For Pre-clustering we used python and followed the SCANPY protocol proposed by the authors to reproduce most of Seurat's clustering analysis -- since, at the time, only this package was able to deal very efficiently with a large number of cells. You can find an updated version of this protocol here:

https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html#

The relevant functions are

```
sc.pp.normalize_per_cell(adata, counts_per_cell_after=1e4)
sc.pp.log1p(adata)
sc.pp.highly_variable_genes(adata, min_mean=0.0125, max_mean=3, min_disp=0.5)
sc.pp.regress_out(adata, ['n_counts'])
sc.pp.scale(adata, max_value=10)
sc.tl.pca(adata, svd_solver='arpack')

sc.pp.neighbors(adata, n_neighbors=30, n_pcs=50)

sc.tl.tsne(adata)

sc.tl.louvain(adata)

sc.tl.rank_genes_groups(adata, 'louvain', method='t-test')
```

You can find the exact output we obtained from this analysis here:
https://www.dropbox.com/sh/8uxbk1qbcsbt8gz/AABggGI9onEs3lQYtNYfq5WLa?dl=0

`/AD_Ctx_clustering_outputs/scanpy_out/`

we used the top marker genes for each lovain cluster cell group (our pre-clusters) to assign cell-type annotations. You will find the genes in this file:

`/AD_Ctx_clustering_outputs/scanpy_out/uns/rank_genes_groups_gene_names.csv`

The cell tags in the file /obs.csv should match the tags in the data that you can obtain from Synapse (https://www.synapse.org/#!Syn- apse:syn18485175), where you will also find corresponding cell-type annotations.

You will see that in this pre-clustering analysis the input was 75,060 cells.

After cell-type annotation and an additional round of QC (see Note on QC below), sub-clustering was performed independently for each cell-type. Sub-clustering was performed for 73,581 of the cells -- you can find TAGs of these cells in the file.

```
/AD_Ctx_clustering_outputs/scanpy_out/qced/cells_for_subcluster_postQC.csv
```

We performed all the additional analyses using R. For sub-clustering, we defined the same parameter values as used in SCANPY, but this time run Seurat using the following function, where the input is a SingleCellExperiment object (https://bioconductor.org/packages/release/bioc/html/SingleCellExperiment.html), with the data corresponding to one cell-type:

```
Run.Seurat.Clustering <- function(Insce) {
  require(Seurat)
  require(scater)
 Insce <- calculateQCMetrics(Insce)
  temp.counts <- Insce@assays[["counts"]]
        temp.seu    <-    CreateSeuratObject(raw.data    =    temp.counts,    meta.data    =
as.data.frame(colData(Insce)))
    temp.seu  <-  NormalizeData(object  =  temp.seu,  normalization.method  =  "LogNormalize",
scale.factor = 10000)
      temp.seu   <-   FindVariableGenes(object   =   temp.seu,   mean.function   =   ExpMean,
dispersion.function = LogVMR)
  temp.seu <- ScaleData(object = temp.seu, vars.to.regress = "total_counts_endogenous")
    temp.seu  <-  RunPCA(object  =  temp.seu,  pc.genes  =  temp.seu@var.genes,  do.print  =  F,
pcs.compute = 50)
   temp.seu  <-  FindClusters(object  =  temp.seu,  reduction.type  =  "pca",  dims.use  =  1:50,
resolution = 0.6, print.output = 0, save.SNN = TRUE, k.param = 30)
  temp.seu <- RunTSNE(object = temp.seu, dims.use = 1:10, do.fast = TRUE)
  temp.seu.markers <- FindAllMarkers(temp.seu, only.pos = T)
  return(list(seu=temp.seu, markers=temp.seu.markers))
}
```

We applied this function to all cell-types and obtained the following output:

```
/AD_Ctx_clustering_outputs/subclustering_out/subclust.out.by.celltype.RData
```

After subsequent QC analyses (see note below) we identified subclusters: Mic4 and Mic5; Ast4, Ast5, Ast6; Opc3 ; Oli6, Oli7, Oli2; Ex13, Ex10, Ex15; End0 as potential low-quality/doublet cells, and remove them from further analysis. Based on markers, we identified subcluster End1 as putative Pericytes, and renamed End2 and End1 and End3 as End2.

The output of these analyses in the 70,634 cells matching the filtered/annotated cells reported in Synapse.

Cell signature genes (subcluster “markers”) were recomputed using only the final, curated cells (n=70,634) by comparing expression values of cells in subcluster X vs cells of the same cell-type not in subcluster X. For flexibility, we implemented this directly using R’s functions: wilcox.test and p.adjust with fdr correction.

You can find the output of this in
`/AD_Ctx_clustering_outputs/subclustering_out/subclust.markers.DE.list.RData`

Genes reported as signatures correspond to those over-expressed in the subcluster -- meeting the criteria: FDR-corrected P value ≤ 0.01, lFC=log2(mean gene expression across cells in sub-cluster X/mean gene expression across cells in other sub-clusters) >= 0.5, and detected in at least 25% of the cells within the given sub-cluster.

Note on QC: We do not have a specific code for doublet detection and/or cell quality control analysis, since, as commonly happens in single-cell analysis, this involved a lot of exploratory work and multiple rounds of analyses and manual curation, as briefly described in the methods section. Since we are aware there are multiple ways to do QC and there no real standards yet, we reported the complete dataset for others to analyze it in any way they see fit. In our analysis, we tried to be stringent, and to remove cells with apparent low quality, with disagreement in cell-type annotation that suggested were potential doublets, and/or those cells of only one individual clustering together.