FastQ Data (10x)

Tar file walkthrough:

- https://bioinformatics.ccr.cancer.gov/btep/wp-content/uploads/Decompressing-files-with-the-tar-command-%c2%b7-AmyStonelakeBTEP-Wiki.pdf

Info on FastQ files:

- https://knowledge.illumina.com/software/general/software-general-reference_material-list/000002211
    - Format Glossary: https://help.basespace.illumina.com/files-used-by-basespace/fastq-files

H5AD data (Atlases)

Conversion for scanpy

- https://mojaveazure.github.io/seurat-disk/articles/convert-anndata.html

https://docs.scarches.org/en/latest/multigrate.html

- Multigrate walkthrough
- https://docs.scarches.org/en/latest/multigrate.html#Add-harmonized-cell-type-labels
    - Must use querying to synchronize celltype labels – we'll want to see what specific kinds we have and use our prior knowledge to align them
- AnnData object contains everything we need, including our labels to query on
- https://docs.scarches.org/en/latest/totalvi_surgery_pipeline.html
    - Tutorial for using scArches and multimodal totalVI base model (CITE-seq used in tutorial)

https://scanpy.readthedocs.io/en/stable/

- Scanpy documentation

https://anndata.readthedocs.io/en/latest/

- Anndata documentation

https://docs.scarches.org/en/latest/

- scarches documentation if we decide it could be helpful

Idea:

- https://github.com/satijalab/seurat-data
- Probably won't work because the datasets were generated by a model to begin with

Potential datasets we could use:

- https://www.10xgenomics.com/resources/datasets?query=&page=1&configure%5BhitsPerPage%5D=50&configure%5BmaxValuesPerFacet%5D=1000&refinementList%5Bproduct.name%5D%5B0%5D=Single%20Cell%20Multiome%20ATAC%20%2B%20Gene%20Expression

Healthy control data (some downloaded):

- https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/2.0.0/pbmc_granulocyte_sorted_10k?
- Multigrate 10x labels?: https://drive.google.com/drive/folders/1287j92xrWg8kbQa-hDCHDq4ECQo5pSxR

Covid 19 single cell datasets:

- RNA seq:
  - http://covid19.cancer-pku.cn/#/summary
  - https://cellxgene.cziscience.com/collections/187a3c52-3eb2-4c2b-9c3f-3f291f13bc3b?explainNewTab
- MULTIOMICS!!
  - https://www.medrxiv.org/content/10.1101/2021.01.13.21249725v1
    - "processed data" mentioned in above paper https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-10026
      - Separate csv includes celltype labels (I think) (this is true)
    - https://www.covid19cellatlas.org/index.patient.html
  - https://www-ncbi-nlm-nih-gov.revproxy.brown.edu/geo/query/acc.cgi?acc=GSE194122
    - CITE-seq (RNA + surface protein abundance) + "multiome" (RNA + ATAC)

Actual MultiGrate repo: https://github.com/theislab/multigrate

Class Discussion:

- Use pre-labeled cell-type data. It would be pretty convoluted (according to Prof Singh) to first make label predictions and then broadcast back and then encode…
- The architecture we currently have was shown to Prof Singh, and she called it logical