CNSDV Task Documentation

Name: Vicky Daiya (vdaiya@iu.edu)

I decided to do the "**Cell Type Distribution Analysis**" task. Here, I have described my approach, assumptions, results and steps to reproduce the results.

I created a separate folder for working on the task.

1. **Getting the data**

   I visited the https://portal.nemoarchive.org/ to get the 7 samples (9 files) by applying the filters as mentioned in the task.

2. **Client to download the files**

   The task mentioned using a client to download the files that were obtained in **Step 1.** For this, I decided to use **portal_client,** which is a Python-based client for downloading data files hosted by an instance of the portal software. In order to do this, I first downloaded the github files for **portal_client** from https://github.com/IGS/portal_client. I then had to install the python module for **portal_client.** For this, I created a python virtual environment and then using "pip3 install ." command installed the module. **portal_client** has a command to download the files using a manifest file. For this, I went back to the URL in **Step 1,** added the files to the cart and then downloaded the manifest file. Using, "portal_client --manifest /path/to/my/manifest.tsv" command, I was able to download the 9 zipped files.

3. **Preparing data to upload on Azimuth**

   After downloading the data, the next step was to upload the data on Azimuth for analysis. Azimuth does not accept the zipped mex files as input. Hence, I wrote a script **prepare_data.R.** This script unzips all the tar files using untar and stores them in **data** folder in the working directory. I got 9 folders for the 9 zipped files. I decided to convert data to Seurat objects as RDS files for Azimuth. Hence, I first read the files in this folders using Read10X which helps reading mex data. Then, I converted them to Seurat object using CreateSeuratObject. Finally, I saved these objects as **.rds** files in the **RDS_files** directory.
   Note: Read10X gave errors for 4/9 folders. Read10X expects a barcodes.tsv, genes.tsv and matrix.mtx files. 4 folders namely filtered_feature_bc_matrix, GW18_motor, GW19_M1_all and GW19_M1_CP didn't follow the file naming convention. Hence, I renamed the files in these folders from **prepare_data.R**.

   Library used: Seurat (Read10X() and CreateSeuratObject())

4. **Uploading data on Azimuth**

   I uploaded the 9 RDS files on Azimuth. After following the steps given in task, I downloaded the 9 TSV files for each dataset. These are stored in **Azimuth TSVs** folder.

5. **Aggregating subtotals**

   For this step, I wrote the **analysis.R** script. This script reads the TSV files, converts them to a **data.table** (for speeding up group by and aggregation operations) and then finds the count and percentage for each **predicted.cluster**. The results for all the 9 TSV files are stored in a single file **Azimuth_analysis.xlsx** with different sheets.

   Library used: data.table, openxlsx

6. **Reproducing Azimuth results locally (Bonus)**

   Azimuth also provided script which was used for producing the results along with a TSV which has the results. I downloaded the script for one of the datasets (the script will be same for all datasets). I created the **azimuth_local.R** script. This script has a function **azimuth_analysis()** which is basically just the script provided by Azimuth. Azimuth's script gives a Seurat object. I got the meta data from the seurat object and using **dplyr**, I created the data frame which has the same format as the TSV file which they offer. When you run the **azimuth_local.R** script, All the RDS files in **RDS_files** folder are analyzed and a TSV file for each RDS file is stored in **Azimuth local TSVs** folder. Thus, the need for uploading data on Azimuth is eliminated.
   The TSVs produced locally and using Azimuth app are identical.
   After getting the local **Azimuth local TSVs**, I performed **Step 5** again to get the xlsx file. This is stored as **Azimuth_local_analysis.xlsx**. Again, this is identical to **Azimuth_analysis.xlsx**.

   Library used: devtools (installing azimuth package from git), BiocManager (downloading Seurat package), Seurat, Azimuth, dplyr, readr

Github link: https://github.com/vickydaiya/CNSDV_assessment

This repository has all the files. **Azimuth_analysis.xlsx** and **Azimuth_local_analysis.xlsx** are the final results obtained. **Azimuth TSVs** and **Azimuth local TSVs** has the TSV files that are obtained after analyzing data on Azimuth.

Steps to reproduce:

1. Get the manifest file from https://portal.nemoarchive.org/
2. Clone the repository
3. Clone https://github.com/IGS/portal_client
4. Create a python virtual environment (optional)
5. Install portal client and run the command (portal_client --manifest /path/to/my/manifest.tsv)

6. Run the **prepare_data.R** file. This will generate the RDS files in **RDS_files** folder
7. Run the **azimuth_local.R** file. This will generate the TSVs after azimuth's analysis in **Azimuth local TSVs** folder.
8. Run the **analysis.R** file, This will generate the final results file in **Azimuth_local_analysis.xlsx**