Project: Refined transcriptomic characterization of cell types in human middle temporal gyrus (2022)

Short Title: Human MTG - 10x (2022)
(Readme file created April 14, 2022)

Project Information at:
https://knowledge.brain-map.org/data/8G5ACH74WSK2R5DET0Z/summary

This data set includes single-nucleus transcriptomes from 166,868 total nuclei derived from five post-mortem human brain specimens. It is used to characterize cell type diversity in the middle temporal gyrus (MTG) for use with multiple projects (described below), and can be considered follow-up to the "Human MTG Smart-Seq (2018)" study (website, database, publication). These nuclei were collected as part of two separate efforts: an Allen Institute-funded project specifically targeting cortical areas and a National Institute of Mental Health grant (NIMH U01 MH114812-01) targeting cells across the whole human brain. Samples were processed using the 10x Chromium Single Cell 3'™ Reagent Kit v3 (link) or v3.1 (link). 10x chip loading and sample processing was done according to the Manufacturer's protocol. Gene expression was quantified using the default 10x Cell Ranger v6 pipeline with the CR6 genome annotation. For clarity samples are referred to as "cells" in this document even though RNA was collected only from each cell's nucleus.

CITE AS

Dataset: Allen Institute for Brain Science (2022). Allen Cell Types Database -- Human MTG 10x [dataset]. Available from celltypes.brain-map.org/rnaseq.

PROTOCOLS

Human Nuclei Isolation V2: https://dx.doi.org/10.17504/protocols.io.ewov149p7vr2/v2
10xV3 Genomics Sample Processing V.2: https://dx.doi.org/10.17504/protocols.io.bq7cmziw

DOWNLOAD INSTRUCTIONS

Data Collection: Human MTG - 10x (2022): processed.
This data set includes the analysis of single-nucleus transcriptomes from 166,868 total nuclei derived from the middle temporal gyrus (MTG) from five post-mortem human brain specimens. The counts data, along with associated cell type assignments and donor metadata, is open access.

Available from Repository: Allen Cell Types Database - Human MTG
Data Files are Accessible from URL: https://portal.brain-map.org/atlases-and-data/rnaseq/Human-MTG-10x_SEA-AD

ASSOCIATED PROJECTS

- **Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD)**: The Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) is a consortium focused on identifying and characterizing early changes in the brain in Alzheimer's disease and normal aging, that is funded by the National Institute on Aging (NIA U19 AG060909-01A1). This project includes 10x collected from human MTG from an aged cohort of 84 donors who span the full spectrum of disease severity. The Transcriptomics Explorer and all of its associated files are based on a taxonomy of "supertypes", which represent a refinement of the cell type taxonomy developed from the great ape study that includes stricter criteria for cluster separability (127 SEAAD refined supertypes).
- **Great ape (GA) study**: This study investigates cellular diversity in MTG across human and several non-human primate species: chimpanzee, gorilla, rhesus macaque, and common marmoset. It includes nuclei collected from individual cortical layers using SMART-Seq v4 (link) and from tissue sections using 10x (link) which provide a relatively unbiased survey of cell types. These files include 10x nuclei collected from human MTG for this study, along with cell type calls for the within-human taxonomy (151 identified cell types).
- **Human cross areal (CA) study**: This study investigates cellular diversity across human cortical areas. It includes nuclei collected from individual layers using SMART-Seq v4 (link) and from tissue section using 10x (link) which provide a relatively unbiased survey of cell types. These files include 10x cells collected from MTG for this study, along with cell type calls for the within-region taxonomy (124 identified cell types). The SMART-Seq v4 nuclei used in the GA and CA studies were collected as part of the "Human MTG Smart-Seq (2018)" study (link), and the data is available for download therein. Metadata for the CA and GA studies for these nuclei will be available elsewhere.
- **Human variation study** (link): This study seeks to characterize variation of gene expression in cell types of adult human cortex, and how this variation relates to genetic signatures. Cells from this study will be assigned cell types based on this SEA-AD taxonomy.

FILES IN THIS DIRECTORY

- **Gene expression data matrices (expression_matrix.csv):** This file contains one row for every cell in the dataset and one column for every gene sequenced. The values of the matrix represent a count of the unique molecular identifier (UMIs) for that gene (row) for that cell (column).

- **Cell metadata (metadata.csv):** This file contains metadata about each cell, including information about which cells were used in each experiment, their various cell type assignments, and associated donor information. There are a lot of columns in this file, which are organized by category below. Missing and "NA" values in any column indicate that that value was not determined or is unknown.
    - Donor and general cell metadata
        - sample_name: unique sample identifier, which links cells across files

        - external_donor_name_label: unique (de-identified) identifier for each human donor
        - donor_sex_label: biological sex of the donor
        - cortical_layer_label: cortical layer targeted for sampling
        - region_label: brain region targeted for sampling (in this case, middle temporal gyrus)
        - full_genotype_label: (these slots are left blank for human)
        - species_label: species the sample came from (in this case, Homo Sapiens)
        - age_label: age of each donor in years
    - Cell metadata related to the SEA-AD project. For compatibility with ingest into the Transcriptomics Explorer, these columns are not prefixed "SEAAD" but all of them refer to the SEA-AD project.
        - QCpass: an indication of whether a cell passed QC ("True"), failed QC ("False"), or was excluded from consideration in the SEA-AD study prior to the start of the analysis ("Not assessed"). Only cells with "True" in this column are included in the Transcriptomics Explorer.
        - cluster** (see note below for **): the SEA-AD cluster name used in the Transcriptomics Explorer. These are also called "supertypes" because they derive from the 'GA_cluster' calls below, after increasing the strictness of parameters used in determining whether clusters are distinct.
        - cell_type_accession_label: Globally unique ID of the cluster in "cluster label", also found in dend.json. Note that additional information related to cluster, subclass and class including cell_set_accession_id, cell_set_label, cell_set_alias, and cell_set_aligned_alias can be found in "nomenclature.zip".
        - cluster_confidence: scores ranging from 0 (low) to 1 (high) indicating the confidence that a given cell is mapped to the assigned SEAAD cluster.
        - subclass**: cell type subclass assigned in the SEA-AD project (for example, "SST", "L6 CT", and "Astrocyte"). Subclasses and classes (below) match standard interlex terms published in mammalian primary motor cortex (publication).
        - subclass_confidence: scores ranging from 0 (low) to 1 (high) indicating the confidence that a cell is mapped to the assigned SEAAD subclass.
        - class**: broad cell class terms ("Neuronal: GABAergic", "Neuronal: Glutamatergic", or "Non-neuronal and Non-neural"). This single column corresponds to SEAAD, GA, and CA taxonomies because every cell assigned a class in more than one taxonomy was assigned the same class in every taxonomy.
        - class_confidence: scores ranging from 0 (low) to 1 (high) indicating the confidence that a given cell is mapped to the assigned SEAAD class. Note that this value is not specified for CA or GA class calls.
    - Cell metadata related to the CA and GA taxonomies. These metadata relate to additional projects using the data included herein but are not directly shown in the Transcriptomics Explorer. The CA cluster, subclass, neighborhood, and QCpass calls correspond to one another (and likewise for GA). The subclass and neighborhood calls almost perfectly line up, but not quite, which is why they are

listed separately. For example, a small number of PVALB nuclei in CA might be labeled as SST in GA (etc.). All class calls for CA, GA, and SEA-AD match.
    - [GA/CA]_QCpass: an indication of whether a cell was included in the analysis ("TRUE") or was excluded from the analysis ("FALSE").
    - [GA/CA]_cluster**: the cell type cluster name (e.g., the result of the clustering algorithm in the GA or CA taxonomy).
    - [GA/CA]_subclass**: cell type subclass (for example, "SST", "L6 CT", and "Astrocyte"). Subclasses and classes (below) match standard interlex terms published in mammalian primary motor cortex (publication).
    - [GA/CA] neighborhood**: level of cell type resolution used for analysis which groups all non-neuronal and non-neural cells together, and divides both GABAergic and glutamatergic cells into two distinct groups (MGE vs. CGE, and IT vs. non-IT, respectively).
    - class**: broad cell class terms (the same column as above)
  - ** Each item of this table tagged with '**' has three columns:
    - [item]_label: name of the item (e.g., "L2/3 IT" would be an example under " SEAAD_subclass_label")
    - [item]_order: order that the item will be displayed on the Transcriptomics Explorer (for "SEAAD_supertype_order") and/or the order that the item appears in it's relevant taxonomy (see "nomenclature.zip" below)
    - [item]_color: color that the item will be displayed on the Transcriptomics Explorer and/or associated manuscripts (as of May 2022)

- **Taxonomy of clusters (dend.json)**: This file contains the serialized cluster hierarchy with all node information embedded in json format. The dendrogram shown at the top of the Transcriptomics Explorer, including the underlying cluster order, is derived from this file. This dendrogram corresponds to the SEA-AD taxonomy, CCN202204130, which correspond to columns without CA or GA prefixes in the cell metadata, and with additional taxonomy metadata listed in "nomenclature.zip"
- **Taxonomy of clusters (dend.RDS)**: The initial serialized cluster hierarchy file with identical information as "dend.json" (dend.json is derived from this file). This file format is compatible with the R programming language and can be read using "readRDS".

- **Cluster trimmed means (trimmed_means.csv):** A table of trimmed means for each gene (rows) in each cluster ("SEAAD_supertype_label"; columns). Trimmed means are calculated by first normalizing gene expression as follows: value = log2(CPM(UMI)+1), where CPM = "counts per million" and then calculating the average expression of the middle 50% of the data (e.g., after excluding the 25% highest and 25% lowest expression values) independently for each gene and each cluster.
  - The first row lists the cluster name (cluster_label), which matches the cell type alias shown in the Transcriptomic Explorer.
  - The first column lists the unique gene identifier (gene), which in most cases is the gene symbol, and which in all cases comes from "genes.gtf.gz".

- **Cluster medians.csv (medians.csv):** This file is the same as "trimmed_means.csv", except that the values shown are the median values calculated independently for each gene across each cluster.

- **UMAP and scVI latent space coordinates (UMAP.csv):** Uniform Manifold Approximation and Projection (UMAP) coordinates for each sample shown on the Transcriptomics Explorer. UMAP is a method for dimensionality reduction of gene expression that is well suited for data visualization (reference). The UMAP coordinates were calculated from a nearest neighbor graph constructed with scanpy.tl.neighbors (reference). More specifically, a 20-dimension latent representation of the entire counts matrix was learned with scVI (reference), with gene dispersion allowed to vary across each cluster label and the donor and number of genes encoded as categorical covariates. This file contains the following columns:
    - sample_name: Unique sample identifier for matching across all relevant files
    - umap_0: First UMAP coordinate
    - umap_1: Second UMAP coordinate
    - scVI_[0-19]: 20 scVI latent space dimensions

- **Seurat object (Seurat_object.RDS):** This file contains Seurat (version 4.0.4) object with the same UMI counts and cell metadata as described above. This file format is compatible with the R programming language and can be read using "readRDS".

- **AnnData object (Full_AnnData_Object.h5ad):** This file contains the same information as expression_matrix.csv and cell_metadata.csv, but saved in a standard AnnData hdf5 (version 0.7.8) format commonly used in python.

- **10x Projected AnnData object (10x_projected_AnnData_Object.h5ad):** This file contains counts and metadata (as above) for the 10x cells that passed QC, but also coordinates for the scVI latent space and UMAP embedding, nearest neighbor graph, and cluster specific colors.

- **scVI Model (scVI_model.zip):** This archive contains the trained scVI model used to project the counts matrix into the latent dimensions (scVI.csv). It contains a folder with three files (see below), which is loaded directly by scVI. This model can, in principle, be used to project additional datasets into the same latent space either using training weights as is or by initializing a new model with its weights.
    - var_names.csv: The genes used in constructing the model (all genes were used).
    - attr.pkl: Model attributes in python pickle format
    - model_params.pt: Model parameters in pytorch format

- **Nomenclature files (nomenclature.zip):** Output files from applying Common Cell type Nomenclature (CCN) to this taxonomy, for linkage to other BICCN taxonomies. More detailed information about how and why to apply the CCN, and its outputs can be found at the Allen Brain Map (here; reference; GitHub). This zip file contains these files:
    - **dend.json**: A duplicate of the dendrogram file, as defined above

    - **nomenclature_table.csv**: This file includes metadata columns from the CCN format (terms defined here) for each node in dend.json, as well as additional cell sets corresponding to subclasses and classes without corresponding nodes.
    - **cell_to_cell_set_assignments.csv**: A table indicating the probability of each cell mapping to each cell set. In this case we define hard probabilities (0 = unassigned to cell set; 1 = assigned to cell set) but this could be adapted to reflect real probabilities calculated elsewhere.

- **File generation scripts (scripts.zip):** This file contains R and python scripts for generating some of the files listed above based on gene expression matrix and cell metadata. More specifically:
    - **c1_initial_dendrogram_and_summarizations.R**: This script generates "trimmed_means.csv", "medians.csv", and an initial version of the dendrogram.
    - **c2_build_annotation_tables.[Rmd/nb.html]**: These scripts generate everything in “nomenclature.zip” and also source “required_scripts.R”.
    - **Build AnnData.[ipynb/html]**: This Jupyter script generates the AnnData object above, along with the scVI latent space and associated UMAP

- **Gene information (genes.gtf.gz):** This file is used for aligning the RNA-seq data to the reference human genome and associated transcriptome mentioned in the introductory paragraph (link). Information about genes and associated transcripts is located within this file, and only rows tagged as “gene” are used in the alignment. Each row contains information including (but not limited to):
    - (header values): chromosome location of gene (or transcript or component)
    - gene_id: Ensembl gene ID
    - gene_type: whether it’s protein coding, one of a few noncoding transcript types, or a pseudogene
    - gene_name: gene symbol (this is what is used in the other files)