Computational Stem Cell Biology
Project 6:
Report Due: Before Tuesday April 27th
Presentation: April 27th and April 29th

In this final project, you will work in teams of 2-4 members either to explore the nuances of network reconstruction (Option 1) or to grapple with a 'real-world' cell fate engineering controversy (Option 2). Both projects entail method development, benchmarking, and analysis of single-cell transcriptomic data.

Option 1: make a better scGRN reconstruction method

You will be asked to develop two of your own methods, one to reconstruct gene regulatory networks and one to reconstruct cell-cell interaction networks. The networks you reconstruct can then be used to postulate mechanisms driving cell fate decisions. Graduate student teams will be asked to develop a third method to predict cell fate potential.

You have been given two sets of files. The first refers to *in silico* data generated by thermodynamically simulating the dynamics of a synthetic GRN followed by a single-cell experiment. Such a dataset will be useful in testing any new methods you develop. The second dataset is a subset of the Bertie data from Project 5.

Files:

1. *In silico* data
   - *synthetic_expX_proj6.rda* : This is the cells-by-genes synthetic single-cell expression matrix (counts, not normalized: you can perform whatever preprocessing steps you deem necessary)
   - *synthetic_TFs_proj6.rda* : This is a list of TFs for this system
   - *synthetic_gs_proj6.rda* : This is the gold standard true network

2. *In vivo* data
   - *bertie_e7.5_proj6.rda* : This is expression and meta data from E7.5 of the same dataset used in Project 5.

3. Ligand-receptors
   - LR_pairs.rda: This data frame lists known and predicted ligand-receptor pairs.

Additionally, you will need to install the R package **minet** to perform CLR. Refer to the minet vignette for specific usage.

Part 1: Bulk and Single Cell GRN reconstruction

1. Simulate a bulk expression matrix. Create 30 samples by randomly selecting without replacement 25 single cells and averaging their expression.

2. Reconstruct the GRN using the synthetic single cell data provided. Reconstruct the GRN using the bulk data from the previous step. For this step, use CLR. The following code might be helpful:

```
mim <- build.mim(sample_by_genes_data,estimator="pearson")
net <- clr(mim)
```

3. Analyze and compare the performance of each method. Discuss. (Look back to project 1 for a refresher on PR curves and AUPR).

Part 2: Method Development

1. Devise your own single-cell GRN reconstruction method. When explaining, be sure to discuss shortcomings of existing methods and how your method addresses these issues.
2. Use your method to reconstruct the GRN of the synthetic single cell data. Compare performance to existing method. Discuss. Remember that performance can include a variety of metrics not limited to precision/recall, runtime, etc.

Part 3: Real Data

1. Load in the Bertie data. Cluster and assign identity via marker genes, or any of the automated methods we have covered in class. Feel free to convert to a Seurat object first.
2. Reconstruct GRN using two separate methods: CLR and your method from part 2.
3. Extract subnetworks. Are any cell type specific? Describe how you came to this conclusion.
4. Perform enrichment analysis to support your findings from #3.
5. Devise a method to infer cell-cell signaling interaction networks and utilize it to reconstruct the cell-cell interaction network of the data.
6. Look for downstream targets of these signaling pathways. Combined with your reconstructed GRN, postulate some specific examples of mechanisms driving cell fate decisions.
7. For graduate teams: devise a method to quantify cell fate potency as we have defined in class, and apply it to the Bertie data. Which cluster has the greatest fate potential, and how does this compare to your cluster annotation?

Grading:

1. Prepare a report that thoroughly addresses all questions, providing sufficient discussion. The report should be formatted in a cohesive manner, should include all generated figures, and should detail any methods you create in a logical, step-wise manner.
2. Include a separate file for any code that is written.
3. Prepare a 12-15 minute presentation focused on Part 2 and Part 3. Your presentation will be evaluated on how clearly you describe your methods, their performance, and the interpretation of their application to the Bertie data, as well as your response to questions.

Option 2: Are iPSC-endothelial blood brain barrier cells really endothelial, and if not, how can you improve them?

The past nine years has seen a rapid increase in the use of pluripotent stem cell derived endothelial cells to model the blood-brain-barrier (BBB). While there have been steady improvements in these derivation protocols, more recently there has been push back from some groups in the field as to the true identity of the engineered populations. Some have argued that the derived populations are more akin to neural epithelial cells than to endothelial cells. Lu et al 2021 have published data that support this argument at the bulk and single cell transcriptional level (see PNAS paper here). In this project, your objective is to (1) develop a pipeline to determine whether iPSC-BBB are more similar to true endothelial cells (EC) of the BBB or to some other cell type, and (2) to develop a method that predicts ways to improve the iPSC-BBB derivation protocol. Graduate student teams will be asked to develop a third method to predict cell fate potential. You should read the PNAS paper to gain a better understanding of the argument and the data in Lu et al. We have also included an article by Lippman et al that gives a different perspective on the question of PSC-BBB identity.

<u>Files:</u>

1.) TS_Ref_2021.h5seurat is single cell RNA-seq data from the Tabula Muris project. Specifically, these are non-myeloid cells from the mouse adult brain, inlcuding endothelial cells. This is your reference data. The cells are annotated under the "cell_type" slot. Note that these are mouse cells.

2.) Lu_Query_2021.h5seurat is single cell RNA-seq data from the Lu et al paper. It includes IMR90-iPSC-derived cells from day 0 to day 50 of differentiation, primary brain endothelial cells and vascular endothelial cells derived from umbilical cord (HUVEC), and cells that are also derived using a modified BBB-protocol (rEC). Note that these are human cells.

3.) oTab.csv: a table that will allow you to convert gene annotations from human to mouse.

To load the .h5seurat files into Seurat, you will need to install the SeuratDisk R package, and invoke:

```
xdata <- LoadH5Seurat("TS_Ref_2021.h5seurat")
```

For more information about SeuratDisk, see https://mojaveazure.github.io/seurat-disk/index.html

Part 1: Develop or adapt a pipeline to determine the similarity of human cells to a mouse reference data set.

Either devise your own cell-typing pipeline, or adapt one of the cell typing methods that we discussed in class. Provide a rationale for your approach. Evaluate the performance of the pipeline by applying it to well annotated-data that is independent of the training data. You will have to find this data yourself. To get you started, you should look at how cell typing methods that we covered in class were evaluated. Your assessment should include some of the metrics that we have discussed including accuracy and AUPR.

Part 2: Are standard PSC-BBB endothelial?

Apply your cell typing pipeline to the Lu et al 2021 data to determine whether the standard PSC-BBB cells represent endothelial cells of the brain. How does their identity shift over the time-course? If you determine that they do not resemble BBB ECs, then what are they? If needed, you can seek out other reference data sets to answer this question. Finally, how do the rECs fare?

Part 3: Devise and apply a method to yield improved protocols for cell fate engineering

Imagine that you want to use iPSC-BBB ECs for your project but you find current cells unsuitable for your needs. How would you go about improving the derivation process? What changes to the protocol would you make? Devise a computational method that will predict perturbations that will yield improved iPSC-BBB ECs. Provide a rationale for your method. The predictions it produces may include either over-expression or repression of transcription factors, as you explored in Project 3. Or the predictions can include modulation of signaling pathways (or a combination of both). However, at a minimum, the method must take as input scRNAseq of your starting population. Acceptable outputs would include a list of scored perturbations. Recall that directed differentiation entails a series of exposures meant to mimic development. Therefore, methods that predict the relative time of perturbation would be valued.

Apply your method to the improve standard iPSC-BBB. Also apply it to the rEC cells to see how they can be further improved. Discuss the predictions. Extra-credit if you can evaluate the predictions in silico by simulation.

Part 4: Define a metric of cell fate potency

For graduate teams: Devise a method to quantify cell fate potency as we have discussed in class. Apply it to the Lu data. How does your metric of fate potency track with time point of differentiation? Apply the method to the mouse reference data. Does you metric correspond to documented potency of the various progenitors in this data (i.e. interneuron, ESC, neuronal stem cell, oligodendrocyte precursor)?

Grading:

1. Prepare a report that thoroughly addresses all questions, providing sufficient discussion. The report should be formatted in a cohesive manner, should include all generated figures, and should detail any methods, and your quantitative evaluation of them, in a logical, step-wise manner.
2. Include a separate file for any code that is written.
3. Prepare a 12-15 minute presentation. Your presentation will be evaluated on how clearly you describe your methods, their performance, and the interpretation of their application to the Lu data, as well as your response to live question