct directed Boolean gene implication networks conducive to intuitive analysis of gene expression at the cellular level.

Approach: We will develop a novel methodology that combines the unobscured cellular resolution of single-cell RNA-seq data, the resampling methodology presented in *Conditional density-based analysis of T cell signaling in single-cell data* by *Krishnaswamy et al.*, and inspiration from the fuzzy Boolean implication network construction methodology of *Boolean implication networks derived from large scale, whole genome microarray datasets* by *Sahoo et al.*

Impact: Networks constructed by our novel workflow will be capable of capturing functional relationships implication Boolean implications resulting from the consideration of rare cell types/states that former methodologies would not be able to capture. Furthermore, our networks promise to be readily interpretable at the cellular and molecular genetics level.

Aim 2. To broaden the scope of this Boolean gene implication network construction method to analyze the changes in gene expression implication through a dynamic biological process.

Approach: We will append our work to consider Trapnelle et al.'s [12] pseudotemporal ordering strategy of single-cells in a dataset that represents cells at various stages in a biological progression. This will be accomplished by determining a clustering of samples, ordering those clusters in pseudotime, and then constructing an implication network for each cluster.

Impact: Breaking down samples by stages in biological progression allows for an unprecedented look into the dynamic implications of gene expression at and between stages in a biological progression.

Aim 3. Visualize for hypothesis generation tool while providing an array of organizational and information tools to aid investigators' navigation.

Approach: We will develop a visual encoding capable of capturing the range and characteristics of implications produced by the approach presented in aim 1. Furthermore, a software tool replete with navigation and organizational tools for the purpose of interactive hypothesis generation will be developed for the R environment.

Impact: This tool promises to make our results even more amiable to hypothesis generation over previous works by combining the interpretability of our results in aim 1 with the navigational and organizational tools of an interactive visualization tool.

Research Strategy

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 [13, 14, 15, 9, 16, 17]. Our proposed methodology promises a glimpse of the 'rules of the game' within this heterogeneity for a given progression across a dynamic cellular biological process. Pervious works suffer, in part, from poor interpretability given the unknown cellular composition of the input datasets. A single-celled perspective in theory does not suffer from the averaging effects of bulk sample transcriptomics data and therefore derived implications in this work will not suffer from the same interpretation issues. Furthermore, aim 2, provides an exciting first look into the dynamics of Boolean implication of a given dataset and biological context.

Innovation: To our knowledge no work has attempted to infer Boolean implication networks from single-cell RNASeq data. Neither have other publications attempted to combine the works of Sahoo's Boolean implication inference methodologies [2, 3] with the DREMI dependency metric [10]. We believe that the combination of these techniques will be able to detect potential bivariate gene implications that would otherwise be masked due to the rarity of certain cellular states in a given dataset. Finally, no work has proposed to look at the dynamics of Boolean gene implication networks in any biological context to our knowledge.

Approach:

Aim 1. To construct directed Boolean gene implication networks conducive to intuitive analysis of gene expression at the cellular level.

Our first goal is to develop a novel methodology to infer Boolean implication gene networks from single-cell RNASeq data. 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. DREMI is an application of mutual information [18] that measures the decrease in uncertainty of one variable given the value of another. To grasp the metric fully, consider the following definitions:

$$\text{Let } \mathbf{X} = \langle X_1, X_2, \dots, X_n \rangle \text{ and } \mathbf{Y} = \langle Y_1, Y_2, \dots, Y_n \rangle \text{ where } (X_i, Y_i) \sim f_{X,Y}(x, y)\ \forall\ i \in \{1,2, \dots, n\}$$

$$\text{Let } \mathbf{x} = \langle x_1, x_2, \dots, x_n \rangle \text{ and } \mathbf{y} = \langle y_1, y_2, \dots, y_n \rangle$$

$$\text{Let } d_i = (x_i, y_i)\ \forall\ i \in \{1,2, \dots, n\} \text{ and } \mathbf{D} = \langle d_1, d_2, \dots, d_n \rangle$$

Namely, $\mathbf{X}$ and $\mathbf{Y}$ are vectors of random variables while $\mathbf{x}$ and $\mathbf{y}$ are vectors of a sample's *X* and *Y* components from the joint distribution $f_{X,Y}(x, y)$.

Imagine that there may exist cell populations that are dominated by a particular cellular state or subtype (we'll just say 'state' here for simplicity). Values in $\mathbf{x}$ and $\mathbf{y}$ may be dominated by a specific sub-range of the spans of $\mathbf{X}$ and $\mathbf{Y}$. If values in the dominant cell state are centralized around some (x, y) value, then many metrics may

conclude that no relationship exists between genes *X* and *Y*. However, rare signatures for *X* and *Y* may offer discerning information regarding the functional dependence of genes *X* and *Y*. DREMI addresses this issue by subsampling according to an estimation of $f_{X|Y}(x|y)$ rather than $f_{X,Y}(x,y)$. We'll refer to this subsampling as $\mathbf{D}'$ along with the following definitions:

Let $\mathbf{X}' = \langle X'_1, X'_2, \dots, X'_n \rangle$ and $\mathbf{Y}' = \langle Y'_1, Y'_2, \dots, Y'_n \rangle$ where $(X'_i, Y'_i) \sim f_{X|Y}(x|y)\ \forall\ i \in \{1,2,\dots,n\}$

Let $\mathbf{x}' = \langle x'_1, x'_2, \dots, x'_n \rangle$ and $\mathbf{y}' = \langle y'_1, y'_2, \dots, y'_n \rangle$

Let $d'_i = (x'_i, y'_i)\ \forall\ i \in \{1,2,\dots,n\}$ and $\mathbf{D}' = \langle d'_1, d'_2, \dots, d'_n \rangle$

DREMI deconvolution is achieved by estimating $f_{Y|X}(y|x)$ 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. They show that this is equivalent to mutual information [Eq. 3] where every sample is weighted by $1/p(x_i)$ [Eq. 4].

$$\hat{f}(x;t) = 1/n \sum_{i=1}^{n} \sum_{k=-\infty}^{\infty} e^{-k\pi t/2} \cos(k\pi x) \cos(k\pi x_i) \quad (1)$$

$$I(Y;X) = H(Y) - H(Y|X) \quad (2)$$

$$I(Y;X) = E\big(-\log\big(p(Y)\big)\big) - E\big(-\log\big(p(Y|X)\big)\big) \quad (3)$$

$$I^c(Y|X) = E\left(\frac{-\log\big(p(Y)\big)}{p(X)}\right) - E\left(\frac{-\log\big(p(Y|X)\big)}{p(X)}\right) \quad (4)$$

The DREMI metric gives a reasonable measure for the strength of functional relationships between gene *X* and gene *Y*. The next step is to extract inference wherever possible. In other words, we want to turn our undirected network into a directed one. A number of previous works attempt to infer directionality in transcriptome data (Sahoo 2008). Previous work from our lab have successfully utilized the approach used by *Sahoo et al*. for the 'fuzzification' and directionality inference of relationships extracted from microarray data [1]. We intend to use the same basic methodology applied to the down-sampled technique described in *Pe'er et al*.

For a scatterplot of $\mathbf{D}'$, we start by discretizing the resampled by labeling a given gene expression value as either 'high' or 'low'. This is done by first ordering all observed values of $\mathbf{x}'$ from smallest to largest and fitting a step function to those ordered values using StepMiner [3]—an algorithm presented in *Sahoo et al*. These step functions aim to minimize the mean squared error (MSE) using an adaptive regression process. Next, the average of the high and low steps in a fitted step function serve to be the decision threshold between fuzzy 'high' and 'low' labels for gene *X*.

To discretize $\mathbf{y}'$, we may be able to take advantage of the P(Y|X) normalization effects that occur in the subsampling proposed by *Pe'er et al*. In our lab's original work, step function fitting was chosen, in part, because the minimization of MSE can still find a suitable boundary decision line in the presence of outliers. However, Pe'er's conditional probability normalization resampling methodology can serve to remove outliers. We propose that a simpler boundary decision can be found:

$$\frac{\min_i(y_i) + \max_i(y_i)}{2} \quad (5)$$

After discretization, we continue Sahoo's workflow to test for sparsity amongst the quadrants formed by the decision boundaries for *X* and *Y*. This is done by testing sample counts in a quadrant against the null hypothesis of uniformity in the distribution of $\mathbf{D}'$ samples. For instance, consider the test for the quadrant

corresponding to low values of both *X* and *Y*. Letting n_{LL}, n_{LH}, n_{HL}, and n_{HH} be the number of samples categorized with *X* and *Y* labels 'low' and 'low', 'low' and 'high, 'high and 'low', and 'high and 'high respectively. Sparsity of the low-low quadrant is determined as follows:

$$\text{total} = n_{LL} + n_{LH} + n_{HL} + n_{HH} \tag{6}$$

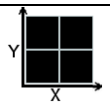
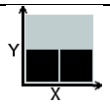
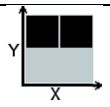
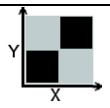
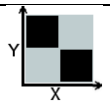
$$\text{expected}_{LL} = \frac{(n_{LL} + n_{LH}) * (n_{LL} + n_{HL})}{\text{total}} \tag{7}$$

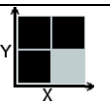
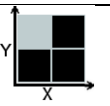
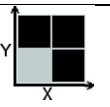
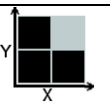
$$\text{statistic}_{LL} = s_{LL} = \frac{(\text{expected}_{LL} - n_{LL})}{\sqrt{\text{expected}_{LL}}} \tag{8}$$

$$\text{error rate}_{LL} = e_{LL} = \frac{1}{2}\left(\frac{n_{LL}}{n_{LL} + n_{LH}} + \frac{n_{LL}}{n_{LL} + n_{HL}}\right) \tag{9}$$

$$\text{quadrant}_{LL} = Q_{LL} = \begin{cases} 0 & s_{LL} > s \;\&\; e_{LL} < \varepsilon \\ 1 & \text{o. w.} \end{cases}, \text{ for thresholds } s \text{ and } \varepsilon \tag{8}$$

Analogous calculations are performed to determine Q_{LH}, Q_{HL}, and Q_{HH}. Now let, X_H, X_L, Y_H, and $Y_L$be Boolean variables that are true when, for a particular biological context, gene *X* or *Y* is highly expressed or not highly expresses (note that $X_H = \neg X_L$ and $Y_H = \neg Y_L$). Depending on the sparsity profile of the quadrants we propose that the following implications can be drawn.

Q_{LL}	Q_{LH}	Q_{HL}	Q_{HH}	Class	Conclusion
1	1	1	1		No implication
1	0	1	0		No implication
0	1	0	1		No implication
1	0	0	1		$X_H \Rightarrow Y_H \wedge X_L \Rightarrow Y_L$
0	1	1	0		$X_H \Rightarrow Y_L \wedge X_L \Rightarrow Y_H$

Q_{LL}	Q_{LH}	Q_{HL}	Q_{HH}	Class	Conclusion
1	1	0	1		$X_H \Rightarrow Y_H$
1	0	1	1		$X_L \Rightarrow Y_L$
0	1	1	1		$X_L \Rightarrow Y_H$
1	1	1	0		$X_H \Rightarrow Y_L$

The set of resulting, gene pair-wise Boolean implications constitute a Boolean implication network between 'fuzzified' expression values in a given biological context. An example of one such network containing 3 genes—A,B, and C—is illustrated in Figure 1.

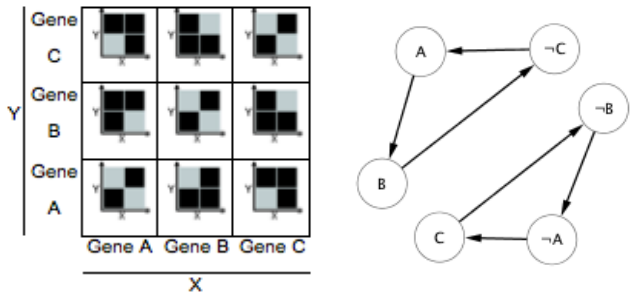


Figure 1| Example implications (Left), and implication network (Right) for a set of three genes A, B, and C.

Note that the conditional probability subsampling methodology proposed does not guarantee symmetric implications about *X* and *Y*. Non-symmetric patterns may be of interest biologically—potentially indicating the presence of a rare expression pattern.

Outcomes, Pitfalls, and Alternative Plans: The primary output of aim 1 is a Boolean implication network between genes. Major foreseeable obstacles for this step of the proposal revolve around the known noisiness of single-cell RNA-seq data. The prevalence of dropout events (missing values) has plagued the analysis of this technology. Applying appropriate normalization techniques and accounting for technical bias with spike-in controlled datasets, however, will greatly increase the signal to noise ratio of our data. Furthermore, massively parallelized single-cell RNA-seq data boasts not only increased sample sizes but also reduced technical bias owing to the utilization of unique molecular identifier (UMI) barcoding. Dropout events can be countered with low rank matrix estimation diffusion techniques or mixture model estimation.

Aim 2. Aside from gene expression profiles, functional relationships can also change between different cellular states and cellular subtypes. The challenge is to broaden our Boolean implication network to analyze the dynamics of a biological process.

Next, we attempt to decouple functional relationships between gene *X* and gene *Y* from conditional dependence upon a factor that *Trapnell et al.* refer to as 'pseudotime'—"a quantitative measure of progress through a biological process" [12]. The interplay of genes is known to change as a cell progresses through dynamic biological processes such as differentiation, cell cycle, or oncogenic transformation. Our approach may not be able to identify strong relationships between *X* and *Y* when considering cells from multiple cellular states if the true, underlying relationships between *X* and *Y* in those states differ. For example, consider the example illustrated by Figure 2. Clusters I and II, when considered together, yield no implication. However, partitioning of the samples gives rise to separate implications. Moreover, if these clusters represent cellular states along a biological progression, we may gain valuable information by analyzing the difference in the class of implications yielded by each partition.

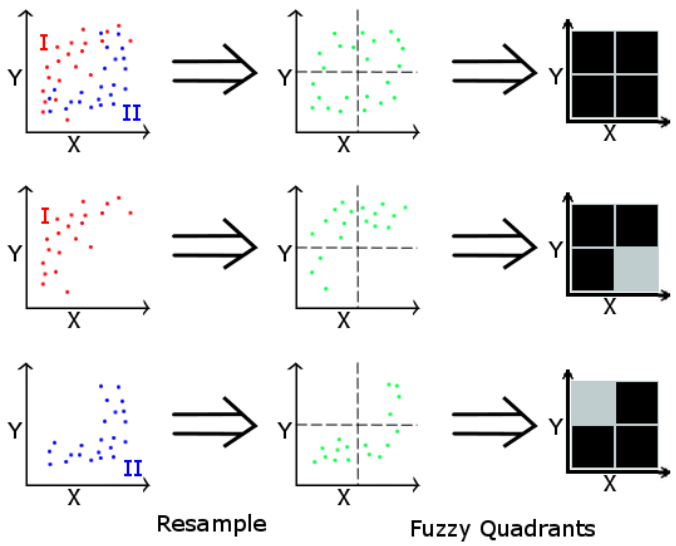


Figure 2| Somewhat contrived example of the convolution of signal that may exist in a dataset due to the combination of biologically differentiable cellular states.

To overcome this convolution of signals through pseudotime, we combine sample clustering with the temporal ordering output of an algorithm presented in *Trapnell et al.*—Monocle [12]. We plan to cluster samples in 2-dimensional independent component analysis (ICA) [19] space using hierarchical clustering. Let sets $C_1, C_2, \dots C_k$ define these output clusters such that $\bigcup_{j=1}^{k} C_j = \{1,2,\dots,n\}$. And then check for agreement between our clustering and Monocle's output as described below.

First, an overview of the algorithm in question: Monocle orders single-cells along an inferred smooth transition function $\vec{\rho}(t) = \vec{\gamma}(t) + \vec{\delta}(t)$ in Euclidean gene expression space. Here $\vec{\gamma}(t)$ represent the true set of expression values between states, and $\vec{\delta}(t)$ captures biological and technical noise. Monocle estimates $\vec{\rho}(t)$ by adapting a methodology that was first introduced by *Magwene et al* [20]. First, independent component analysis (ICA) [19] is performed on the data, and fitted with a minimum spanning tree (MST) in 2-dimensional independent component space. The diameter of the MST is taken as an estimate of $\vec{\rho}(t)$. The algorithm goes on to find potential orderings of samples in the data relative to $\vec{\rho}(t)$ using a PQ tree. A PQ tree defines a family of orderings of discrete elements in a set; it is a tree with two types of nodes—a Q node whose children are ordered (although reversible), and a P type node where children are permutable. So, a PQ tree is created with a single Q node denoted Q_{Main}. All vertices along the diameter of the MST with degree greater than 2 are deemed 'indecisive', and 'decisive' otherwise. The 'indecisive backbone' of the diameter is located—the longest sequence of vertices for which the endpoints are indecisive. All decisive vertices along the indecisive backbone are added to Q_{Main} in an ordered fashion. Then for any indecisive vertex along the indecisive backbone, a P node is appended to the tree and the indecisive vertex is added as a child of that P node. This same approach is then applied recursively to each branch of the indecisive vertex. Possible orderings of samples in pseudotime are given by those orders extractable from the final PQ tree.

Using the ordering proposed by the PQ tree with the shortest total distance in component space to define our samples' ordering in pseudotime, we can define an order between sample clusters. We'll let π be the ordered set of indices $\{1,2,\dots,n\}$ output by Monocle that orders our samples in pseudotime. Clusters $C_1, C_2, \dots C_k$ can then be ordered by $\Pi(j) = \frac{\sum_i \pi(i)}{|C_j|}, \forall i \in C_j$. We will say that a given clustering agrees with the MST output by Monocle if the variance of $\pi(i)$ per cluster is sufficiently small. We purposely supply this crude definition of 'agree' because clustering can be a highly customizable process by method selection and parameterization therein. Furthermore, we believe that the clustering achieved by hierarchical clustering and the ordering of

individual samples by Monocle will largely agree because, if two samples are assigned to the same cluster, they are relatively close to each other and Monocle will therefore likely assign indices that are relatively close to each other as well.

At this point, we will assume that $C_1, C_2, \ldots C_k$ constitute different cellular states separated by different progressions through pseudotime. By constructing Boolean implication networks by the approach outlined for aim 1 for each individual cluster, we are, in a sense, looking at relational dependencies between genes at different slices of pseudotime. More interestingly, maybe, are the Boolean implication networks that result from considering samples that belong to pairs of adjacent clusters, where adjacency is defined according to $\Pi(j)$ and Monocle's MST. Effectively we can construct the rules of genetic interplay for both major cellular states in a dynamic biological process as well as for the transitions between those states.

Outcomes, Pitfalls, and Alternative Plans: The primary output of aim 2 is a series of Boolean implication networks that represented of clustered and pairs of clustered samples. Secondary output includes the pseudotemporal ordering of clusters and the differentiation tree of cell states hypothesized by Monocle. Do to the partitioning required to constitute different states in pseudotime, sample-size will most likely be a limiting factor for success. Publicly accessible GEO dataset GSE65525 provides ~3000 UMI-barcoded, differentiating, embryonic mouse stem cells RNA-seq samples. This extreme number of samples will not only provide sufficient samples that can be separated by biological differentiation, but UMI barcoding can also prove to reduce noise.

Aim 3. Visualize for hypothesis generation tool while providing an array of organizational and information tools to aid investigators' navigation.

We will build an interactive visualization that will allow investigators to navigate the networks constructed by the approaches outlined in aim 1—complete with appropriate organizational and informational tools that can help the user with gene regulation oriented hypothesis generation.

First we wish to develop a concise visual encoding capable of conveying all classes of Boolean implication between gene pairs possible as output of aim 1 of this proposal. A traditional Boolean implication network visualization, as seen in Figure 1, requires $2m$ nodes ('high' expression and 'low' expression labels per gene). Such a separation can complicate hypothesis generation. For this reason, we strive for a visualization consisting of a one-to-one mapping between genes and nodes. However, because traditional Boolean implication network nodes capture the state of Boolean variables, the proposed consolidation requires a translation of this information into another encoding. We propose capturing the state of a given variable in an implication in the directed edge between nodes as seen in the center panel of Figure 3. Finally we propose to further consolidate the information contained in this graph by consolidating implication classes between two genes using edge glyphs as seen in the right panel of Figure 3.

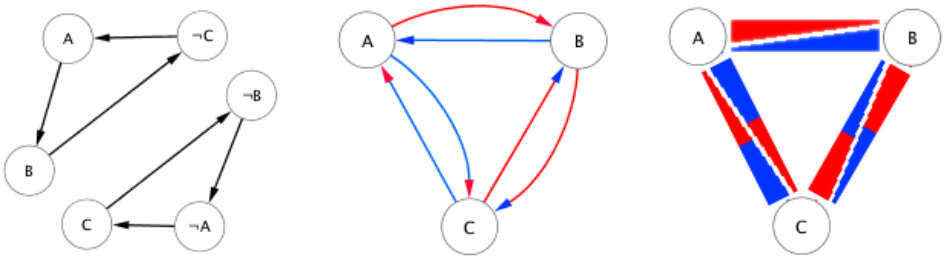


Figure 3| Example Boolean implication network from Figure 1 (Left). A node-consolidated visualization of the same example (Middle). Colors denote high (red) and low (blue) labels of genes in an implication relationship. Colors of arrow stems map to the upstream gene in an implication while arrowheads map to the downstream gene in an implication. And our proposed glyph encoding of edges (Right) which captures the same information.

Additional visual encodings:

- Edge thickness or transparency will be proportional to DREMI $I^c(Y|X)$.
- Node color will denote user-specified gene groupings.
- Pairs of directed Boolean implication classes will be encoded according to the following chart:

Our application will be provided as an R package developed using the web interface framework Shiny [21]. Using a web interface framework like Shiny that can also be run locally provides the flexibility of future extension into a web service without sacrificing benefits of local applications (latency, data security, etc…). Scatterplots and heatmaps will be realized using the ggplots [22] package and network data structures, manipulations (such as layout), and visualizations will be handled using the igraphs package [23].

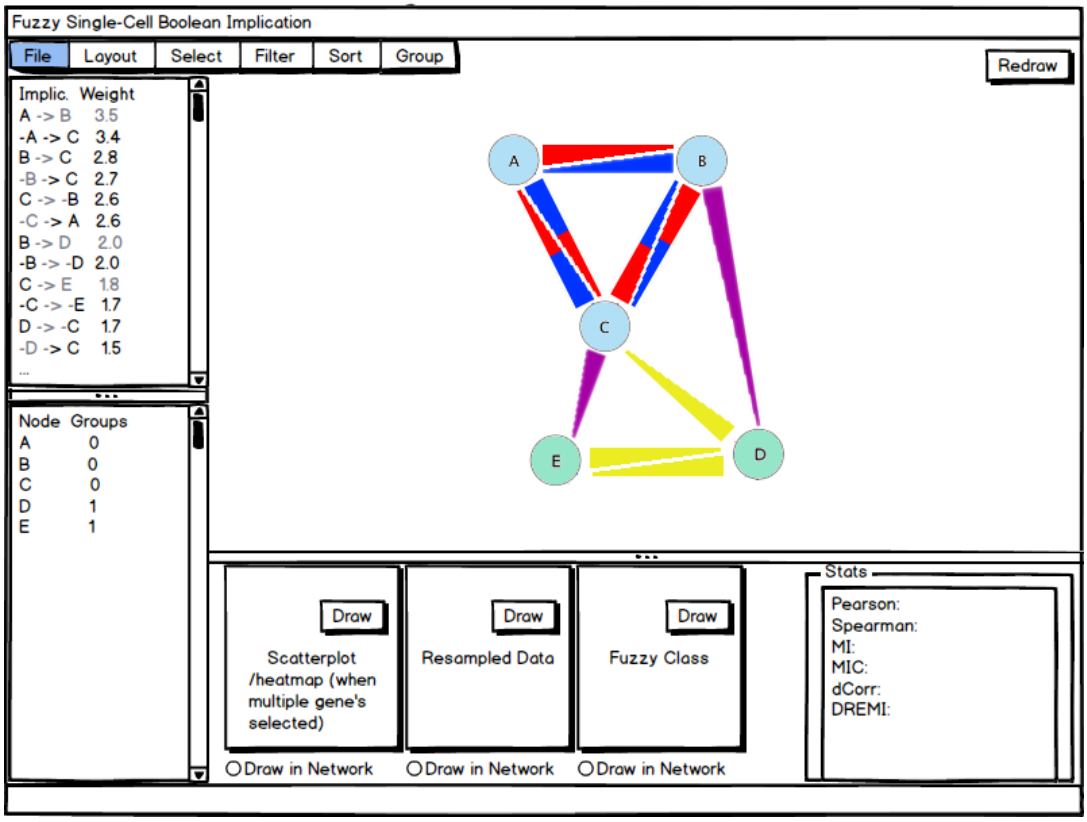


Figure 4| Mockup of proposed interactive visualization application.

Proposed features:

- Import/export of data matrices.
- Boolean implication network with aforementioned visual encodings.
    - Click and drag interaction support.
    - Automatic node layout algorithm support.
- Node organization tools: criteria/manual selection, filter, sort, and grouping.
- Edge organization tools: criteria/manual selection, filter, and sorting.

- Sample organization tools: criteria/manual filtering.
- Node information tools: histogram
- Node group information tools:
    - Color-coding user specified
    - Use biomaRt [24] to query gene selections against public genetics databases.
    - Heatmap
- Edge information tools:
    - Visualization of underlying scatterplots, resampled data, and implication class progression.
    - Various statistics that can be calculated on the fly for a given relationship.
- Pearson, Spearman, mutual information (MI) [18], maximum information coefficient (MIC) [25], distance correlation (dCorr) [26], and DREMI.
- Sample information tools:
    - Color-coding (to be viewed in scatterplots).
- Computationally intensive operations should be user initiated—never automatic.

Outcomes: The primary output of aim 3 is a visualization software package developed for the R programing environment that will aid investigators in hypothesis generation, navigation, and organization of the data structures output from aims 1 and 2.

Validation

Validation of our construction can be made by comparison with known biological pathways found in the KEGG Pathway database. Gene products in these pathways have a causal relationship. If latency is sufficiently low, edges of our implication networks should capture those causal relationships.

One can also think of an implication network as a large machine learning undertaking. With this perspective, classical cross-validation between split training and testing sets may help us glean the predictive utility of our construction.

To take full advantage of the resolution of this technology, the datasets selected for analysis 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 datasets would also be well suited for evaluation of aim 2. Furthermore, the number of samples per subtype will impact the accuracy of estimated gene dependency metrics.

We will validate our methods on stem cell differentiation datasets with a large number of samples to test our analytics. There are a plethora of publically available datasets that meet our criteria. The datasets here represent the range of sample sizes available.

The DREMI metric specifically will be evaluated in the presence of a rare cellular subtype. GSE60749 is an examination of pluripotent stem cells (PSCs) in *mus musculus* brain tissues [17]. A cluster analysis by *Kumar et al*. find that there were two clusters. One composed of 98% of the samples and the other only 8% or 14 individual cells. This breakdown may constitute a rare cellular state suitable to evaluate the utility of DREMI-based subsampling techniques.

GSE64016 provides 460 human embryonic stem cells (hESC)—213 H1 single cells and 247 H1-Fucci labeled single cells. With a palatable number of samples, and labeled cycle stage data, this dataset can help evaluate the accuracy of pseudotemporal ordering.

GSE65525 provides ~3000 UMI-barcoded, differentiating, embryonic mouse stem cells from a massively parallelized microfluidics-based single-cell sequencing technology called DropSeq [27]. This dataset may be able to alleviate obstacles that arise from insufficient sample sizes.

References

[1] A. W. M. S. H. C. K. H. A. Yates and R. Machiraju, "Visualizing Multidimensional Data with Glyph SPLOMs," Eurographics Conference on Visualization (EuroVis), 2014.

[2] S. D, A. J. Dill DL FAU Gentles, R. Gentles AJ FAU Tibshirani, S. K. Tibshirani R FAU Plevritis and P. SK, Boolean implication networks derived from large scale, whole genome microarray datasets., Department of Computer Science, Stanford University, Stanford, CA 94305, USA. FAU - Dill, David L, pp. --.

[3] S. D, R. Dill DL FAU Tibshirani, S. K. Tibshirani R FAU Plevritis and P. SK, Extracting binary signals from microarray time-course data., Department of Electrical Engineering, Stanford University, USA. FAU - Dill, David L, pp. --.

[4] R. H, R. S. Reynolds R FAU Varghese and V. RS, Increasing the efficiency of fuzzy logic-based gene expression data analysis., Intelligent Systems Laboratory, Department of Electrical and Computer Engineering, University of Maine, Orono, Maine 04469, USA. ressom@eece.maine.edu FAU - Reynolds, Robert, pp. --.

[5] M. AA, K. Nemenman I FAU Basso, C. Basso K FAU Wiggins, G. Wiggins C FAU Stolovitzky, R. Stolovitzky G FAU Dalla Favera, A. Dalla Favera R FAU Califano and C. A, ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context., Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA. adam@dbmi.columbia.edu FAU - Nemenman, Ilya, pp. --.

[6] W. PJ and W. Y, A fuzzy logic approach to analyzing gene expression data., Bioinformatics, Department of Molecular Biology, Parke-Davis Pharmaceutical Research, Warner-Lanbert, Ann Arbor 48105, USA. FAU - Wang, Y, pp. --.

[7] Y. Cai, B. Fendler and G. S. Atwal, "Utilizing RNA-Seq data for cancer network inference," in Genomic Signal Processing and Statistics,(GENSIPS), 2012 IEEE International Workshop on, 2012.

[8] K. AA, K. JK, S. V, M. JC and T. SA, The technology and biology of single-cell RNA sequencing., European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK; Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. Electronic address: saraht@ebi.ac.uk.,, pp. --.

[9] Q. P, S. C. Simonds EF FAU Bendall, K. D. J. Bendall SC FAU Gibbs, R. V. Gibbs KD Jr FAU Bruggner, M. D. Bruggner RV FAU Linderman, K. Linderman MD FAU Sachs, G. P. Sachs K FAU Nolan, S. K. Nolan GP FAU Plevritis and P. SK, Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE., Department of Radiology, Stanford University, Stanford, CA, USA. pqiu@mdanderson.org FAU - Simonds, Erin F, pp. --.

[10] K. S, S. MH, M. M, B. SC, L. O, S. E, P. D and N. GP, Systems biology. Conditional density-based analysis of T cell signaling in single-cell data., Baxter Laboratory in Stem Cell Biology, Department of Microbiology and Immunology, Stanford University, Stanford, CA, USA., pp. --.

[11] J. Zhang, L. Yu, H. Zhu and G. Sun, "Density and Non-Grid based Subspace Clustering via Kernel Density Estimation," dicode-project.eu, 2015.

[12] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," Nat Biotech, vol. 32, no. 4, pp. 381-386, #apr# 2014.

[13] B. F, N. KN, C. FP, P. V, S. A, T. F. A.-O. 0000000224191943, T. SA, M. JC and S. O. A.-O. 0000000288187193, Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells., European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK., pp. --.

[14] A. el AD, M. D. Davis KL FAU Tadmor, E. F. Tadmor MD FAU Simonds, J. H. Simonds EF FAU Levine, S. C. Levine JH FAU Bendall, D. K. Bendall SC FAU Shenfeld, S. Shenfeld DK FAU Krishnaswamy, G. P. Krishnaswamy S FAU Nolan, D. Nolan GP FAU Pe'er and P. D, viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia., Department of Biological Sciences, Columbia Initiative for Systems Biology, Columbia

University, New York, New York, USA. FAU - Davis, Kara L, pp. --.

[15] A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suvà, A. Regev and B. E. Bernstein, "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma," Science, 2014.

[16] H. L, B. F and T. FJ, Diffusion maps for high-dimensional single-cell analysis of differentiation data. LID - btv325 [pii], Institute of Computational Biology, Helmholtz Zentrum Munchen 85764 Neuherberg, Germany and Department of Mathematics, Technische Universitat Munchen 85748 Garching, Germany Institute of Computational Biology, Helmholtz Zentrum Munchen 85764 Neuherberg, Germany and Department of Mathematics, Technische Universitat Munchen 85748 Garching, Germany., pp. --.

[17] K. RM, C. P, S. AK, S. R, D. AJ, L. H, Z. J, P. K, G. D, T. JJ, F. TC, R. A, D. GQ and C. JJ, Deconstructing transcriptional heterogeneity in pluripotent stem cells., 1] Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts 02115, USA [2] Howard Hughes Medical Institute, Department of Biomedical Engineering, Center of Synthetic Biology, Boston University, Boston, Massachusetts 02215, USA., pp. --.

[18] C. Shannon, "A mathematical theory of communication," Bell System Technical Journal, The, vol. 27, no. 3, pp. 379-423, July 1948.

[19] H. A and O. E, Independent component analysis: algorithms and applications., Neural Networks Research Centre, Helsinki University of Technology, Finland. aapo.hyvarinen@hut.fi FAU - Oja, E,, pp. --.

[20] M. PM, J. Lizardi P FAU Kim and K. J, Reconstructing the temporal ordering of biological samples using microarray data., Department of Ecology and Evolutionary Biology, Yale University School of Medicine, New Haven, CT, USA. FAU - Lizardi, Paul, pp. --.

[21] W. C. et al., "Package 'shiny': Web Application Framework for R," https://cran.r-project.org/web/packages/shiny/shiny.pdf, 2015.

[22] W. W, S.-C. F and R. M, GOplot: an R package for visually combining expression data with functional analysis. LID - btv300 [pii], Department of Cardiovascular Development and Repair and Bioinformatics Unit, Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid, Spain.,, pp. --.

[23] G. Csardi and T. Nepusz, "The igraph software package for complex network research," InterJournal, vol. Complex Systems, p. 1695, 2006.

[24] K. RJ, S. K. A. F. Haider, J. H. S. F. Zamora, G. Z. J. F. Proctor, G. P. G. F. Spudich, J. S. G. F. Almeida-King, D. A.-K. J. F. Staines, P. S. D. F. Derwent, A. D. P. F. Kerhornou, P. K. A. F. Kersey, P. K. P. F. Flicek and F. P, Ensembl BioMarts: a hub for data retrieval across taxonomic space., European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. rhoda@ebi.ac.uk FAU - Kahari, Andreas,, pp. --.

[25] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher and P. C. Sabeti, "Detecting Novel Associations in Large Data Sets," Science, vol. 334, no. 6062, pp. 1518-1524, 2011.

[26] G. J. Sz{\'e}kely, M. L. Rizzo and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," The Annals of Statistics, vol. 35, no. 6, pp. 2769-2794, 2007.

[27] K. AM, M. L, A. I, T. N, V. A, L. V, P. L, W. DA and K. MW, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells., Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. Electronic address: marc@hms.harvard.edu.,, pp. --.

[28] A. K. Shalek, R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, J. J. Trombetta, D. Gennert, A. Gnirke, A. Goren, N. Hacohen, J. Z. Levin, H. Park and A. Regev, "Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells," Nature, vol. 498, no. 7453, pp. 236-240, #jun# 2013.

[29] D. P, E. Gong J FAU Syrkin Wurtele, J. A. Syrkin Wurtele E FAU Dickerson and D. JA, Modeling gene expression

networks using fuzzy logic., Virtual Reality Applications Center, Iowa State University, Ames 50011-3060, USA. FAU - Gong, Jian, pp. --.

[30] S. I and K. SA, Activities and sensitivities in boolean network models., Cancer Genomics Laboratory, University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, USA. FAU - Kauffman, Stuart A, pp. --.

[31] M. GK, K. W. B. F. McCue, G. P. M. K. F. Schroth, J. S. G. F. Gertz, R. M. G. J. F. Myers, B. J. M. R. F. Wold and W. BJ, From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing., Division of Biology, California Institute of Technology, Pasadena, California 91125, USA; FAU - Williams, Brian A,, pp. --.

[32] Z. I. Botev, J. F. Grotowski, D. P. Kroese and others, "Kernel density estimation via diffusion," The Annals of Statistics, vol. 38, no. 5, pp. 2916-2957, 2010.

[33] Z. A, M.-M. AB, C. S, L. P, L. M. G, J. A, M. S, M. H, H. L, B. C, R. C, C.-B. G, H.-L. J and L. S, Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq., Division of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, S-171 77 Stockholm, Sweden. sten.linnarsson@ki.se jens.hjerling-leffler@ki.se.,, pp. --.