The Molecular Language of the Body: Discovery of cell-to-cell communication molecules from RNAseq data

Lucia Guerri, Miranda Darby, Jingwen Gu, Amanda Bell, Van Truong, Saba Nafees

WHY?

- Fasten hypothesis generation in neuroscience by exploiting molecular knowledge from immunology.
- The study of cell-to-cell communication molecules has been particularly strong in immunology.
    - The same pairs of communication molecules employed by immune cells are employed differently by other tissues
- Cell-to-cell communication molecules between brain cells remains largely unknown
    - We can leverage RNAseq immune data to discover potential cell communication pairs in neuroscience

HOW?

- Use a RNA-seq input file for immune cell expression data
- Filter GO terms for membrane proteins only
- Pair the molecules based on known relationships
- Verify output
  - High score for known pairs of cell-to-cell communication molecules
- Our tool will identify brain cell molecule pairs with high interaction score

BRAINSTORM TOPICS:

- Binary Y/N for gene expression: For first pipeline only. Discuss cutoff (avoid complex normalization if possible) across data sets.
- scRNAseq as input: use expression of cell type clusters vs. single cell data. If clusters, then use mean of top quartile/quantile (?) expression value for each gene, to address technical loss of data. If single cell, think about how to address the cell pairs, and how to process that potentially heavy data.
- Machine learning: discover new potential pairs of molecules of cell-to cell-communication, instead of being limited by current knowledge to guide the scores. Discuss normalization strategy and challenges to exploit the actual level of expression, instead of binary Y/N expression.
- Soluble signals: how to deal with soluble signals and their receptors (membrane-bound and cytoplasmic) in this pipeline.
- Mouse vs human data: less “noise” vs higher N (=power)

GOALS:

- Supervised pipeline:
  - **RNAseq** analysis
  - **R programming language** (Supplemented with machine learning in Python)
  - **Scoring systems (statistics)**
  - Creative thinking to informatically define two datasets as “a pair” (for the pairs of cell types that talk to each other)
  - Basic understanding of biology and cell types
- Machine-learning partially-unsupervised tool:
  - **Machine-learning**! (ideally with Shiny)
  - WGCNA
  - All points for the supervised pipeline
- Adapt single-cell RNAseq to use as input:
  - **Single-cell RNAseq** analysis
  - All points for the supervised pipeline

TO DO

- Create unified dataset (Amanda and Miranda)
  - Set threshold of X # samples need to have this many reads
-