Literature List

The long and winding road of reprogramming-induced rejuvenation:
https://www.nature.com/articles/s41467-024-46020-5

Transfer learning enables predictions in network biology:
https://www.nature.com/articles/s41586-023-06139-9#citeas

- Used a rank based representation where genes are ranked by their expression in that cell normalized by their expression across the entire 30M single cell transcriptomes available. This is a representation of gene specificity to a particular cell type.

scGPT: toward building a foundation model for single-cell multi-omics using generative AI:
https://www.nature.com/articles/s41592-024-02201-0

The transcriptional landscape of age in human peripheral blood:
https://www.nature.com/articles/ncomms9570

Single cell transcriptomics and genomic changes in the aging human brain:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10659272/

A Single-cell and Spatial RNA-seq Database for Alzheimer's Disease (ssREAD)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10515769/

Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's Disease
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8766217/

Pre-trained Models

GeneCorpus trained on 30M single cell datasets:
https://huggingface.co/datasets/ctheodoris/Genecorpus-30M
scGPT https://huggingface.co/metehergul/scgpt

Brainstorming

Could potentially fine-tune these models that understand gene-gene coexpression patterns across many cell types for a particular purpose:

- Maybe something mechanistic i.e. predicting whether a gene belongs in a particular biological pathway (that was previously unknown) within a particular cell type. These could be aging related pathways. If we have a broader understanding of what genes are involved in pathways in a cell type specific manner, then we have a larger context for genes related to aging and specific mechanisms they may be involved in that we previously were unaware of.

- Can we somehow use the model to figure out whether ablating a particular gene (or combination of genes) between datasets profiled in “old versus young” or “diseased vs healthy” people would increase the cosine similarity of the embeddings? Looking at single or combinatorial gene ablations to understand what may be important in reversing a phenotype?
    - If we fine-tune the model to understand the difference between young or old OR control and disease, we can then ablate genes (i.e. knock them out) to see how much the resulting cell embeddings shift towards the “non-control” phenotype. Through this we can predict which genes might be involved in disease progression or aging.
    - We can also do the opposite and “over-express” genes by artificially increasing their gene rank

According to this article, there may be some difficulties with zero-shot learning for scGPT and GeneFormer: https://www.biorxiv.org/content/10.1101/2023.10.16.561085v2.full This should be less of a concern in our case where we are fine-tuning the model.

Running with the idea of fine-tuning the model on control vs perturbation dataset…
We can fine-tune the model to discern healthy transcriptomes from Parkinsons or dementia transcriptomes. I’ll have to find a repository of diseased single cell data.

This seems promising for Alzheimer’s: https://bmblx.bmi.osumc.edu/ssread/singlecell

    - Upon further inspection, that data portal is quite difficult to traverse and download the files… I think I will find a GEO dataset that I can download a tar file from with control/disease samples and convert the dataset into a format that is compatible with Geneformer.

Promising AD snRNA-seq dataset:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE174367

Promising parkinsons dataset:
https://singlecell.broadinstitute.org/single_cell/study/SCP1402/a-molecular-census-of-midbrain-dopaminergic-neurons-in-parkinsons-disease-preprint-data#study-download

https://linnarssonlab.org/loompy/semantics/index.html this seems to be the library that I should use for conversion into the format expected by Geneformer

***If this fine tuning task proves to be successful, as an added bonus we can run knockout or overexpression experiments to understand how the absence/presence of a particular gene within an input affects the output embeddings.

With regard to fine-tuning:

- We can take a look at combining all the datasets in a cell type agnostic manner and we can also try fine-tuning on a specific cell type and see how the various approaches perform?
- I think for a first pass, it might make sense to focus on one abundant cell type within a particular disease area and see how this approach performs

We could extend this to profiling mice (or humans) with a perturbation after taking a drug and potentially understand the genomic implications that factor into drug efficacy

Approach

Started doing some data exploration to understand the data and understand how to convert into the proper format required as input to GeneFormer. Once I've figured this out I will attempt to fine-tune the model on the transformed input data.

I've run into a CUDA error which I will need to navigate around, here is an article discussing this issue:
https://www.ml-illustrated.com/2020/02/11/how-to-setup-macs-cuda-for-tensorflow-pytorch.html

However my Macbook does not fit these criteria, so I will have to move development into a cloud setting where I can use a GPU.

My first pass with the Alzheimer's dataset was not fruitful, the model was hardly better than random. This may be due to a limited amount of cells to fine-tune on.

I made a second pass with the Parkinson's dataset and the model appears to be picking up on some signal to discriminate between healthy and Parkinson's afflicted single cell transcriptomes. I chose oligodendrocytes as the cell type to focus on due to the abundance of these cell type samples compared to other cell types.

Now that we have a fine-tuned version of Geneformer to predict Parkinson's afflicted cells, we will have to also consider simpler baselines to justify the use of this more advanced model.

My thoughts are to use something very simple like a logistic regression model and something middle of the road like a neural network and compare on the basis of AUROC, F1 score, and other metrics.
The Geneformer paper wasn't rich in details about how they trained their comparative baseline models. I will take the approach of using the tokenized "gene sets" as the features, and padding them to ensure equal dimensions.