Carter Lab Root

#Research_Root

Carter Network vs Spheroid Gene Spacial processing

HnE slide to Spacial Transcriptome Annotated Data Aggregation

Spacial Transcriptomics Lit Review

Notes:

Theres about 59 Visium datasets, but they average 60gb per sample most of which is in the fastq (Eg: of 68G, 49 is fastq)

- gunzip compression doesn't save much (~1Gb)
- Removed them for the 6 samples on server went from ~430 gb to 123gb (~20gb avg now)
  Theres another 34 Xenium samples that have the same issue
  imaging mass cytometry (IMC) records protien expression not RNA -- should i collect on this

TODOs

Download all Breast data
talk to Douglas abt which genes he want to vis
yashwin -- could spacial data predict mrd?

Main Hypothesis RN:

First run a CNN to pull features that will be fed as the node attributes into a GNN which will then incorporate spatial context into an MLP that will output an expresional map:
Can run pre-training esp for CNN on the lower res data and on non-cancerous non-tissue specific data to improve generalizability later bc want to find larger functional components
then train with higher res to get more accurate RNA exprsn profiles

alt is using a Vision Transformer (ViT) might be easier implementation but i think it might struggle given the issues with data

Questions

- Just want pancreatic cancer data?
    - More broadly
- what kind of features from spacial transcriptomics tells us about protease separation
    - what is the relevance fo protease separation
- Is in situ gene expression data give the same information as the HD spatial?
- can we use in situ gene expression data
    - Use RNA fluorescence (RNA-FISH) to gen high spatial res but only on specific genes
    - Basically how important is the lower expression genes/ ones not caught by RNA-FISH?
        - OR/ALSO how imp is the larger spacial context
- What types of variance is there in H&E slides?
    - Variation in 3d slicing?
    - Variantion in staining image rotation/orientation
    - Is there any standardization?

Maybe identification of stroma cells in pancreatic cancer could be a good starting point? effects tumor growth and therapy responce

Architecture Options

A. Convolutional Neural Network (CNN) + Multi-Layer Perceptron (MLP)
Vision Transformer (ViT) for H&E Feature Extraction + MLP for Gene Prediction
Spatial Graph Neural Network (GNN) + CNN/ViT

- **GNN for Spatial Relations**: Since spatial transcriptomics data is inherently spatial, using a GNN is a promising choice to capture spatial dependencies between spots.
    - Represent each spot as a node in a graph.
    - Use a GNN layer to capture spatial dependencies by learning from neighboring nodes (i.e., nearby spots in the H&E image).
- **CNN/ViT for Feature Extraction**: Use a CNN or ViT to extract features from each H&E patch, then use these features as node attributes in the GNN.
- **Output Layer**: The GNN's output for each node (spot) can be fed into a final MLP layer for gene expression prediction.
- This setup is powerful for spatially dependent data and is potentially more accurate in representing the biological context of gene expression.
  U-Net with Regression Head (Pixel-to-Spot Mapping)
- this probs not going to work
  https://www.youtube.com/watch?v=j3VNqtJUoz0&ab_channel=DeepFindr
  ^ vision transformers video ViT

Torch Geometric (for GNNs on spatial data) and **Transformers** (ViTs) in PyTorch for implementation.

• **scikit-image** and **OpenCV** for H&E image pre-processing.

• **Scanpy** or **Seurat** for handling gene expression data, as these packages offer utilities for dimensionality reduction and gene selection.

Zed feedback

label propogatoin unsupervised learning to help with poorly labeled data

Idk

- **Tangram:**
  - **Description:** Tangram maps bulk and single-cell RNA-seq data to spatial transcriptomics data.
  - **Limitations:** May not be directly applicable for deconvolution without spatial data.
  - **Usage:** Tangram Documentation
- **cell2location:**
  - **Description:** A tool for integrating single-cell and spatial transcriptomics data.
  - **Limitations:** Focused on spatial data; adaptation might be required.
  - **Usage:** cell2location GitHub

A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival
https://pmc.ncbi.nlm.nih.gov/articles/PMC6988279/#Sec8

Integrative multiomics-histopathology analysis for breast cancer classification
https://pmc.ncbi.nlm.nih.gov/articles/PMC8630188/#Sec8

https://github.com/hms-dbmi/breastCaPathologyTranscriptomics/tree/main/tumornorm_subtype

Predicting Breast Cancer Gene Expression Signature by Applying Deep Convolutional Neural Networks From Unannotated Pathological Images
https://pmc.ncbi.nlm.nih.gov/articles/PMC8673486/#:~:text=Abstract

Multimodal Deep Learning for Subtype Classification in Breast Cancer Using Histopathological Images and Gene Expression Data
https://arxiv.org/html/2503.02849#:~:text=Multimodal%20Deep%20Learning%20for%20Subtype,Images%20and%20Gene%20Expression%20Data

Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype
https://pmc.ncbi.nlm.nih.gov/articles/PMC6120869/#:~:text=developed%20an%20image%20analysis%20approach,accuracy%29.%20Sampling%20considerations%20in

Download Instructions

https://hbctraining.github.io/Accessing_public_genomic_data/lessons/accessing_public_experimental_data.html how to download geo data ncbi

```
library(GEOquery)#directlycall

eSet <- getGEO("[**GSE211956**](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE211956)",

destdir = '.',

getGPL = F)
```

DO this:

```
wget --recursive --no-parent -nd
ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE50nnn/GSE50499/suppl/
```

ST-preprocess Pipeline
TNBCtype:

JupyterNotebooks

```
jupyter-submit -p carter-compute -c 4 -m 128G -I
```

```
srun --partition=carter-compute --cpus-per-task=4 --mem=128G --pty bash
```