Abigail Pajulas
Pradeep Varathan
B573 - Fall 2019

Developing an R package to study the immune system in an Alzheimer's disease dataset

Abstract

Alzheimer's Disease is a chronic neurodegenerative disorder that affects 24 million people worldwide. Lymphocytes, such as T regulatory and T helper 17 cells, have been implicated in disease pathogenesis and progression. Previous research has demonstrated that the Treg population is significantly diminished in AD patients compared to healthy subjects. Furthermore, there is an imbalance of circulating Tregs and Th17 cells in AD patients. By designing an R program, we will be able to characterize differential gene expression of lymphocyte-related genes from RNA-seq and microarray data. To elucidate the role of these lymphocytes in Alzheimer's disease, we will use published and publicly available microarray data on human brains of 27 control and 52 AD subjects. Our analysis will mainly focus on the differential immune-related gene expression within the regions of the brain. Specifically, our R package can be used to identify innate and adaptive immune cell types. Transcriptomic activity within the three groups would support that lymphocytes such as Tregs or Th17 cells, play important regulatory roles in AD pathogenesis and progression.

Introduction

In our project, we are interested in designing an R package to characterize gene expression to profile innate and adaptive immune cells in microarray and RNA-seq datasets. The specific dataset used in our analysis was recovered from a GEO dataset on human Alzheimer disease individuals' brain that is publically-available.

Transcriptome analysis of Alzheimer Brains in humans

GEO dataset from Patel et al., 2018.

Patel et al. made the microarray geoset publically available in 2018. Microarray data profiling humans brains from 27 healthy controls, 33 asymptomatic but highly likely to develop AD individuals, and 52 AD subjects were used to assess differential and co-expression of genes in various regions of the brain. The researchers obtained total RNA from entorhinal cortex, temporal cortex, frontal cortex, and cerebellum brain regions. The dataset has samples from 46 males and 64 females with an age range of 43 - 105.

In this project, we will focus on a thorough analysis of innate and adaptive immune system contributions to AD pathology by looking at changes between asymptomatic individuals developing AD.

Review of the Literature

Alzheimer's Disease (AD) is a chronic neurodegenerative disorder that affects 24 million people worldwide. Its symptoms are characterized as behavioral and cognitive disabilities such as loss of memory, language problems, and mood swings (McGeer et al., 1989). Neuropathologically, AD is characterized by amyloid plaque forming around neurons containing tangles in the brain which impairs the neural communication and connectivity (Mishra et al., 2017). Immune cells, such as microglial cells, play important roles in clearing cellular debris from damaged areas and promoting an immune response (Mishra et al., 2017). In AD, immune cells such as lymphocytes are suggested to play regulatory roles in disease pathogenesis and progression (Oberstein et al., 2018). In this short report, we will characterize the role of innate and adaptive immune cells in Alzheimer's Disease and propose an R program that can be used to screen RNA-seq and microarray data to characterize neuroinflammatory environment.

Cells of the innate immune system are important for hunting and clearing pathogens. Innate immune cells such as phagocytes have been demonstrated to contribute to Alzheimer's disease (Mishra et al., 2017). Neuroinflammation has been linked with microglia and astrocytes are demonstrated to activate and produce an inflammatory response to aid in clearing the damaged area in people with AD (Mishra et al., 2017). Phagocytes like macrophages and microglial cells engulf and destroy pathogens. Whereas, granulocytes recognize and send signals to recruit phagocytic cells. Localized inflammation in the brain recruits and activates the complement system to aid in microglial cells in clearing cellular debris. Cytokines secreted by microglial cells and other cells in the adaptive immune response influences complement system activation and function.

Cells of the adaptive immune system which include T cells and B cells are highly specific and diverse. These cells can downregulate total and resting regulatory T cells (Tregs) numbers and functions. Downregulation of the adaptive immune system has been shown to occur in people with Alzheimer's Disease and Multiple Sclerosis (Ciccocioppo et al., 2019). Tregs are T-lymphocytes involved in maintaining immune peripheral tolerance by removing self-reactive B and T cells. It does this by regulating apoptosis of lymphocytes that respond against the body's own tissues are other harmless material such as food. Ciccocioppo et al. (2019) demonstrated that total and resting Treg numbers are significantly decreased in AD patients compared to that of healthy controls. This suggests that impairment in Treg tolerance may play a role in AD pathology. There has been a correlation between Tregs and proinflammatory T helper cells. With a decrease in Treg number, there is an increase in proinflammatory-secreting T helper subsets. Oberstein et al. (2018) showed that in early stages of AD, there is a decrease in Tregs and an increase in Th17 cells. They hypothesize that this switch in addition to other factors lead to hyperphosphorylated Tau proteins.

Although previous studies have identified the role of the innate and adaptive immune system to regulate AD-like symptoms, these studies have not identified potential secretory mediators that might contribute to AD pathogenesis and progression. In order to elucidate the role of the immune system in AD, we will characterize the differential gene expression of lymphocyte-related genes in human brain datasets. Specifically, our R package will screen significant changes in gene expression between two RNA-seq or microarray datasets for cell types belonging to CD4, CD8, T helper subsets, monocytes, and natural killer cell types. By identifying these subsets, we hope to distinguish between innate and adaptive immune system contributions to AD pathology.

Research Question

To address and highlight key innate and adaptive immune cell contributions to disease progression, we have developed an R package to analyze microarray and RNA-seq datasets for significant cell-specific gene expression.

1. Which cell types are present in AD symptomatic patients compared to control?
2. Are these cell types innate or adaptive immune cells?

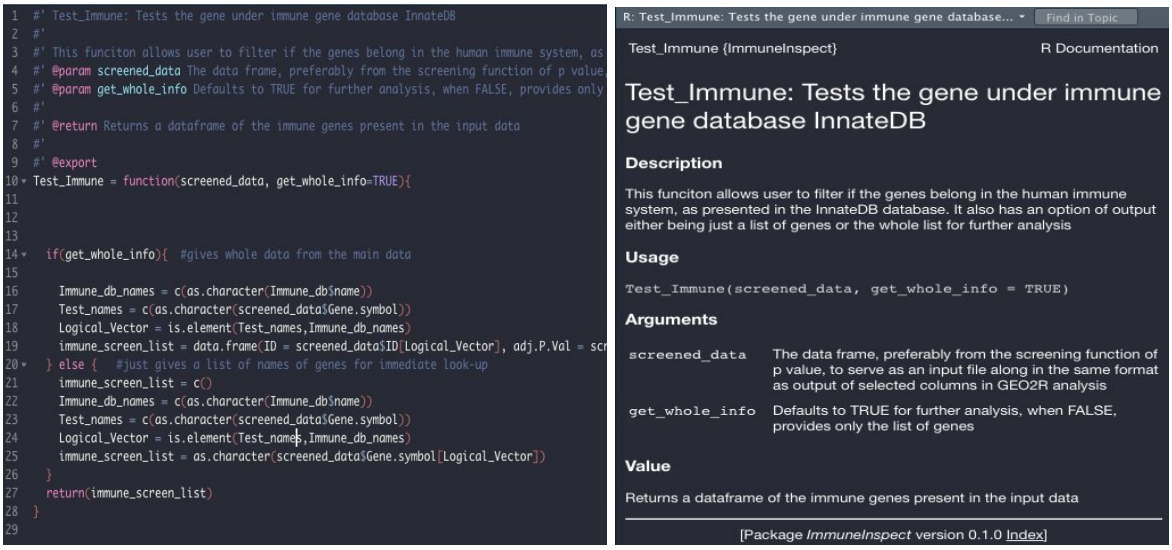
Methods

Our R package is called ImmuneInspect. In order to run our package on various GEO datasets, you must first install our ImmuneInspect package and load it on to the library. Using GEO2R analyzed csv datasets, we can save the data in to dataframes in R studio environment. Using the R studio script in Figure 1 will allow for the analysis of any GEO dataset using our ImmuneInspect package.

This package can be used to analyze microarray and RNA seq GEO datasets after analysis from GEO2R for immune cell contributions based on InnateDB, a database with an immune system gene list (Figure 1). After screening for relevant immune system genes, we used dice.org to retrieve csv files of cell type-specific gene lists (Figure 2). After screening and filtering our data, we have created visual graphs and plots to more easily display our analysis results (Figure 4-6).

```
library("ImmuneInspect")
library(plyr)
library(dplyr)
library(d3heatmap)

setwd("~/Documents/ProgramIntro_Assignments/Project/")

file1 = read.table("AD_vs_AsymAD.txt",header = T)
file2 = read.table("AD_vs_Control.txt",header = T)
file3 = read.table("AsymAD_vs_Control.txt",header = T)

#Screen
screen1 = Screen_pval(file1)
screen2 = Screen_pval(file2)
screen3 = Screen_pval(file3)

#Test Immune
immune1 = Test_Immune(screen1)
immune2 = Test_Immune(screen2)
immune3 = Test_Immune(screen3)

#Retrive info
ret1 = Retrive_from_db(immune1)
ret2 = Retrive_from_db(immune2)
ret3 = Retrive_from_db(immune3)

#Test Adaptive
adt1 = Test_Adaptive(immune1)
adt2 = Test_Adaptive(immune2)
adt3 = Test_Adaptive(immune3)

#Make Plots
Make_plots(ret1)
Make_plots(ret2)
Make_plots(ret3)

#Creting Heatmap
Create_HeatMap(immune1,immune2,names_exp=c("AD_Asym","AD_Cont"),cut_num = 0.8)
Create_HeatMap(immune1,immune3,names_exp = c("AD_Asym","Asym_Cont"),cut_num = 0.8)
Create_HeatMap(immune2,immune3,names_exp = c("AD_Cont","Asym_Cont"),cut_num = 0.8)
```

Figure 1. R Studio script to run different GEO datasets using our ImmuneInspect package.



```
#' Test_Immune: Tests the gene under immune gene database InnateDB
#'
#' This funciton allows user to filter if the genes belong in the human immune system, as
#' @param screened_data The data frame, preferably from the screening function of p value,
#' @param get_whole_info Defaults to TRUE for further analysis, when FALSE, provides only
#'
#' @return Returns a dataframe of the immune genes present in the input data
#'
#' @export
Test_Immune = function(screened_data, get_whole_info=TRUE){



  if(get_whole_info){  #gives whole data from the main data

    Immune_db_names = c(as.character(Immune_db$name))
    Test_names = c(as.character(screened_data$Gene.symbol))
    Logical_Vector = is.element(Test_names,Immune_db_names)
    immune_screen_list = data.frame(ID = screened_data$ID[Logical_Vector], adj.P.Val = scr
  } else {   #just gives a list of names of genes for immediate look-up
    immune_screen_list = c()
    Immune_db_names = c(as.character(Immune_db$name))
    Test_names = c(as.character(screened_data$Gene.symbol))
    Logical_Vector = is.element(Test_names,Immune_db_names)
    immune_screen_list = as.character(screened_data$Gene.symbol[Logical_Vector])
  }
  return(immune_screen_list)
}
```

R: Test_Immune: Tests the gene under immune gene database... | Find in Topic

Test_Immune {ImmuneInspect} R Documentation

Test_Immune: Tests the gene under immune gene database InnateDB

Description

This funciton allows user to filter if the genes belong in the human immune system, as presented in the InnateDB database. It also has an option of output either being just a list of genes or the whole list for further analysis

Usage

```
Test_Immune(screened_data, get_whole_info = TRUE)
```

Arguments

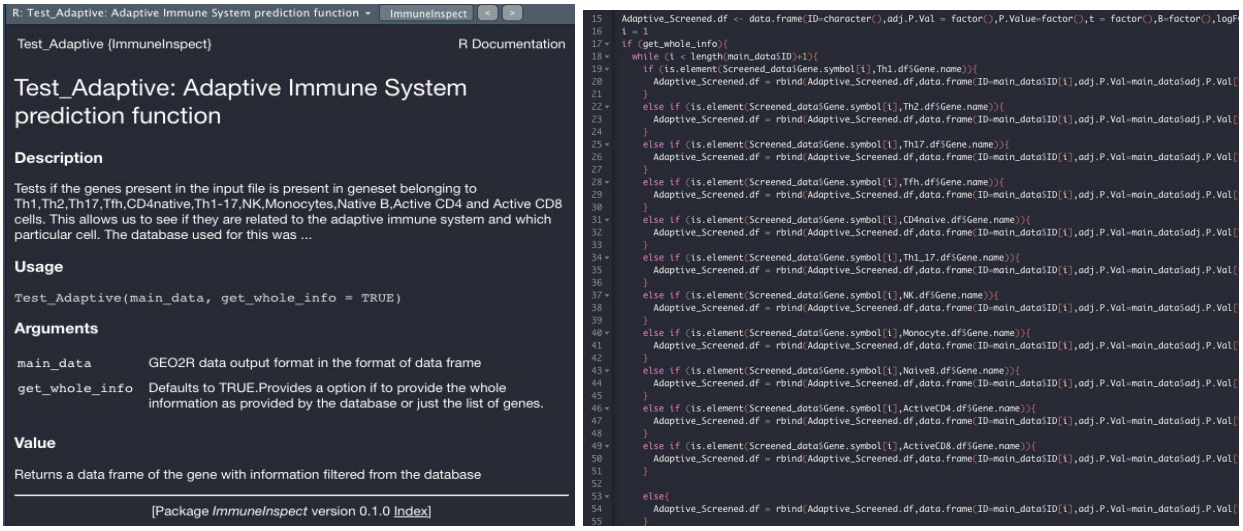
screened_data	The data frame, preferably from the screening function of p value, to serve as an input file along in the same format as output of selected columns in GEO2R analysis
get_whole_info	Defaults to TRUE for further analysis, when FALSE, provides only the list of genes

Value

Returns a dataframe of the immune genes present in the input data

[Package *ImmuneInspect* version 0.1.0 Index]

Figure 2. Test_Immune function is used to compare our analyzed GEO dataset with an Innate database to get the genes associated with the immune system.



R: Test_Adaptive: Adaptive Immune System prediction function | ImmuneInspect

Test_Adaptive {ImmuneInspect} R Documentation

Test_Adaptive: Adaptive Immune System prediction function

Description

Tests if the genes present in the input file is present in geneset belonging to Th1,Th2,Th17,Tfh,CD4native,Th1-17,NK,Monocytes,Native B,Active CD4 and Active CD8 cells. This allows us to see if they are related to the adaptive immune system and which particular cell. The database used for this was ...

Usage

```
Test_Adaptive(main_data, get_whole_info = TRUE)
```

Arguments

main_data	GEO2R data output format in the format of data frame
get_whole_info	Defaults to TRUE.Provides a option if to provide the whole information as provided by the database or just the list of genes.

Value

Returns a data frame of the gene with information filtered from the database

[Package *ImmuneInspect* version 0.1.0 Index]

```
Adaptive_Screened.df <- data.frame(ID=character(),adj.P.Val = factor(),P.Value=factor(),t = factor(),B=factor(),logF
i = 1
if (get_whole_info){
  while (i < length(main_data$ID)+1){
    if (is.element(Screened_data$Gene.symbol[i],Th1.df$Gene.name)){
      Adaptive_Screened.df = rbind(Adaptive_Screened.df,data.frame(ID=main_data$ID[i],adj.P.Val=main_data$adj.P.Val[
    }
    else if (is.element(Screened_data$Gene.symbol[i],Th2.df$Gene.name)){
      Adaptive_Screened.df = rbind(Adaptive_Screened.df,data.frame(ID=main_data$ID[i],adj.P.Val=main_data$adj.P.Val[
    }
    else if (is.element(Screened_data$Gene.symbol[i],Th17.df$Gene.name)){
      Adaptive_Screened.df = rbind(Adaptive_Screened.df,data.frame(ID=main_data$ID[i],adj.P.Val=main_data$adj.P.Val[
    }
    else if (is.element(Screened_data$Gene.symbol[i],Tfh.df$Gene.name)){
      Adaptive_Screened.df = rbind(Adaptive_Screened.df,data.frame(ID=main_data$ID[i],adj.P.Val=main_data$adj.P.Val[
    }
    else if (is.element(Screened_data$Gene.symbol[i],CD4naive.df$Gene.name)){
      Adaptive_Screened.df = rbind(Adaptive_Screened.df,data.frame(ID=main_data$ID[i],adj.P.Val=main_data$adj.P.Val[
    }
    else if (is.element(Screened_data$Gene.symbol[i],Th1_17.df$Gene.name)){
      Adaptive_Screened.df = rbind(Adaptive_Screened.df,data.frame(ID=main_data$ID[i],adj.P.Val=main_data$adj.P.Val[
    }
    else if (is.element(Screened_data$Gene.symbol[i],NK.df$Gene.name)){
      Adaptive_Screened.df = rbind(Adaptive_Screened.df,data.frame(ID=main_data$ID[i],adj.P.Val=main_data$adj.P.Val[
    }
    else if (is.element(Screened_data$Gene.symbol[i],Monocyte.df$Gene.name)){
      Adaptive_Screened.df = rbind(Adaptive_Screened.df,data.frame(ID=main_data$ID[i],adj.P.Val=main_data$adj.P.Val[
    }
    else if (is.element(Screened_data$Gene.symbol[i],NaiveB.df$Gene.name)){
      Adaptive_Screened.df = rbind(Adaptive_Screened.df,data.frame(ID=main_data$ID[i],adj.P.Val=main_data$adj.P.Val[
    }
    else if (is.element(Screened_data$Gene.symbol[i],ActiveCD4.df$Gene.name)){
      Adaptive_Screened.df = rbind(Adaptive_Screened.df,data.frame(ID=main_data$ID[i],adj.P.Val=main_data$adj.P.Val[
    }
    else if (is.element(Screened_data$Gene.symbol[i],ActiveCD8.df$Gene.name)){
      Adaptive_Screened.df = rbind(Adaptive_Screened.df,data.frame(ID=main_data$ID[i],adj.P.Val=main_data$adj.P.Val[
    }

    else{
      Adaptive_Screened.df = rbind(Adaptive_Screened.df,data.frame(ID=main_data$ID[i],adj.P.Val=main_data$adj.P.Val[
    }
```

Figure 3. Test_Adaptive function is used to compare our analyzed GEO dataset with an a series of cell-type specific data frames that contain gene names. The function uses each gene from our analyzed GEO dataset to compare it to our cell-type specific data frames, so that we can clearly group our analyzed genes into a specific cell type.

Data Analysis Results

The package provides information on which highly differentially expressed genes are present in the immune system, their localizations, type of cells and also provides a comparative analysis using a heatmap. Using GEO2R on GSE118553 dataset, we were able

to analyse the data on the Alzheimer's microarray data on how the influence of immune system genes play a role in the pathology of the disease.

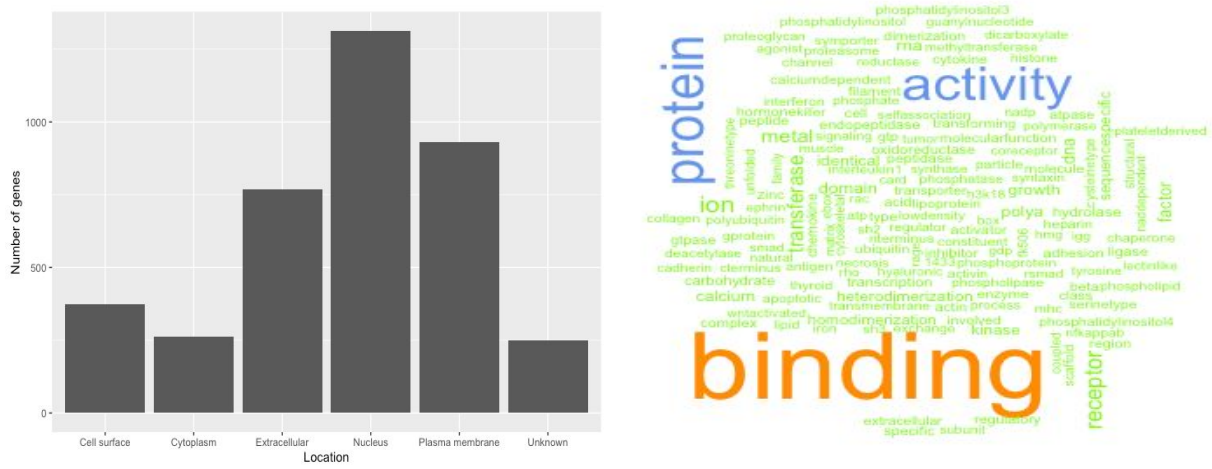


Figure 4. The results from the Make_Plots function in the package provides information about the localization of the genes and the important and the most repeated GO function of the top differentially expressed genes.

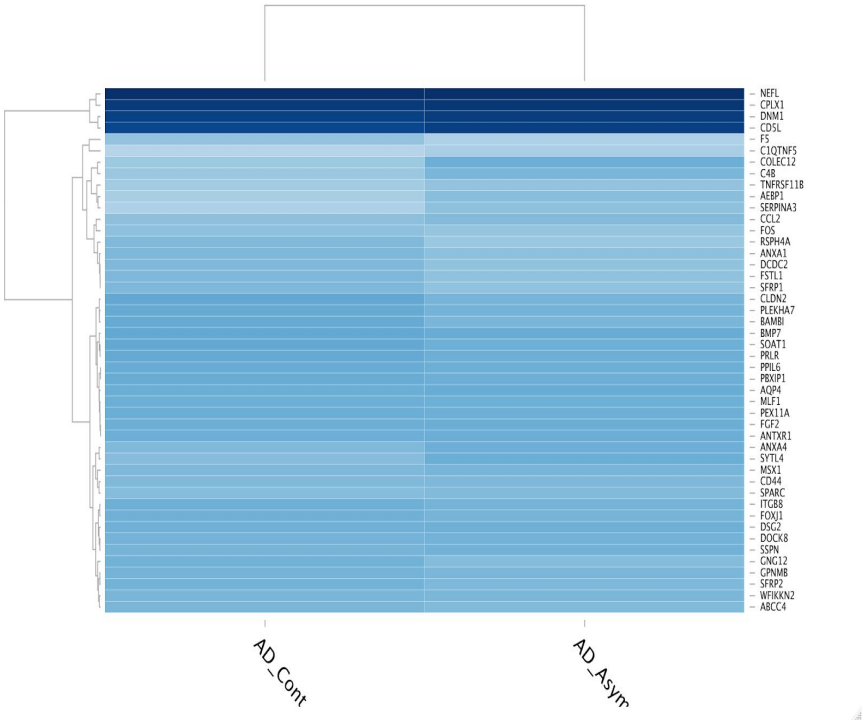


Figure 5. The heat map provides us information on how the most significantly, differentially expressed genes using fold change between healthy controls and AD

symptomatic individuals. This heatmap can clearly show the differential gene expression between the two different conditions: controls and asymptomatic individuals.

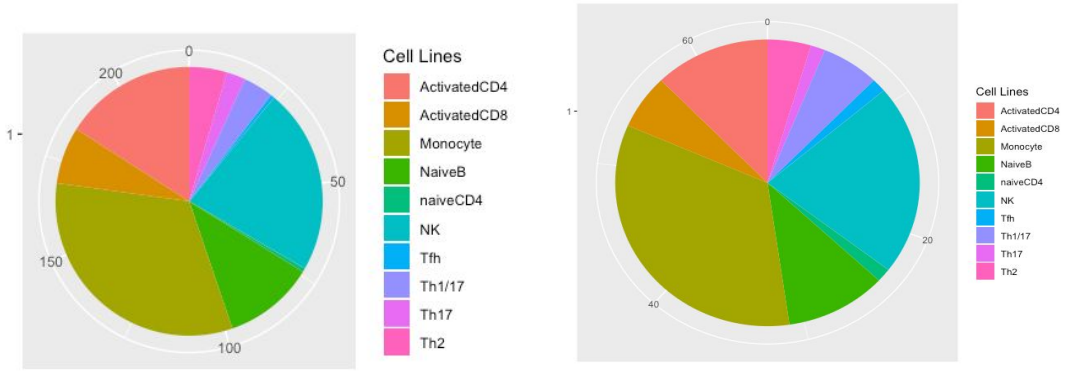


Figure 6. The pie chart represents the division of the immune genes present in the various types of immune cells thereby providing us information on which immune system (immune & adaptive) the genes belong to.

Conclusion and Limitations

Altogether, our data supports the tau protein and amyloid plaque hypothesis. In our pie charts, we found that there was an overall abundance in the frequency of monocytes and natural killer cells (Figure 6). The tau protein and amyloid plaque hypothesis has demonstrated monocytes such as microglial cell and natural killer cell infiltration in individuals with AD. Moreover, we find that there's really no difference between AD symptomatic and asymptomatic monocytes and natural killer cell infiltration. This data suggests that the higher frequency of monocytes and natural killer cells present in brain's indicates the likelihood of developing and having AD.

Our analyses are not without limitations. First and foremost, our dataset analysis is limited to what is publicly accessible in the GEO dataset webpage. Moreover, we had limited access to cell-type specific gene expression data. Access to data on immune cell type specific genes may provide more useful distinctions between our gene expression sorting. Although our data does not directly access gene expression and cell type frequencies in different regions of an AD brain compared to that of healthy control. Researchers interested in uncovering the contribution and presence of specific cell types in various regions can use our package to analyse the GEO dataset by Patel et al. (2018).

To access and download our ImmuneInspect package, please visit:
https://github.com/PradoVarathan/ImmuneInspect

References

Ciccocioppo et al., (2019) The Characterization of Regulatory T-cell profiles in Alzheimer's Disease and Multiple Sclerosis. *Nature*: *Scientific Reports*

Oberstein et al., (2018) Imbalance of Circulating Th17 and Regulatory T cells in Alzheimer's Disease: A case control study. *Frontiers in Immunology*

Cao, W & Zheng, H. (2018) Peripheral immune system in aging and Alzheimer's disease. *Molecular Neurodegeneration*

Mishra et al., (2017) Alzheimer's Disease and Immune Response: A brief overview. *J Alzheimers Dis Parkinsonism*

McGeer et al., (1989) Immune system response in Alzheimer's disease. *Can J Neurol Sci.*

Sommer et al., (2017) The Trojan horse - neuroinflammatory impact of T cells in neurodegenerative diseases. *Molecular Neurodegeneration*