COVID-19 Project: RSKC Analysis for Fidelity Scores Proteases and Receptors Subset 2

Rachel Kwan and Jonathan Zaslavsky

2021-07-13

Relevant Packages

```
# Load packages
library(here) # To read in data from directory
library(tidyverse) # For ggplot2, dplyr
library(magrittr) # For set_colnames() and set_rownames()
library(ggpubr) # For making publication-ready plots based on ggplot
library(RSKC) # For RSKC clustering
library(Rtsne) # To run t-SNE (dimensionality reduction)
library(factoextra) # For fviz_nbclust() and clustering analysis/visualization
library(reshape2) # For melt()
library(RColorBrewer) # For color palettes
library(viridis) # For color palettes
library(gplots) # For heatmap.2()
library(dendextend) # For manipulating visual appearance of dendrograms
library(ggdendro) # For constructing dendrograms based on ggplot
library(see) # For geom_violinhalf()
library(dabestr) # For visualizing estimation statistics

# Set the seed
set.seed(72613)
```

Import and Prepare Dataset

```
# Load full dataset
all_fidelity <- read.csv(here("Data", "ALL_Fidelity.csv"))

# List of 3 SARS-CoV-2 receptors:
# ACE2, DPP4, NRP1
# List of 3 SARS-CoV-2 proteases:
# TMPRSS2, FURIN, ADAM17
# Define a vector containing all SARS-CoV-2 related genes of interest.
genes_prorec_subset <- c("ACE2", "DPP4", "NRP1",
                         "TMPRSS2", "FURIN", "ADAM17")

# Filter full dataset to include only the genes of interest and drop the
# columns that correspond to "Entrez" and "Alias".
```

```r
fidelity_subset <- all_fidelity %>%
  filter(Gene %in% genes_prorec_subset) %>% # Includes protease/receptor genes
  select(-c(2,3))

# Tidy the data frame by separating the previous columns into
# two new columns for the brain region and cell subtype. Assign each of the
# values to a new column for the fidelity scores. The tidy data
# columns each correspond to a variable and each observation is a new row.
fidelity_subset.long <-  pivot_longer(fidelity_subset, cols = -Gene,
                                      names_to = c("Brain.Region", "Cell.Subtype"),
                                      names_sep = "_",
                                      values_to = "Fidelity")

# Pivot the data frame to a wide view by assigning new columns corresponding to
# the gene_celltype for each brain region. Assign the fidelity scores as
# the values.
fidelity_subset.wide <- pivot_wider(fidelity_subset.long,
                                    names_from = c("Gene", "Cell.Subtype"),
                                    values_from = "Fidelity",
                                    names_sep = "_")

# Define a vector containing brain regions of interest. This list comes from the
# Oldham Lab website (https://oldhamlab.ctec.ucsf.edu/).
regions <- c("FCX", "PCX", "TCX", "OCX", "STR", "DI", "HIP", "AMY", "CB")

# Choose only the desired brain regions from the data frame.
fidelity <- fidelity_subset.wide %>%
  filter(Brain.Region %in% regions) %>%
  slice(match(regions, Brain.Region)) %>% # Reorder rows to match entries in "regions" vector
  select(-contains("Percentile")) # Remove NA columns

# Use the brain regions to name the rows and remove the brain region column.
myFidelity <- fidelity %>%
  column_to_rownames("Brain.Region")
```

Perform Robust and Sparse K-Means Clustering (RSKC) and t-SNE Together

This while loop contains sections for RSKC, elbow plot, obtaining weighted data, and t-SNE.

```r
set.seed(72613)

while (T) {

  # Assign desired number of clusters to 'clust_vect'.
  clust_vect <- c(2, 4, 6, 8)

  # Assign an empty list to 'rskc_list'.
  rskc_list <- list()

  # Assign 8 colours to 'col_vect'.
  col_vect <- c("#FF0000",
                "#0000FF",
                "#00FF00",
                "#A020F0",
                "#FFA500",
                "#FFFF00",
                "#A65628",
                "#F781BF")

  # Assign a value of 0 to 'counter'.
  counter <- 0

  # Assign an empty list to 'tsne_list'.
  tsne_list <- list()

  # Assign an empty list to 'weight_list'.
  weight_list <- list()

  # For 'i' -- the current number of clusters -- in 'clust_vect'...
  for (i in clust_vect) {
    # i = 2
    # Increment 'counter' with a value of 1.
    counter = counter + 1

    ###### RSKC ######

    # Perform RSKC for whatever-the-value-of-'i'-is many clusters using
    # 'myFidelity', which has brain region as rows and gene_celltype as columns.
    # Assign RSKC's output as an entry in 'rskc_list'.
    rskc_list[[counter]] <- RSKC(myFidelity,
                                 ncl = i,
                                 alpha = 0.1,
                                 L1 = sqrt(ncol(myFidelity)))


    # Convert the row names of 'myFidelity' to a column
    # called 'Brain.Region' and store it in 'gene_and_region'.
    gene_and_region <- myFidelity %>%
      rownames_to_column("Brain.Region")
```

```r
# For the current object in 'rskc_list' convert the cluster labels
# into characters, and assign them to a new column called 'cluster_labels'
# in 'gene_and_region'.
gene_and_region$cluster_labels <- rskc_list[[counter]]$labels %>%
  as.character()


# Order the weights for the current item in 'rskc_list' from largest
# to smallest, extract the names of the genes in this order,
# convert this object into a data frame, and store this info in
# an object 'weight_df' in a column called 'gene'.
weight_df <- sort(rskc_list[[counter]]$weights,
                  decreasing = T) %>%
  names() %>%
  as.data.frame() %>%
  rename('gene' = ".")


# Assign the ordered weights for the current item in 'rskc_list' into
# 'weight_df', in a column called 'weight',
weight_df$weight <- sort(rskc_list[[counter]]$weights,
                         decreasing = T) %>%
  unname()


# Impose a factor order on the contents of 'weight_df$gene' in
# the current order.
weight_df$gene <- factor(weight_df$gene,
                         weight_df$gene)


# Create a bar graph of the RSKC weights for each gene ordered from
# largest to smallest. Assign this graph to an object, 'weight_bars'.
weight_bars <- weight_df %>%
  ggplot(aes(x = gene, y = weight)) +
  theme_classic() +
  geom_bar(stat = 'identity') +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.6)) +
  scale_y_continuous(expand = c(0, 0)) +
  ggtitle(paste0("RSKC Weights for K = ",i)) +
  xlab("") +
  ylab("Weights\n")


# Assign 'weight_bars' as the current entry into 'weight_list'.
weight_list[[counter]] <- weight_bars


###### Elbow Plot ######

# For the purposes of an elbow plot, we want to run RSKC more times than
# specified in 'clust_vect' above, so create new vector 'elbow_clust_vect'
elbow_clust_vect <- c(2, 3, 4, 5, 6, 7, 8, 9)

# Assign another empty list to 'elbow_rskc_list'.
elbow_rskc_list <- c()


# Only produce elbow plot when we are on the last run of the while loop,
# since we only need one elbow plot
```

```r
    if (i == clust_vect[length(clust_vect)]) {

      elbow_counter = 0

      for (n in elbow_clust_vect) {
        # Increment 'counter' with a value of 1.
        elbow_counter = elbow_counter + 1

        # Perform RSKC for whatever-the-value-of-'n'-is many clusters using
        # 'myFidelity', which has brain region as rows and gene_celltype as columns.
        # Assign RSKC's output as an entry in 'elbow_rskc_list'.
        elbow_rskc_list[[elbow_counter]] <- RSKC(myFidelity,
                                                 ncl = n,
                                                 alpha = 0.1,
                                                 L1 = sqrt(ncol(myFidelity)))
      }

      # Create empty vector 'between_ss' to store weighted between sum of squares (WBSS)
      # values from RSKC output
      between_ss <- matrix(ncol = 1, nrow = length(elbow_clust_vect))

      # Use for loop to add WBSS values to 'between_ss'
      # Some RSKC outputs may give more than one WBSS value, so take the last one
      for (n in 1:length(elbow_clust_vect)) {
        between_ss[n] <- elbow_rskc_list[[n]]$WBSS[length(elbow_rskc_list[[n]]$WBSS)]
      }

      # Create a new dataframe 'objective_function' with the WBSS values and
      # their corresponding number of clusters.
      objective_function <- data.frame(elbow_clust_vect, between_ss) %>%
        rename(k = elbow_clust_vect, WBSS = between_ss)

      # Create the elbow plot and assign it to 'elbow_plot'
      elbow_plot <- ggplot(objective_function, aes(x = k, y = WBSS)) +
        geom_line() +
        geom_point() +
        scale_x_continuous(breaks = elbow_clust_vect) +
        labs(x = "Number of Clusters",
             y = "Total Weighted Between Sum of Squares")

    }


###### Apply weights from RSKC to myFidelity ######

# Create vector of the weights obtained from RSKC and assign them to 'weights'.
# Make empty matrix 'weighted_fidelity' for new weighted fidelity scores.
weights <- as.matrix(rskc_list[[1]]$weights)

# Multiply 'myFidelity' columns containing gene_celltype by corresponding
# weights obtained from RSKC, and assign to 'weighted_fidelity'
weighted_fidelity <- sweep(t(myFidelity), MARGIN = 1, weights, `*`) %>%
  t()
```

```r
    ###### t-SNE (on weighted data) ######

    # Run t-sne on weighted fidelity scores, and assign to 'tsne'
    set.seed(72613)
    tsne <- Rtsne(weighted_fidelity, perplexity = 2)

    # Create new data frame 'tsne_out' which contains the two dimensions
    # obtained from t-SNE and corresponding regions
    tsne_out <- tsne$Y %>%
      data.frame(regions) %>%
      rename(Brain.Region = regions, V1 = X1, V2 = X2) # Rename columns

    # Merge 'tsne_out' with 'gene_and_region' according
    # to their shared 'Brain.Region' column, and assign to
    # 'tsne_genes_regions_clusts'.
    tsne_genes_regions_clusts <- merge(tsne_out,
                                       gene_and_region,
                                       by = "Brain.Region")

    # Create a tSNE scatter plot where each point is colour-coded according to
    # its designated RSKC cluster and assign this figure to 'tsne_scatter'.
    tsne_scatter <- ggplot(tsne_genes_regions_clusts,
                           aes(V1,
                               V2,
                               fill = cluster_labels)) +
      geom_point(shape = 21, size = 3) +
      scale_fill_manual(values = col_vect[1:i],
                        labels = 1:i,
                        name = "Cluster") +
      theme(axis.text.x = element_blank(),
            axis.ticks.x = element_blank(),
            axis.text.y = element_blank(),
            axis.ticks.y = element_blank(),
            panel.background = element_rect(fill = NA,
                                            colour = "white"),
            panel.border = element_blank(),
            axis.line = element_line(),
            legend.position = "bottom",
            legend.background = element_rect(fill = NA,
                                             colour = NA),
            legend.title.align = 0.5) +
      guides(fill = guide_legend(nrow = 2,
                                 ncol = 4,
                                 byrow = TRUE)) +
      labs(x = "V1", y = "V2")

    # Assign 'tsne_scatter' as an entry in 'tsne_list'.
    tsne_list[[counter]] <- tsne_scatter

}

# Break out of the while-loop when for-loop is done.
break
```
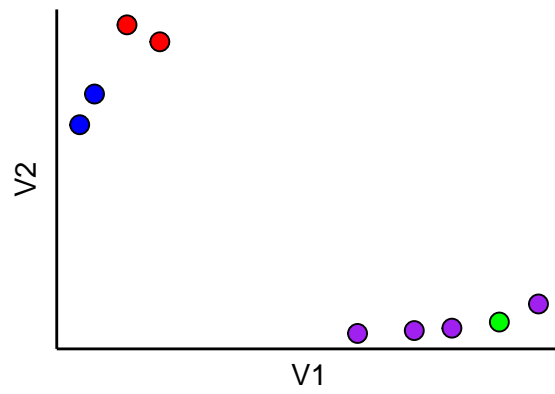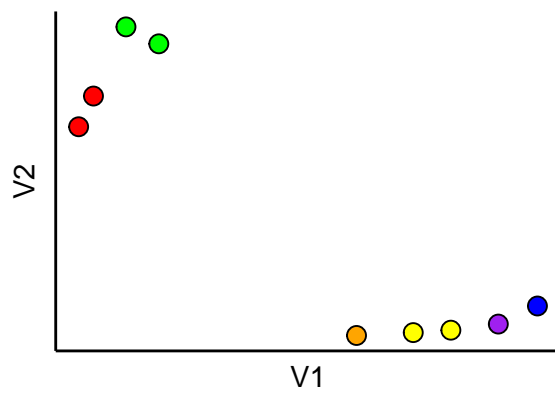
```
}
```

RSKC Weights for K = 2

RSKC Weights for K = 4

RSKC Weights for K = 6
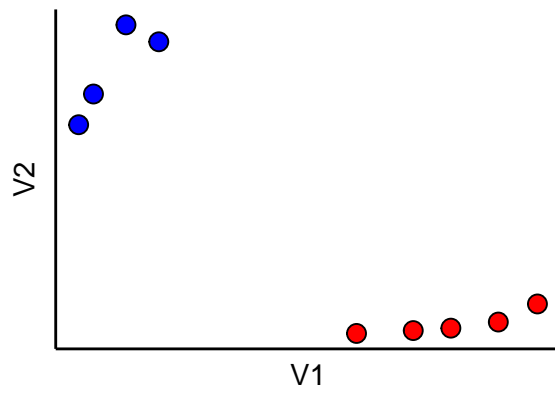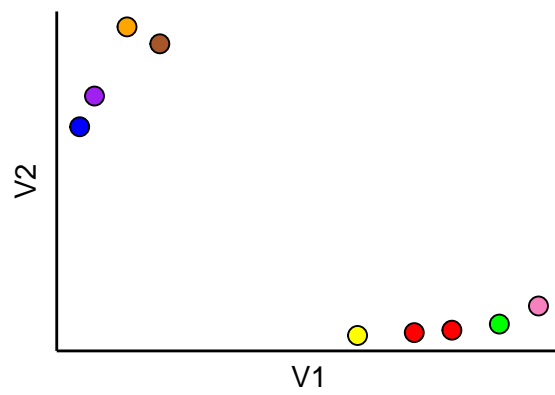
RSKC Weights for K = 8
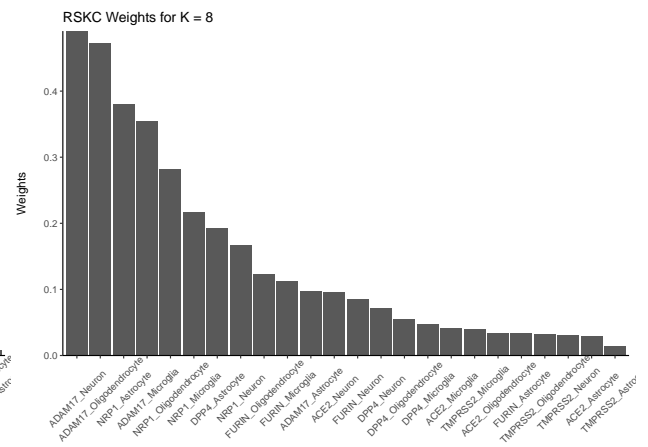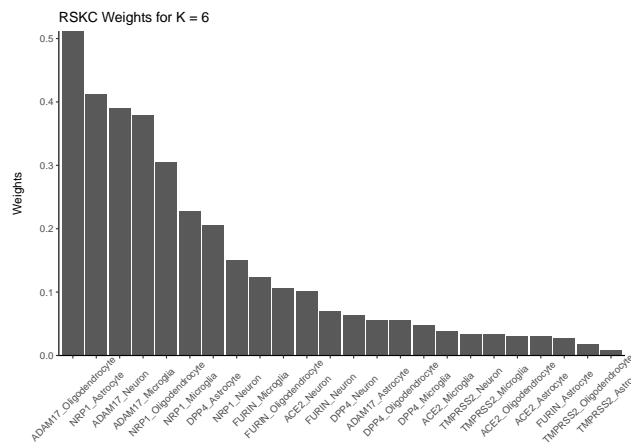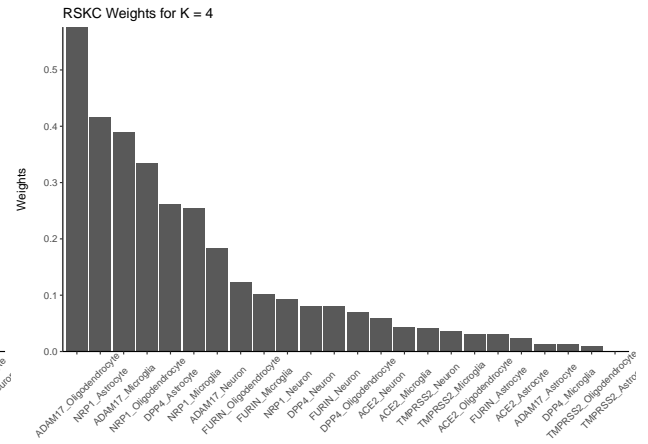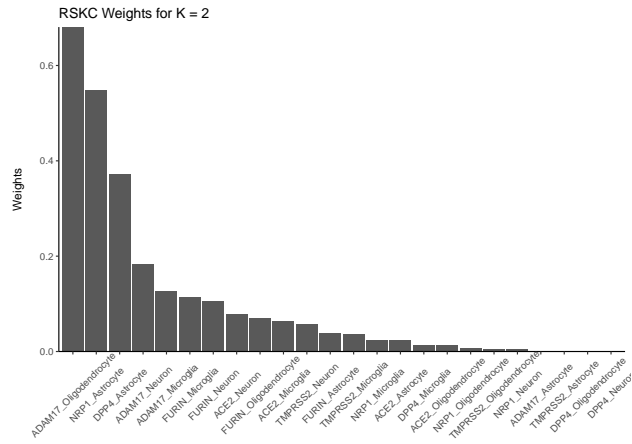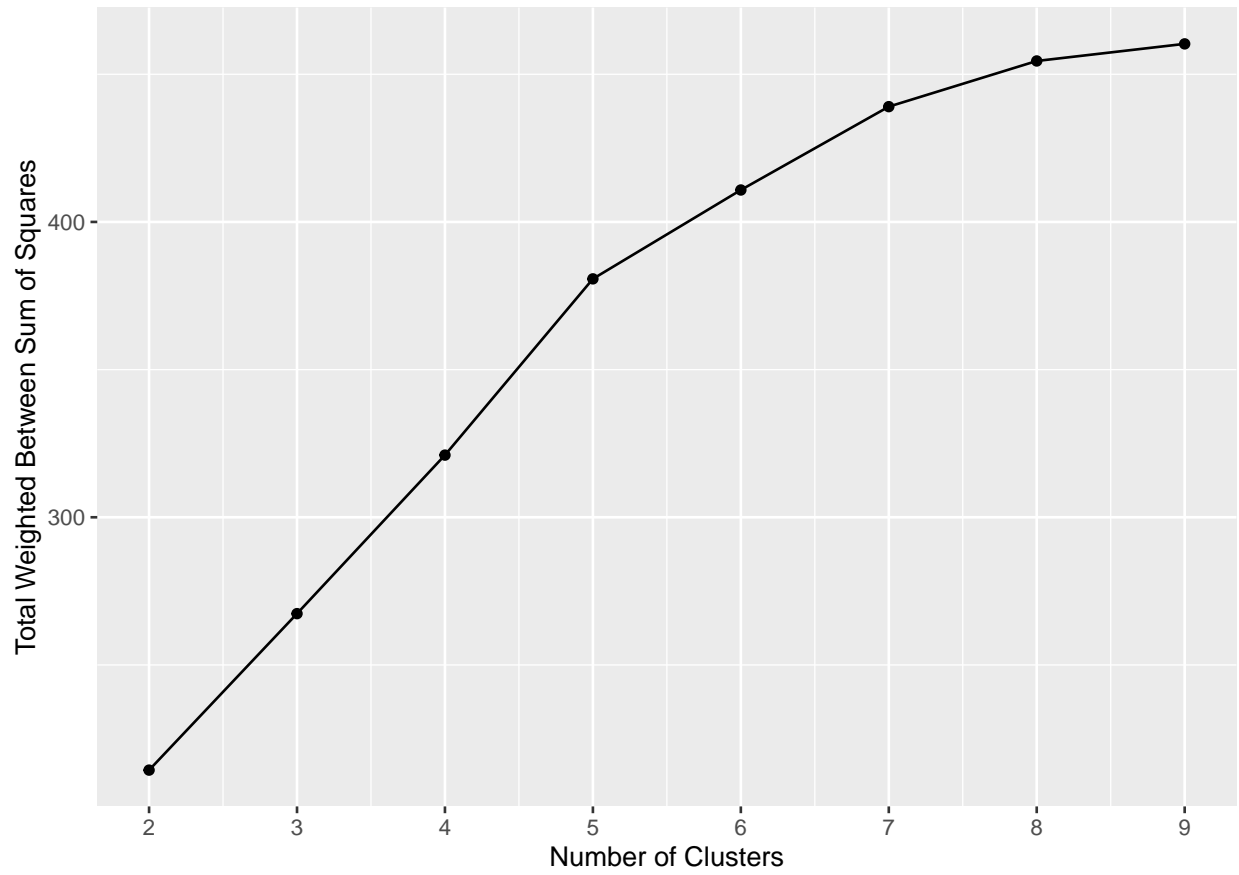
RSKC (10 Runs)

This chunk serves as a snapshot of the RSKC over 10 runs (i.e. different set.seed values) in order to visualize the variation in cluster label assignments for each observation. The 10 runs were carried out for each of $K = 2, 3, 4, 5, 6$.

```r
set.seed(72613)

while (T) {

  # Assign the values of 2, 3, 4, 5, 6 to 'clust_vect'.
  clust_vect <- c(2, 3, 4, 5, 6)

  # Assign empty lists to 'rskc.results.list' and 'rskc.weighted.list'
  # to store the results and the RSKC weighted data frames that result
  # from the clustering.
  rskc_results_list <- list()
  rskc_weighted_list <- list()

  # Assign 6 colours to 'col_vect'.
  col_vect <- c("#FF0000",
                "#0000FF",
                "#00FF00",
                "#A020F0",
                "#FFA500",
                "#FFFF00")

  # Assign an empty list to 'tsne_list_2', 'tsne_list_3', 'tsne_list_4', and
  # 'tsne_list_6'.
  tsne_list_2 <- list()
  tsne_list_3 <- list()
  tsne_list_4 <- list()
  tsne_list_5 <- list()
  tsne_list_6 <- list()

  # Assign an empty list to 'weight_list'.
  weight_list <- list()

  # Create empty data frames to store the cluster assignments and cluster weights
  # for each of the 10 runs.
  rskc_region_labels_2 <- data.frame("Region" = rownames(myFidelity))
  rskc_region_labels_3 <- data.frame("Region" = rownames(myFidelity))
  rskc_region_labels_4 <- data.frame("Region" = rownames(myFidelity))
  rskc_region_labels_5 <- data.frame("Region" = rownames(myFidelity))
  rskc_region_labels_6 <- data.frame("Region" = rownames(myFidelity))
  rskc_region_weights <- data.frame("Case" = colnames(myFidelity))

  # For 'i' -- the current number of clusters -- in 'clust_vect'...
  for (i in clust_vect) {

    ###### RSKC (10 Runs) ######

    # Create a vector of seeds for all 10 runs.
    set.seed(72613)
```

```r
x = rdunif(10, a = 1, b = 1000000)

for (counter in 1:10) {

  # Set the seed.
  set.seed(x[counter])

  # Perform RSKC for whatever-the-value-of-''-is many clusters using
  # 'myFidelity', which has brain region as rows and gene_celltype as columns.
  # Assign RSKC's output as an entry in 'rskc_results_list'.
  rskc_results_list[[counter]] <- RSKC(myFidelity,
                                       alpha = 0.1,
                                       ncl = i,
                                       L1 = sqrt(ncol(myFidelity)))

  # Use the following if statements to add the cluster assignments for run i
  # to the corresponding 'rskc_region_labels_2', 'rskc_region_labels_3',
  # 'rskc_region_labels_4', 'rskc_region_labels_5', 'rskc_region_labels_6'.
  if (i == 2) {
    rskc_region_labels_2[counter + 1] <- rskc_results_list[[counter]]$labels
    colnames(rskc_region_labels_2)[counter + 1] <- paste("Run_", counter, sep = "")
  }

  if (i == 3) {
    rskc_region_labels_3[counter + 1] <- rskc_results_list[[counter]]$labels
    colnames(rskc_region_labels_3)[counter + 1] <- paste("Run_", counter, sep = "")
  }

  if (i == 4) {
    rskc_region_labels_4[counter + 1] <- rskc_results_list[[counter]]$labels
    colnames(rskc_region_labels_4)[counter + 1] <- paste("Run_", counter, sep = "")
  }

  if (i == 5) {
    rskc_region_labels_5[counter + 1] <- rskc_results_list[[counter]]$labels
    colnames(rskc_region_labels_5)[counter + 1] <- paste("Run_", counter, sep = "")
  }

  if (i == 6) {
    rskc_region_labels_6[counter + 1] <- rskc_results_list[[counter]]$labels
    colnames(rskc_region_labels_6)[counter + 1] <- paste("Run_", counter, sep = "")
  }

  # Add the variable weights for run i to the 'rskc_region_weights'
  rskc_region_weights[counter + 1] <- rskc_results_list[[counter]]$weights
  colnames(rskc_region_weights)[counter + 1] <- paste("Run_", counter, sep = "")

  # For the current object in 'rskc_list' convert the cluster labels
  # into characters, and assign them to a new column called 'cluster_labels'
  # in 'fidelity'.
  fidelity$cluster_labels <- rskc_results_list[[counter]]$labels %>%
    as.character()
```

```r
###### Apply weights from RSKC to myFidelity ######

# Create vector of the weights obtained from RSKC and assign them to 'weights'.
# Make empty matrix 'weighted_fidelity' for new weighted fidelity scores.
weights <- as.matrix(rskc_results_list[[1]]$weights)

# Multiply 'myFidelity' columns containing gene_celltype by corresponding
# weights obtained from RSKC.
weighted_fidelity <- sweep(t(myFidelity), MARGIN = 1, weights, `*`) %>%
  t()

###### tSNE (on weighted data) ######

# Run tsne on weighted fidelity scores, and assign to 'tsne'
set.seed(72613)
tsne <- Rtsne(weighted_fidelity, perplexity = 2)

# Create new df 'tsne_out' which contains the two dimensions obtained from tSNE
# and corresponding regions
tsne_out <- tsne$Y %>%
  data.frame(regions) %>%
  rename(Brain.Region = regions, V1 = X1, V2 = X2) #rename columns

# Merge 'tsne_out' with 'fidelity' according
# to their shared 'Brain.Region' column, and assign to
# 'tsne_genes_regions_clusts'.
tsne_genes_regions_clusts <- merge(tsne_out,
                                   fidelity,
                                   by = "Brain.Region")

# Create a tSNE scatter plot where each point is colour-coded according to
# its designated RSKC cluster and assign this figure to 'tsne_scatter'.
tsne_scatter <- ggplot(tsne_genes_regions_clusts,
                       aes(V1,
                           V2,
                           fill = cluster_labels)) +
  geom_point(shape = 21, size = 3) +
  scale_fill_manual(values = col_vect[1:i],
                    labels = 1:i,
                    name = "Cluster") +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        panel.background = element_rect(fill = NA,
                                        colour = "white"),
        panel.border = element_blank(),
        axis.line = element_line(),
        legend.position = "bottom",
        legend.background = element_rect(fill = NA,
                                         colour = NA),
        legend.title.align = 0.5) +
  guides(fill = guide_legend(nrow = 2,
```

```
                            ncol = 4,
                            byrow = TRUE)) +
    labs(x = "V1", y = "V2")

  # Use if statements to assign 'tsne_scatter' as an entry in 'tsne_list_2',
  # 'tsne_list_3', 'tsne_list_4', or 'tsne_list_5', 'tsne_list_5' depending
  # on the current i.
   if (i == 2) {
    tsne_list_2[[counter]] <- tsne_scatter
  }

  if (i == 3) {
    tsne_list_3[[counter]] <- tsne_scatter
  }

  if (i == 4) {
    tsne_list_4[[counter]] <- tsne_scatter
  }

  if (i == 5) {
    tsne_list_5[[counter]] <- tsne_scatter
  }

  if (i == 6) {
    tsne_list_6[[counter]] <- tsne_scatter
  }

  ###### For Bubble Plot ######

  # Make a separate data frame that includes the cluster labels appended to
  # 'fidelity' when:
  # K = 2
  if ((i == 2) & (counter == 1)) {
    fidelity_2_clusts <- fidelity
  }

  # K = 3
  if ((i == 3) & (counter == 1)) {
    fidelity_3_clusts <- fidelity
  }

  # K = 4
  if ((i == 4) & (counter == 1)) {
    fidelity_4_clusts <- fidelity
  }

  # K = 5
  if ((i == 5) & (counter == 1)) {
    fidelity_5_clusts <- fidelity
  }

  # K = 6
  if ((i == 6) & (counter == 1)) {
```

```
        fidelity_6_clusts <- fidelity
      }

    }

  }

  # Break out of the while-loop when for-loop is done.
  break

}
```

V2

V1

Cluster   1   2   3   4   5

V2

V1

Cluster   1   2   3   4   5

V2

V1

Cluster   1   2   3   4   5

V2

V1

Cluster   1   2   3   4   5

V2

V1

Cluster   1   2   3   4   5

V2

V1

Cluster   1   2   3   4   5

V2

V1

Cluster   1   2   3   4   5

V2

V1

Cluster   1   2   3   4   5

V2

V1

Cluster   1   2   3   4   5

V2

V1

Cluster   1   2   3   4   5

V2

V1

Cluster
● 1 ● 2 ● 3 ● 4
● 5 ● 6

V2

V1

Cluster
● 1 ● 2 ● 3 ● 4
● 5 ● 6

V2

V1

Cluster
● 1 ● 2 ● 3 ● 4
● 5 ● 6

V2

V1

Cluster
● 1 ● 2 ● 3 ● 4
● 5 ● 6

V2

V1

Cluster
● 1 ● 2 ● 3 ● 4
● 5 ● 6

19
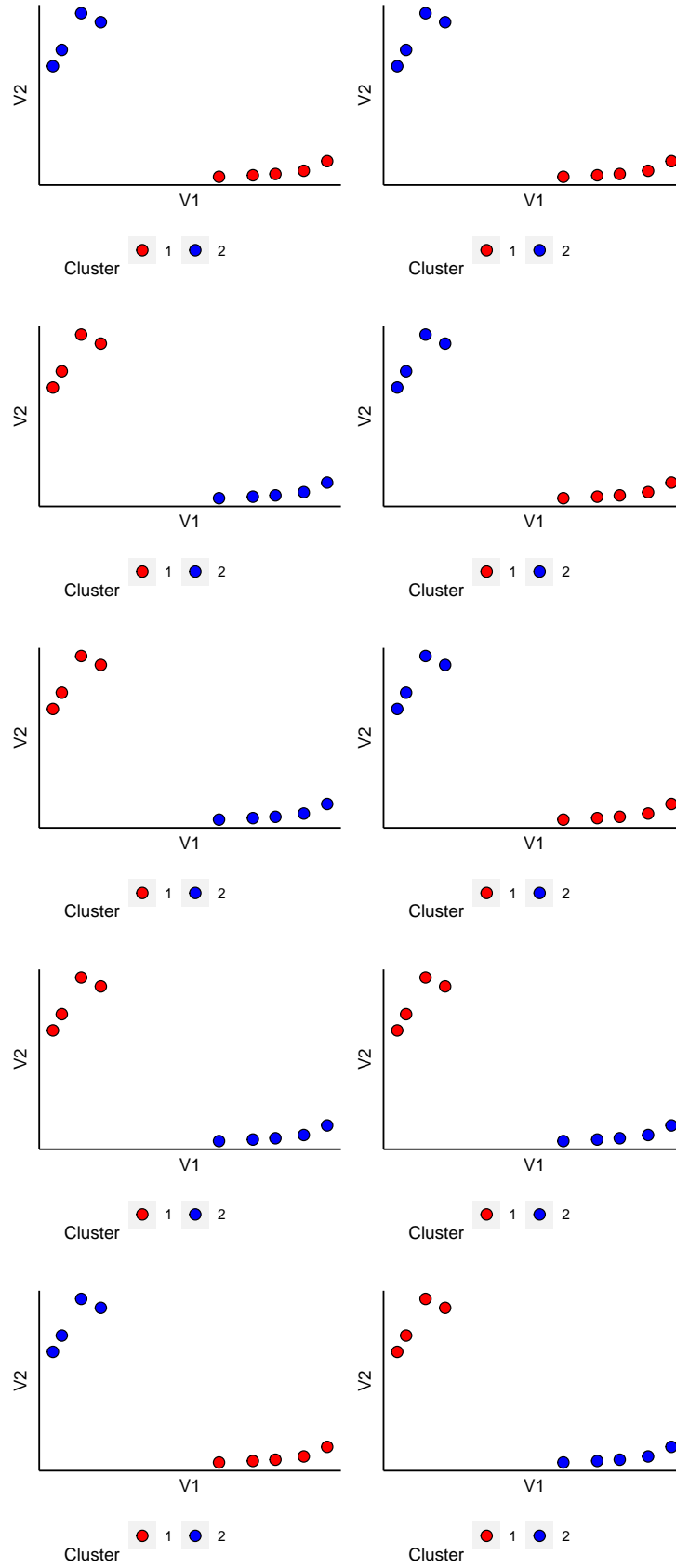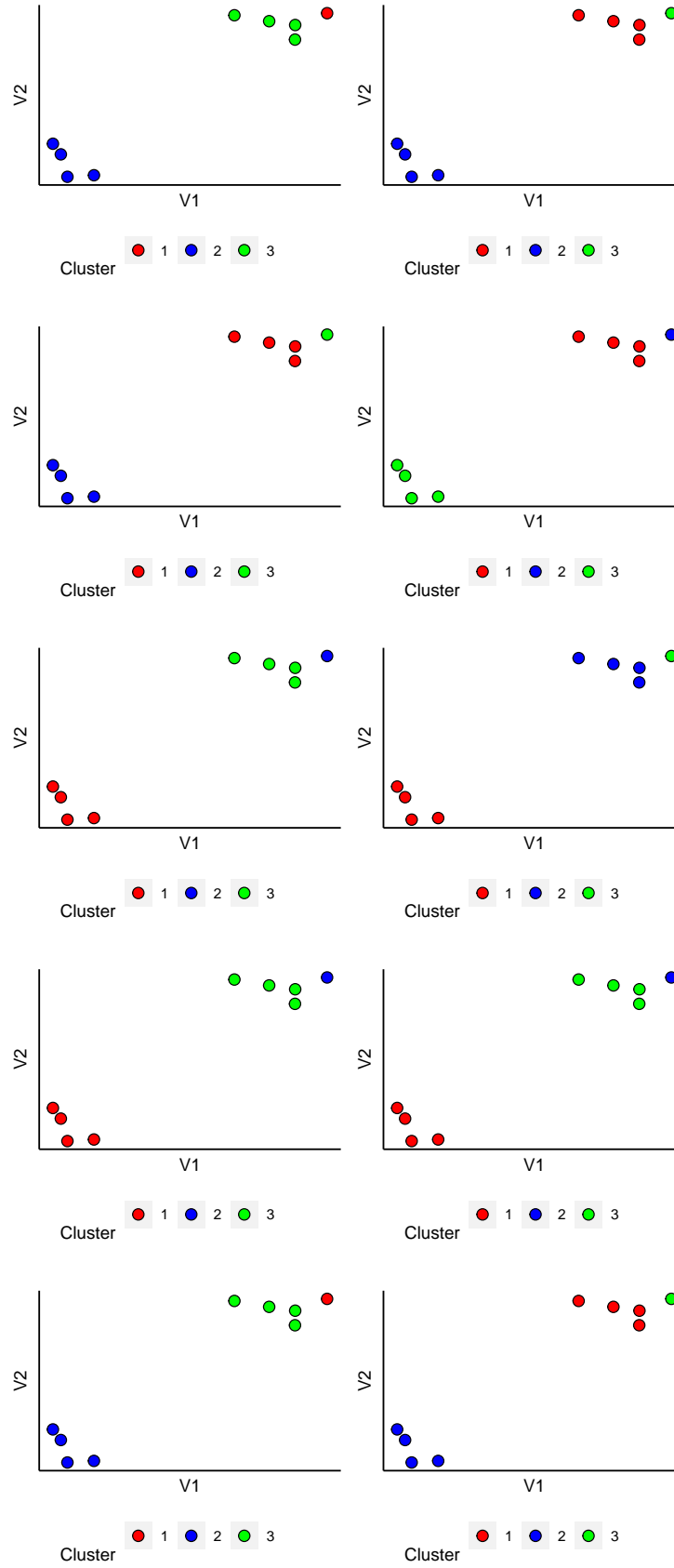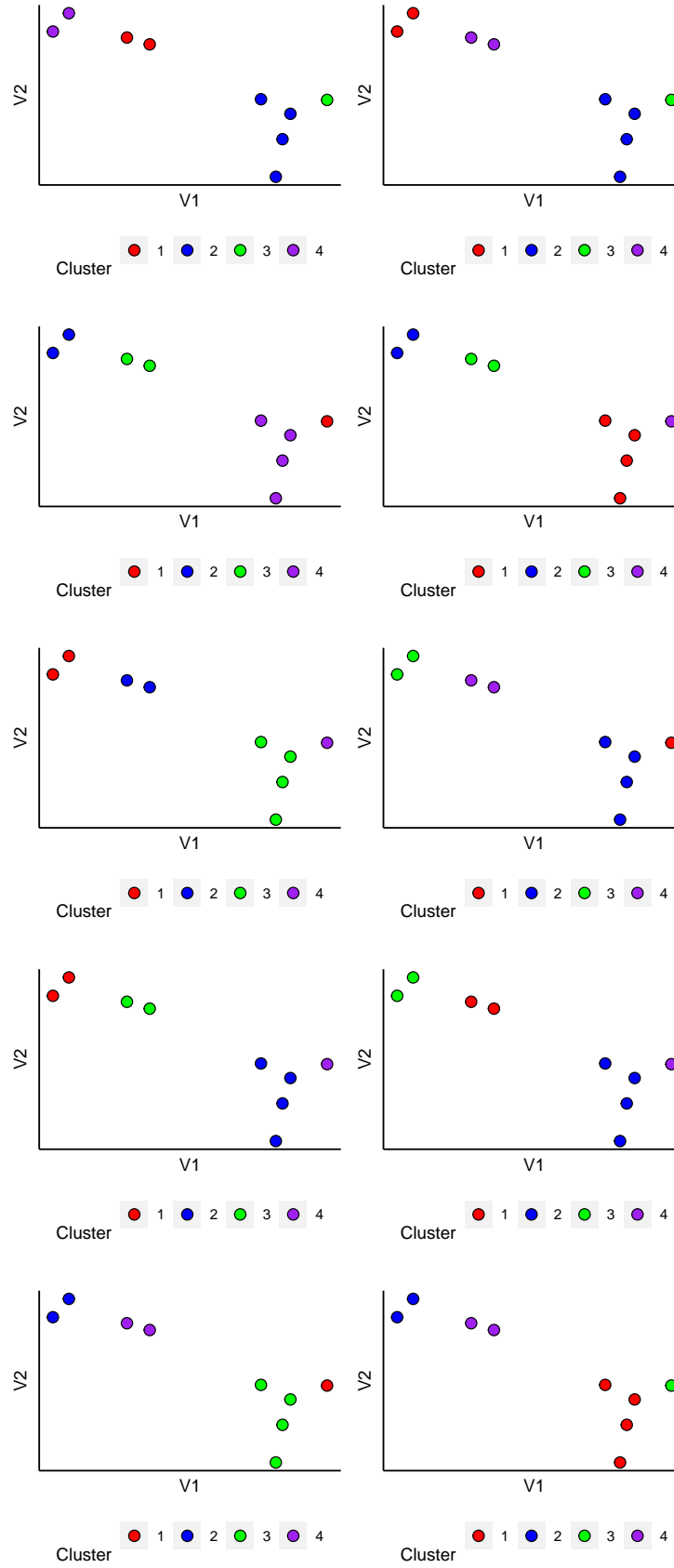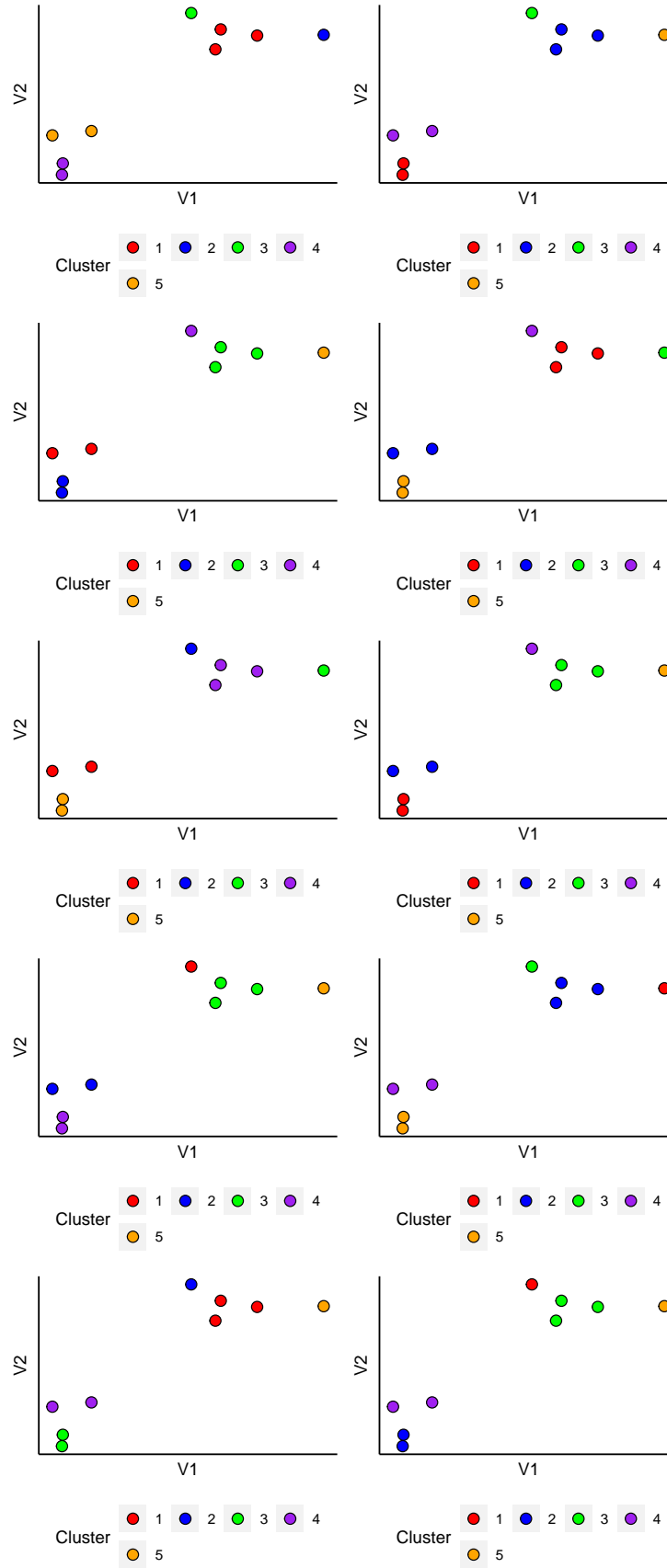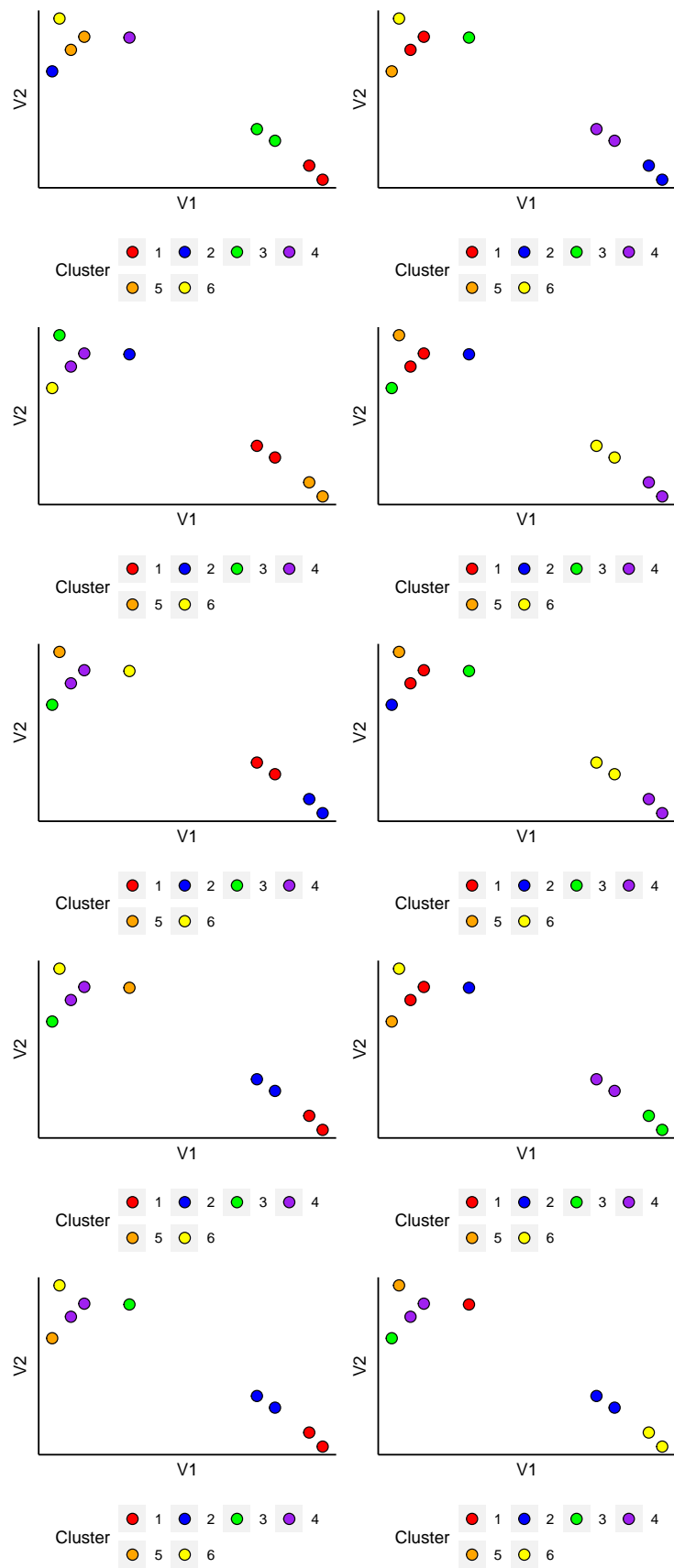
RSKC (100 Runs)

Having previously selected the number of clusters to use for RSKC, this portion of the code aims to evaluate the proportion of brain regions clustered together over 100 runs (i.e. 100 different set.seed values).

```r
# Create empty lists to store the results and the RSKC weighted data frames
# that result from the clustering.
rskc.results.list <- list()
rskc.weighted.list <- list()

# Create empty data frames to store the cluster assignments and cluster weights
# for each of the 100 runs.
rskc.region.labels <- data.frame("Region" = rownames(myFidelity))
rskc.region.weights <- data.frame("Case" = colnames(myFidelity))

# Create a vector of seeds for all 100 runs.
set.seed(72613)
x = rdunif(100, a = 1, b = 1000000)

for (i in 1:100) {

  # Set the seed.
  set.seed(x[i])

  # Perform RSKC clustering on the data; the number of clusters is selected
  # a priori.
  rskc.results.list[[i]] = RSKC(myFidelity,
                                alpha = 0.1,
                                ncl = 3,
                                L1 = sqrt(ncol(myFidelity)))

  # Add the cluster assignments for run i to the 'rskc.region.labels'.
  rskc.region.labels[i + 1] = rskc.results.list[[i]]$labels
  colnames(rskc.region.labels)[i + 1] = paste("Run_", i, sep = "")

  # Add the variable weights for run i to the 'rskc.region.weights'
  rskc.region.weights[i + 1] = rskc.results.list[[i]]$weights
  colnames(rskc.region.weights)[i + 1] = paste("Run_", i, sep = "")

  # Create a list of data frames containing the clustering data multiplied by
  # the corresponding RSKC variable weights.
  rskc.weighted.list[[i]] = sweep(myFidelity, 2,
                                  rskc.results.list[[i]]$weights, "*")
}
```

Calculating Proportion of Times Brain Regions are Clustered Together

```r
# Transpose the data frame with the cluster labels from the 100 runs.
rskc.region.labels.t <- rskc.region.labels %>%
  column_to_rownames("Region") %>%
  t() %>%
  data.frame()
```

20

```r
# Create an empty 9x9 data frame
rskc.cluster.regions.wide <- data.frame(matrix(ncol = 9, nrow = 9)) %>%
  # Set the column and row names as the regions (ordered alphabetically)
  set_colnames(regions) %>%
  set_rownames(regions)

# Add in the number of matches for each ith row/jth column combination to
# create the adjacency matrix.
for (i in 1:length(regions)) {
  for (j in 1:length(regions)) {

    rskc.cluster.regions.wide[i, j] = sum(rskc.region.labels.t[[regions[i]]] == rskc.region.labels.t[[re

  }
}

# Convert the adjacency matrix to long format for various plotting purposes
rskc.cluster.regions.long <- rskc.cluster.regions.wide %>%
  # Give the region row names their own column
  rownames_to_column("Region_1") %>%
  # Lengthen the data with melt() so we have three columns:
  # Region_1, Region_2 and the total number of matches.
  melt(id.vars = "Region_1", variable.name = "Region_2", value.name = "Matches") %>%
  # Convert the number of matches to a proportion
  mutate(Matches = Matches / 100)
```

Heat Maps to Visualize Proportion of Shared Clusters

```r
# Factor the regions with levels corresponding to the specified brain region
# names from the 'regions' vector.
region.rskc.cluster.matches.ordered <- rskc.cluster.regions.long %>%
  # Factor Region_1
  mutate(Region_1 = factor(Region_1, levels = regions)) %>%
  # Factor Region_2
  mutate(Region_2 = factor(Region_2, levels = regions))

# Plot the adjacency matrix as a heat map
region_heatmap1 <- ggplot(region.rskc.cluster.matches.ordered,
                          aes(x = Region_1, y = Region_2, fill = Matches)) +
  geom_tile() +
  scale_fill_gradientn(colours = brewer.pal(n = 9, name = "YlOrRd")) +
  labs(x = NULL,
       y = NULL,
       fill = "Proportion of Matches on \n100 RSKC Runs") +
  theme_minimal() +
  theme(panel.border = element_blank(),
        panel.background = element_blank(),
        axis.text.x = element_text(size = 12, colour = "black",
                                   angle = 45,
                                   vjust = 0.6),
```
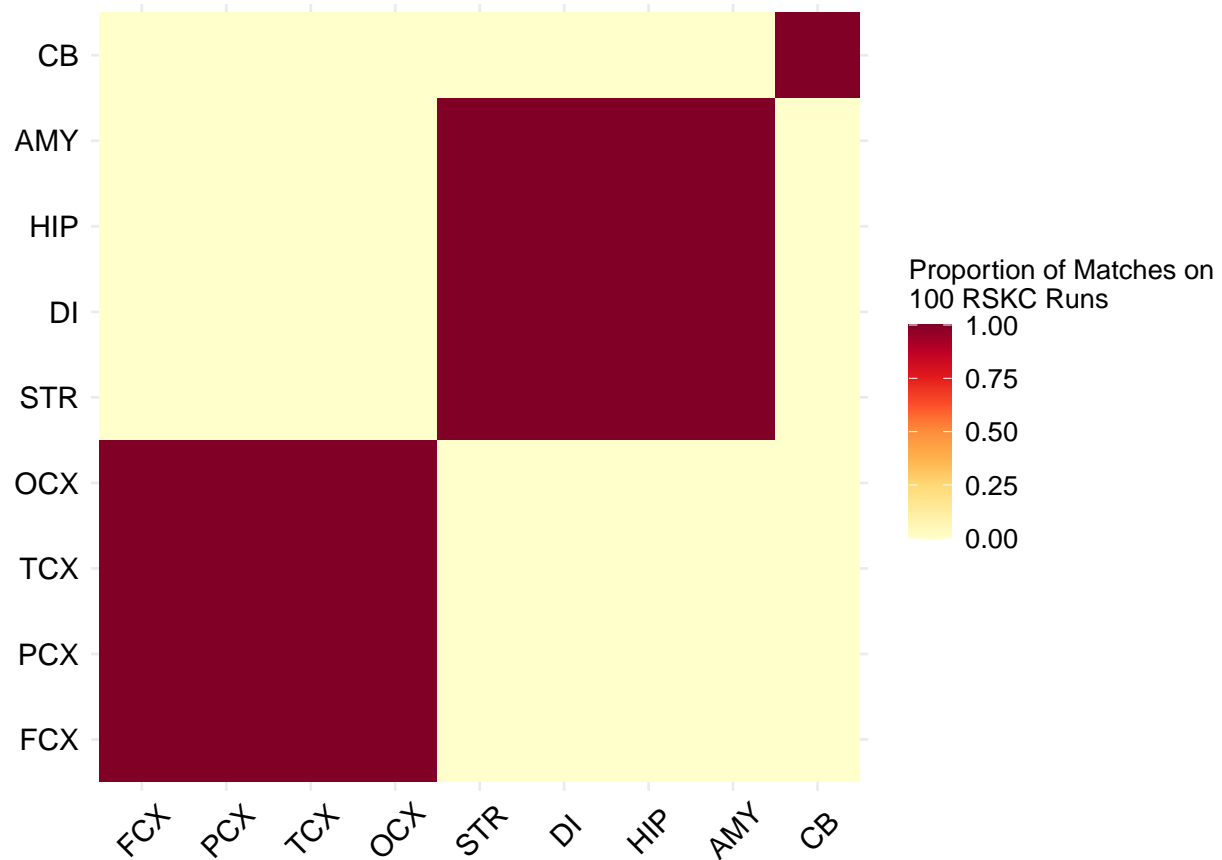
21

```
        axis.text.y = element_text(size = 12, colour = "black"),
        legend.text = element_text(size = 11, colour = "black"),
        axis.ticks = element_blank())
region_heatmap1
```

Unclustered Heat Map


```
# Convert all values in the 'rskc.cluster.regions.wide' matrix into
# proportions.
rskc.region.prop <- apply(rskc.cluster.regions.wide, 2, function(x) return(x/100)) %>%
  as.matrix()

# Assign a dendrogram for the matrix of matched clustering of RSKC regions.
rskc_dendro <- set(as.dendrogram(hclust(dist(rskc.region.prop))), "branches_lwd", 3)

# Define a colour 'ylOrRd' from brewer palette.
ylOrRd <- colorRampPalette(brewer.pal(n = 9, name = "YlOrRd"))

heatmap.2(rskc.region.prop,
          scale = "none",
```

```
col = ylOrRd(100),
Rowv = rskc_dendro,
Colv = rskc_dendro,
key = TRUE,
key.xlab = "Proportion of Matches on \n 100 RSKC Runs",
trace = "none", # Remove the histogram trace from heat map
density.info = "none", # Remove the histogram from color key
srtCol = 45, # Rotate column labels on heat map
margins = c(5, 10),
# Include only first and last labels on x-axis of legend
key.xtickfun = function(){
  breaks = pretty(parent.frame()$breaks)
  breaks = breaks[c(1, length(breaks))]
  list(at = parent.frame()$scale01(breaks),
       labels = breaks)
}
)
```

Hierarchical Clustered Heat Map
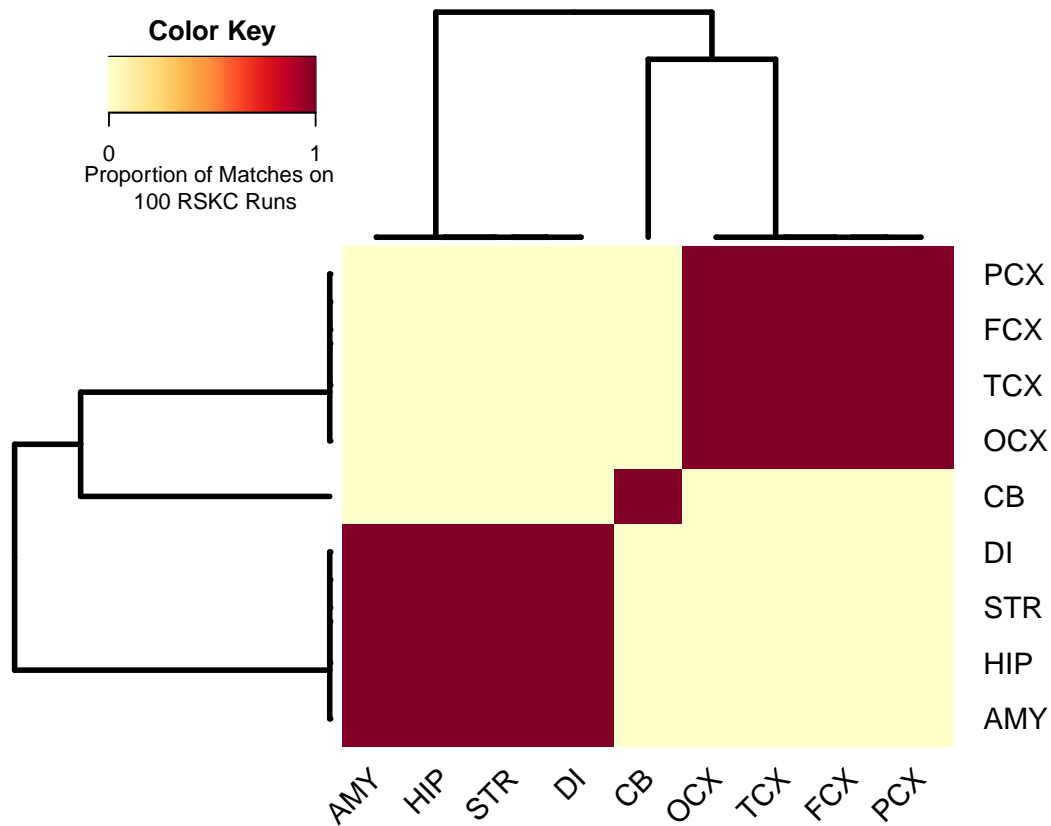
Select Gene and Cell Combinations of Interest Based on RSKC Weights (100 Runs)

```
# Take the average weight for each gene_celltype across 100 RSKC runs.
weights_100_runs <- rowMeans(rskc.region.weights[-1]) %>%
  as_vector()

# Obtain list of top eight largest average weights and their corresponding
# gene_celltype.
rskc_weights_top_8 <- rskc.region.weights %>%
  mutate(Average_Weight = weights_100_runs) %>%
  select("Case", "Average_Weight") %>%
  arrange(desc(Average_Weight)) %>%
  head(8) %>%
  rename(Gene_Cell = Case)
rskc_weights_top_8
```

```
##                 Gene_Cell Average_Weight
## 1 ADAM17_Oligodendrocyte      0.5800867
## 2          NRP1_Astrocyte      0.4990396
## 3   NRP1_Oligodendrocyte      0.3941349
## 4          DPP4_Astrocyte      0.3130029
## 5          NRP1_Microglia      0.2475625
## 6           ADAM17_Neuron      0.1621562
## 7         ADAM17_Microglia      0.1369581
## 8  FURIN_Oligodendrocyte      0.1151876
```

```
# Select only the top eight features from the 'fidelity_3_clusts' data frame.
fidelity_top_features <- fidelity_3_clusts[, names(fidelity_3_clusts) %in%
                                             rskc_weights_top_8$Gene_Cell]

# Append the brain region and cluster labels to the 'fidelity_top_features'
# data frame.
fidelity_features_clusts <- bind_cols(fidelity_3_clusts[, c(1, ncol(fidelity_3_clusts))],
                                      fidelity_top_features)

# Create a duplicate data frame containing column for cluster labels that are
# reassigned based on the order of appearance in the clustered heat map (i.e.
# clusters in order starting at 1 when reading the heat map left to right).
fidelity_features_clusts <- fidelity_features_clusts %>%
  select(cluster_labels) %>%
  mutate(cluster_labels2 = case_when(
    cluster_labels == 1 ~ 2,
    cluster_labels == 2 ~ 3,
    cluster_labels == 3 ~ 1)) %>%
  select(cluster_labels2) %>%
  add_column(fidelity_features_clusts)
```

Prepare Data Frame for Fidelity Statistics Based on Clusters

```r
# Compute the mean fidelity score for each gene_celltype for each cluster label.
# Pivot data frame for further plotting steps.
fidelity_clust_bubble_mean <- fidelity_features_clusts %>%
  group_by(cluster_labels2) %>%
  summarize_at(vars(as_vector(rskc_weights_top_8$Gene_Cell)), mean) %>%
  mutate(cluster_labels2 = as.factor(cluster_labels2)) %>%
  melt(id = c("cluster_labels2")) %>%
  rename("Mean_Fidelity" = value, "Gene_Cell" = variable)

# Make a vector that contains the selected gene and corresponding cell type
# in order from largest to smallest RSKC weights.
gene_level_order <- as_vector(rskc_weights_top_8$Gene_Cell)

# Modify the order of the gene_celltype in the 'fidelity_clust_bubble' data
# frame to match the desired order from 'gene_level_order'.
fidelity_clust_bubble_mean$Gene_Cell <- as.character(fidelity_clust_bubble_mean$Gene_Cell)
fidelity_clust_bubble_mean$Gene_Cell <- factor(fidelity_clust_bubble_mean$Gene_Cell,
                                                levels = rev(gene_level_order))
```

Bubble Plot to Visualize Important Gene and Cell Types in Clusters

```r
# Create bubble plot showing cluster label vs. gene and cell type. Each point
# is coloured by the mean fidelity score.
bubble_clust <- ggplot(fidelity_clust_bubble_mean, aes(x = Gene_Cell,
                                                        y = cluster_labels2)) +
  geom_point(aes(fill = Mean_Fidelity), colour = "black", pch = 21, size = 7) +
  coord_flip() +
  # Blue-white-red colour gradient
  scale_fill_gradient2(low = "blue",
                       mid = "white",
                       midpoint = 0,
                       high = "red") +
  # scale_fill_viridis(direction = -1) + # Viridis colour gradient
  labs(x = "Gene and Cell Type", # Arranged in smallest to largest RSKC weights
       y = "Cluster",
       fill = "Mean Fidelity Score") +
  theme(axis.text.x = element_text(colour = "black", size = 10),
        axis.text.y = element_text(colour = "black", size = 10),
        axis.ticks.x = element_blank(),
        axis.ticks.y = element_blank(),
        axis.title.x = element_text(size = 12, margin = margin(t = 10,
                                                               r = 0,
                                                               b = 0,
                                                               l = 0)),
        axis.title.y = element_text(size = 12, margin = margin(t = 0,
                                                               r = 10,
                                                               b = 0,
                                                               l = 0)),
        legend.title = element_text(size = 10),
```

```
        legend.key = element_blank(),
        panel.grid.major = element_line(colour = "grey", size = 0.2),
        panel.background = element_blank(),
        panel.border = element_blank(),
        legend.position = "right")
bubble_clust
```
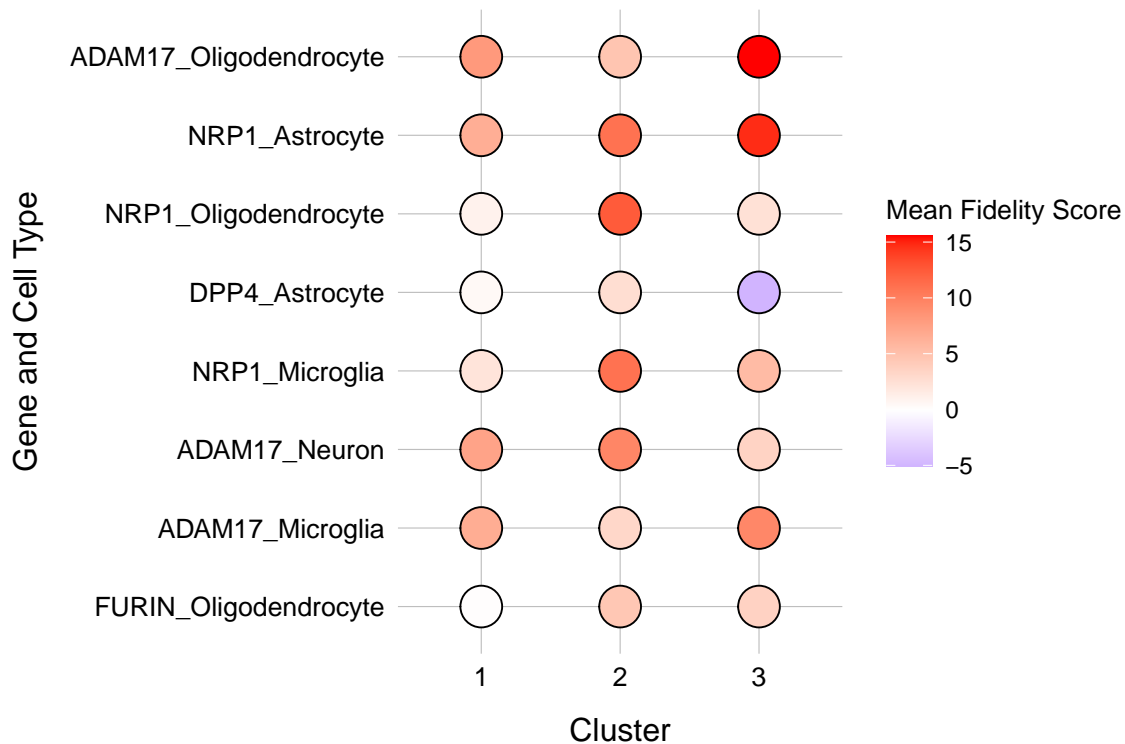

## Prepare Data Frame for Fidelity Statistics Based on Brain Regions

```
# Compute the mean fidelity score for each gene_celltype for each brain region.
# Pivot data frame for further plotting steps.
fidelity_region_bubble_mean <- fidelity_features_clusts %>%
  group_by(Brain.Region) %>%
  summarize_at(vars(as_vector(rskc_weights_top_8$Gene_Cell)), mean) %>%
  melt(id = c("Brain.Region")) %>%
  rename("Mean_Fidelity" = value, "Gene_Cell" = variable)

# Modify the order of the gene_celltype in the 'fidelity_region_bubble_mean'
# data frame to match the desired order from 'gene_level_order'.
fidelity_region_bubble_mean$Gene_Cell <- as.character(fidelity_region_bubble_mean$Gene_Cell)
fidelity_region_bubble_mean$Gene_Cell <- factor(fidelity_region_bubble_mean$Gene_Cell,
                                    levels = rev(gene_level_order))

# Rearrange order of dendrogram leaves such that cortices start at the left.
rskc_dendro2 <- rotate(rskc_dendro, c(7, 8, 9, 6, 5, 1:4))
```

```
# Plot dendrogram made previously for heat map, with rearranged order of brain
# regions
dendro_plot <- plot(rskc_dendro2 %>% set("branches_lwd", 4))
```
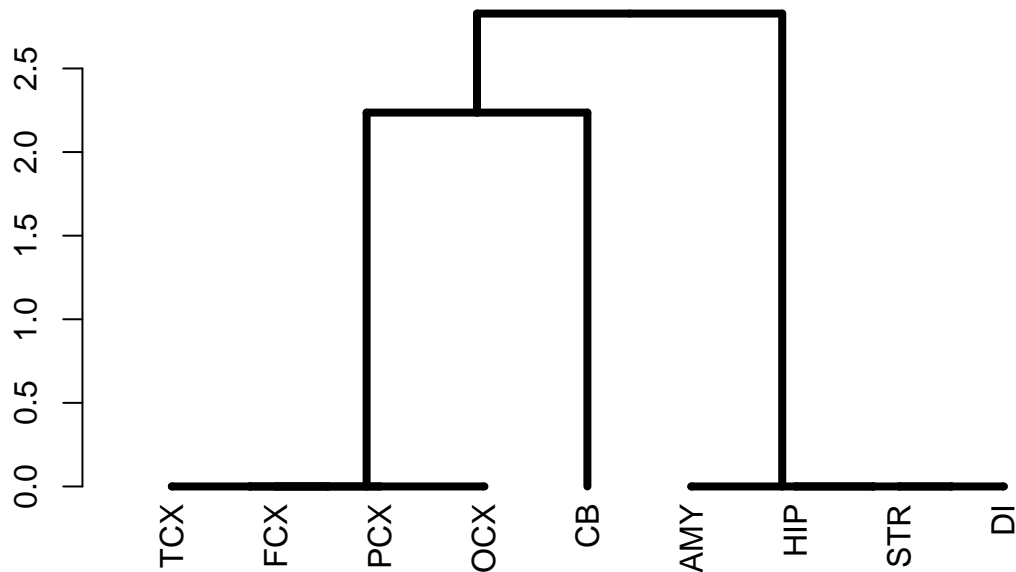


```
# Make a vector that contains the brain regions in order of the four clusters.
region_clust_order <- as.list(labels(rskc_dendro2))

# Modify the order of the brain region in the 'fidelity_region_bubble_mean' data
# frame to match the desired order from 'region_clust_order'.
fidelity_region_bubble_mean$Brain.Region <- as.character(fidelity_region_bubble_mean$Brain.Region)
fidelity_region_bubble_mean$Brain.Region <- factor(fidelity_region_bubble_mean$Brain.Region,
                                                   levels = region_clust_order)
```

Bubble Plot to Visualize Important Gene and Cell Types per Brain Region

```
# Create bubble plot showing brain region vs. gene and cell type. Each point
# is coloured by the mean fidelity score.
bubble_region <- ggplot(fidelity_region_bubble_mean, aes(x = Gene_Cell,
                                                         y = Brain.Region)) +
  geom_point(aes(fill = Mean_Fidelity), colour = "black", pch = 21, size = 7) +
  coord_flip() +
  # Blue-white-red colour gradient
  scale_fill_gradient2(low = "blue",
```

```
                      mid = "white",
                      midpoint = 0,
                      high = "red") +
 # scale_fill_viridis(direction = -1) + # Viridis colour gradient
 labs(x = "Gene and Cell Type", # Arranged in smallest to largest RSKC weights
      y = "Brain Region",
      fill = "Mean Fidelity Score",
      size = "Size") +
 theme(axis.text.x = element_text(colour = "black", size = 10, angle = 45,
                                  margin = margin(t = 10,
                                                  r = 10,
                                                  b = 0,
                                                  l = 0)),
       axis.text.y = element_text(colour = "black", size = 10),
       axis.ticks.x = element_blank(),
       axis.ticks.y = element_blank(),
       axis.title.x = element_text(size = 12),
       axis.title.y = element_text(size = 12, margin = margin(t = 0,
                                                              r = 10,
                                                              b = 0,
                                                              l = 0)),
       legend.title = element_text(size = 10),
       legend.key = element_blank(),
       panel.grid.major = element_line(colour = "grey", size = 0.2),
       panel.background = element_blank(),
       panel.border = element_blank(),
       legend.position = "right")
bubble_region
```
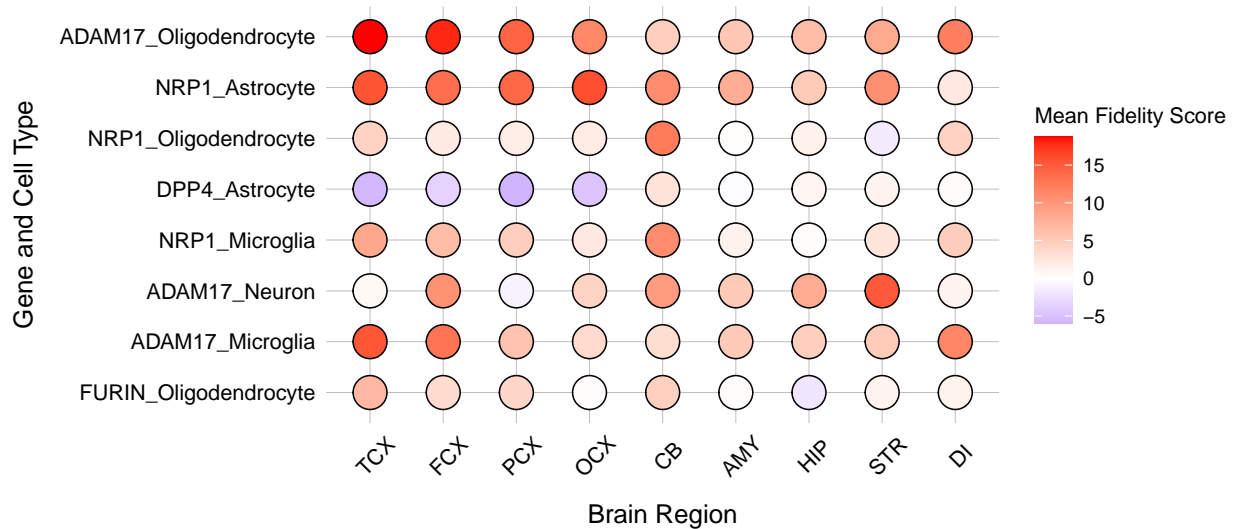
Visualizing Summary Statistics for Fidelity Score Data

Violin Plot for Visualizing Cell Type Fidelity Scores

```r
# Prepare a data frame containing separate columns for the cell type and
# fidelity score.
fidelity_celltype <- pivot_longer(fidelity_3_clusts, cols = -c("Brain.Region", "cluster_labels"),
                                  names_to = c("Gene", "Cell.Subtype"),
                                  names_sep = "_",
                                  values_to = "Fidelity")

# Create violin plot showing the fidelity score distributions for each cell
# type.
celltype_viol <- ggplot(fidelity_celltype, aes(x = Cell.Subtype, y = Fidelity,
                                               fill = Cell.Subtype)) +
  geom_violin(trim = FALSE) +
  geom_boxplot(width = 0.1, fill = "white") +
  labs(x = "Cell Type",
       y = "Fidelity Score",
       fill = "Cell Type") +
  theme(axis.text.x = element_text(colour = "black", size = 12),
        axis.text.y = element_text(colour = "black", size = 12),
        axis.title.x = element_text(size = 14, face = "bold"),
        axis.title.y = element_text(size = 14, face = "bold"),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(size = 0.5, colour = "black"))
celltype_viol
```
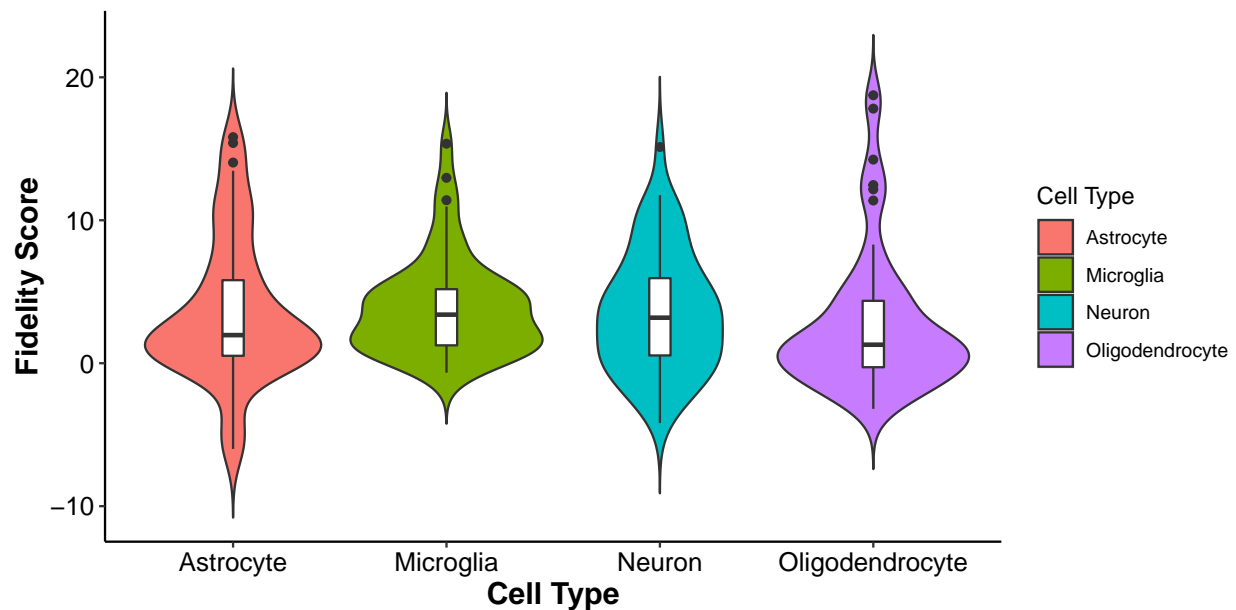
## Raincloud Plots for Fidelity Scores by Cluster

```r
# Define colour palette for the clusters.
cluster_colours <- c("#4477AA", "#66CCEE", "#228833")#, "#CCBB44", "#EE6677", "#AA3377")

# Create raincloud plot showing the fidelity score distributions for each
# cluster.
celltype_raincloud <- ggplot(fidelity_celltype, aes(x = cluster_labels,
                                                     y = Fidelity,
                                                     fill = cluster_labels)) +
  geom_violinhalf(position = position_nudge(x = 0.15, y = 0)) +
  coord_flip() +
  scale_fill_manual(values = cluster_colours) +
  scale_colour_manual(values = cluster_colours) +
  scale_y_continuous(breaks = c(-5, 0, 5, 10, 15)) +
  geom_jitter(aes(colour = cluster_labels),
              width = 0.1, size = 1, alpha = 0.5, show.legend = FALSE) +
  geom_boxplot(width = 0.1, outlier.shape = NA, alpha = 0.5) +
  labs(x = "Cluster",
       y = "Fidelity Score",
       fill = "Cluster") +
  theme(axis.text.x = element_text(colour = "black", size = 12),
        axis.text.y = element_text(colour = "black", size = 12),
        axis.title.x = element_text(size = 14, face = "bold"),
        axis.title.y = element_text(size = 14, face = "bold"),
        legend.position = "none",
        legend.key = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(size = 0.5, colour = "black"))
celltype_raincloud
```
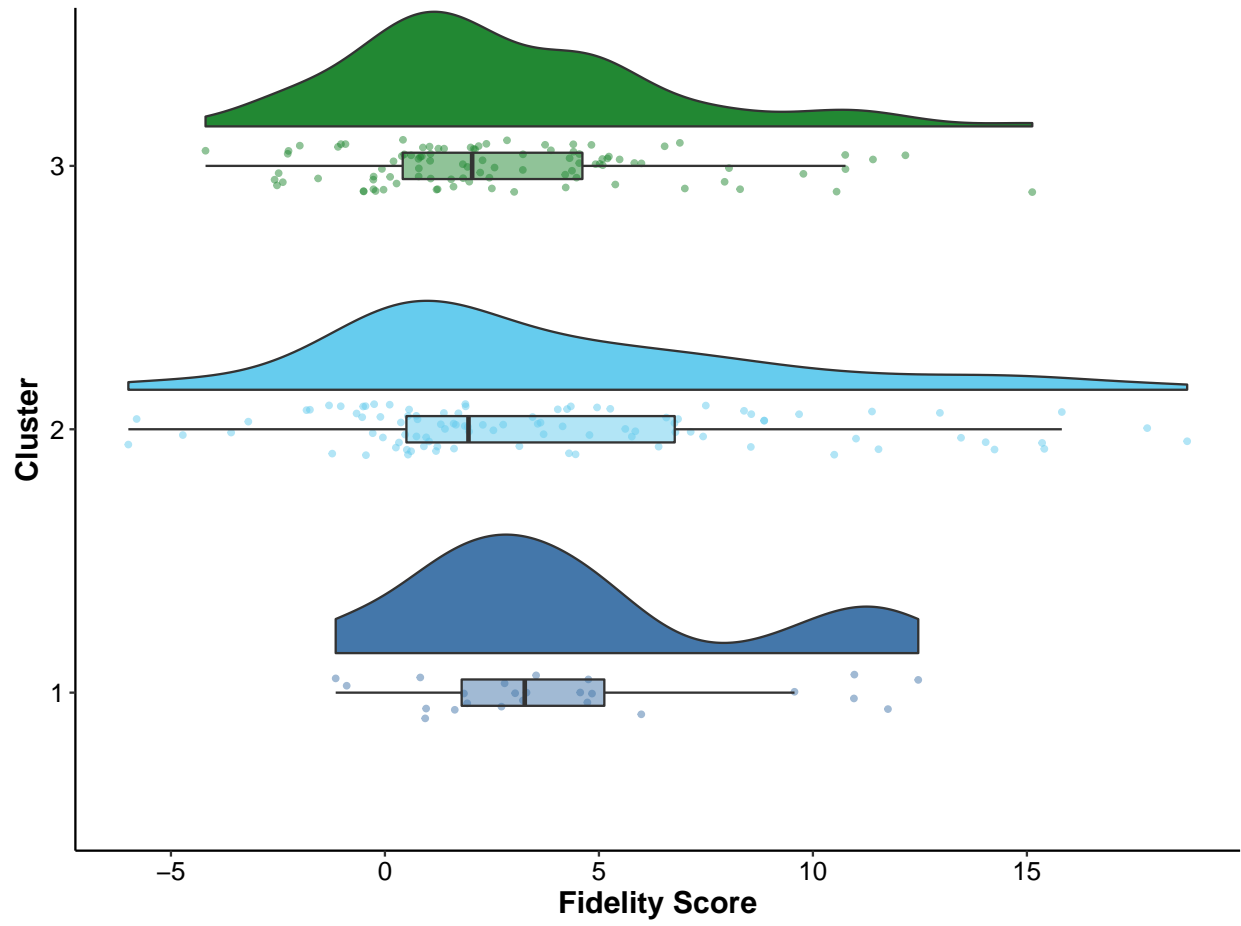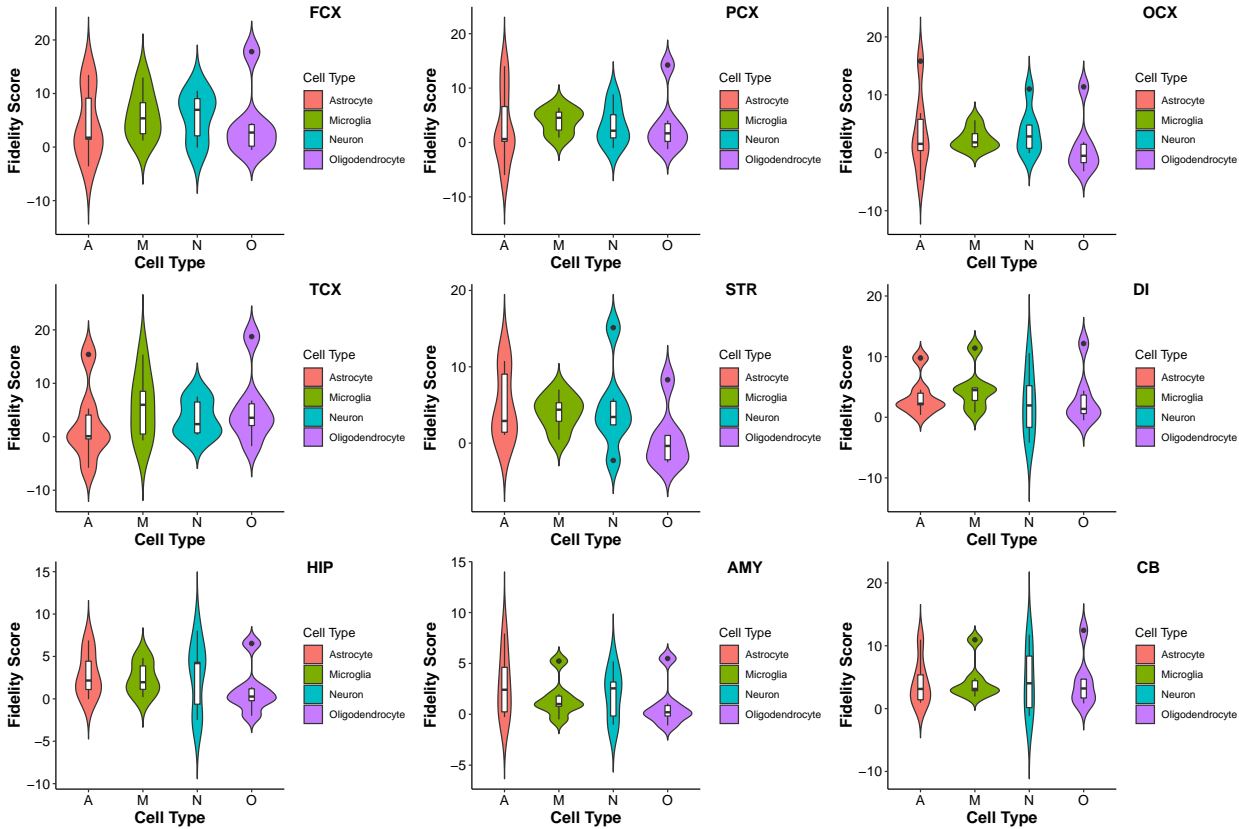
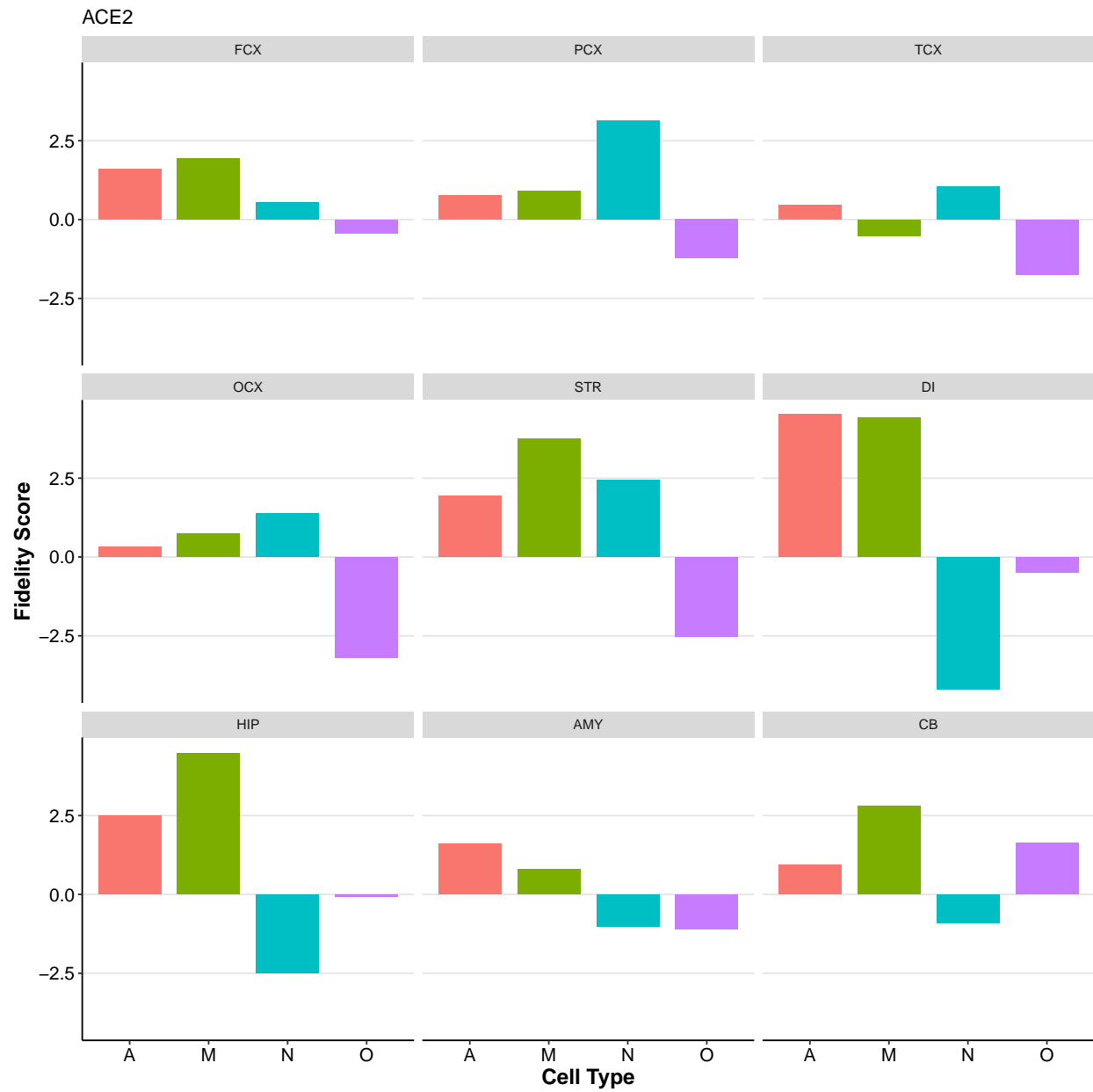# Violin Plot of Brain Region Fidelity Scores by Cell Type for Each Gene
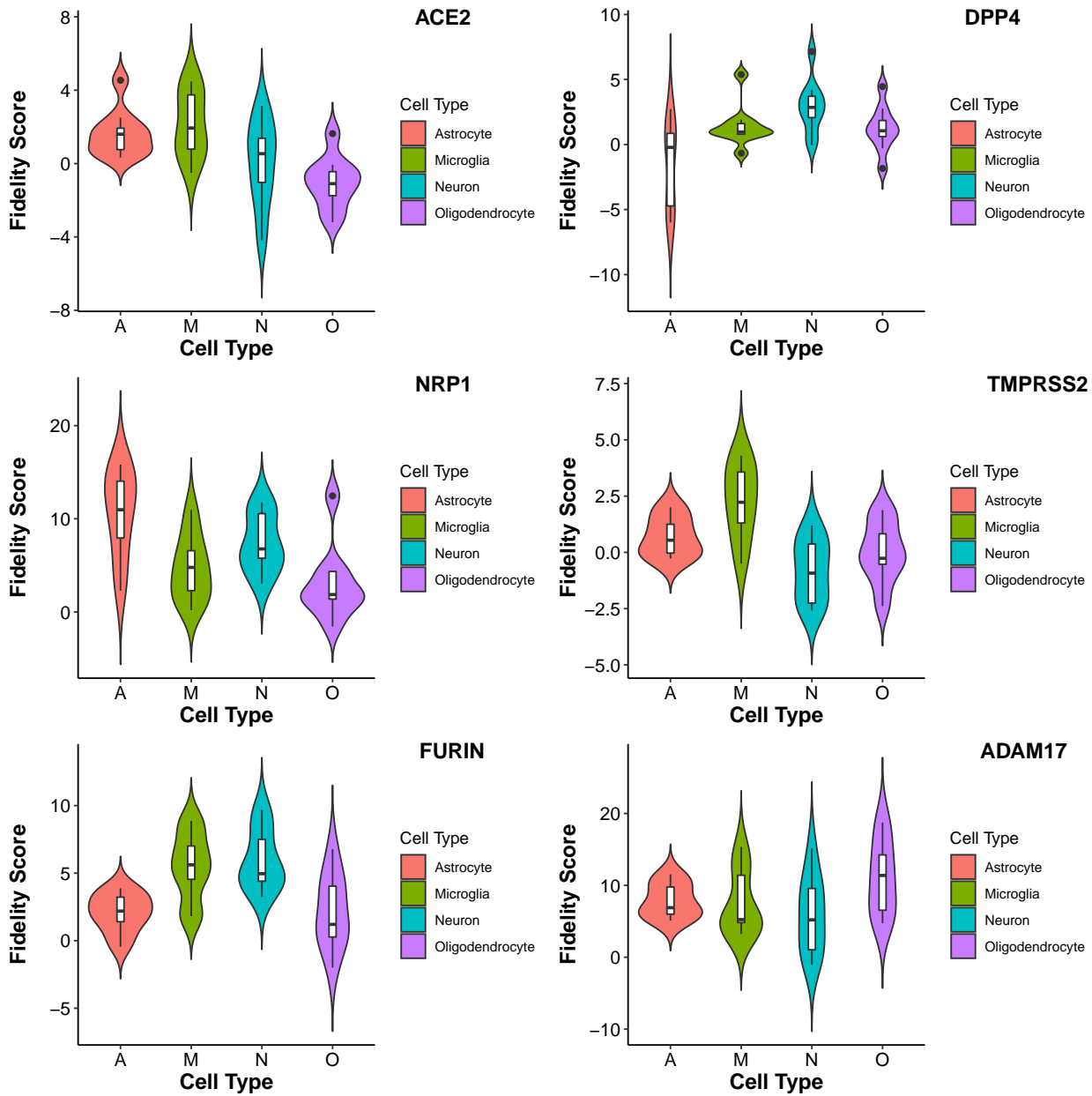
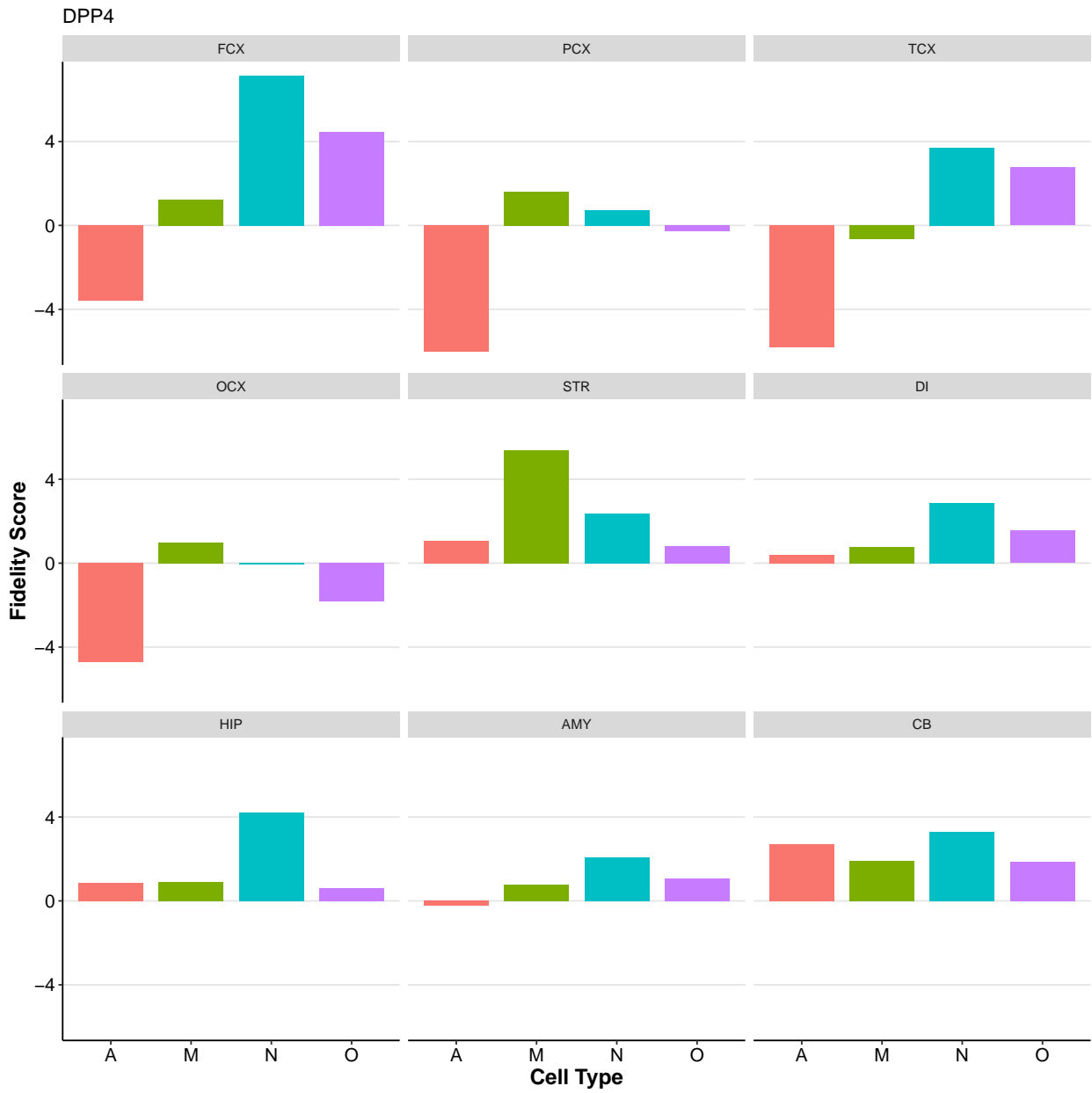# Violin Plot of Gene Fidelity Scores by Cell Type for Each Brain Region

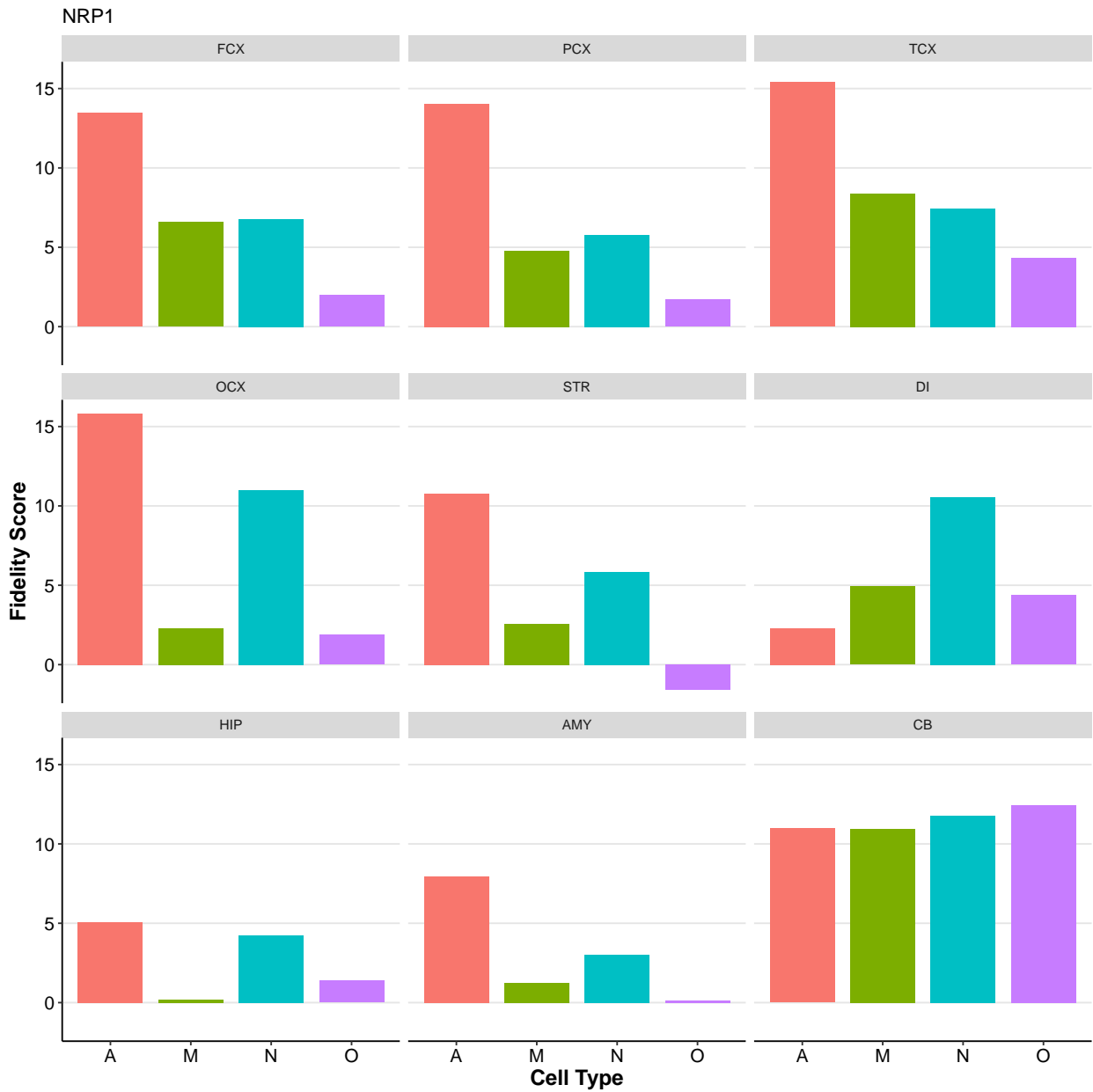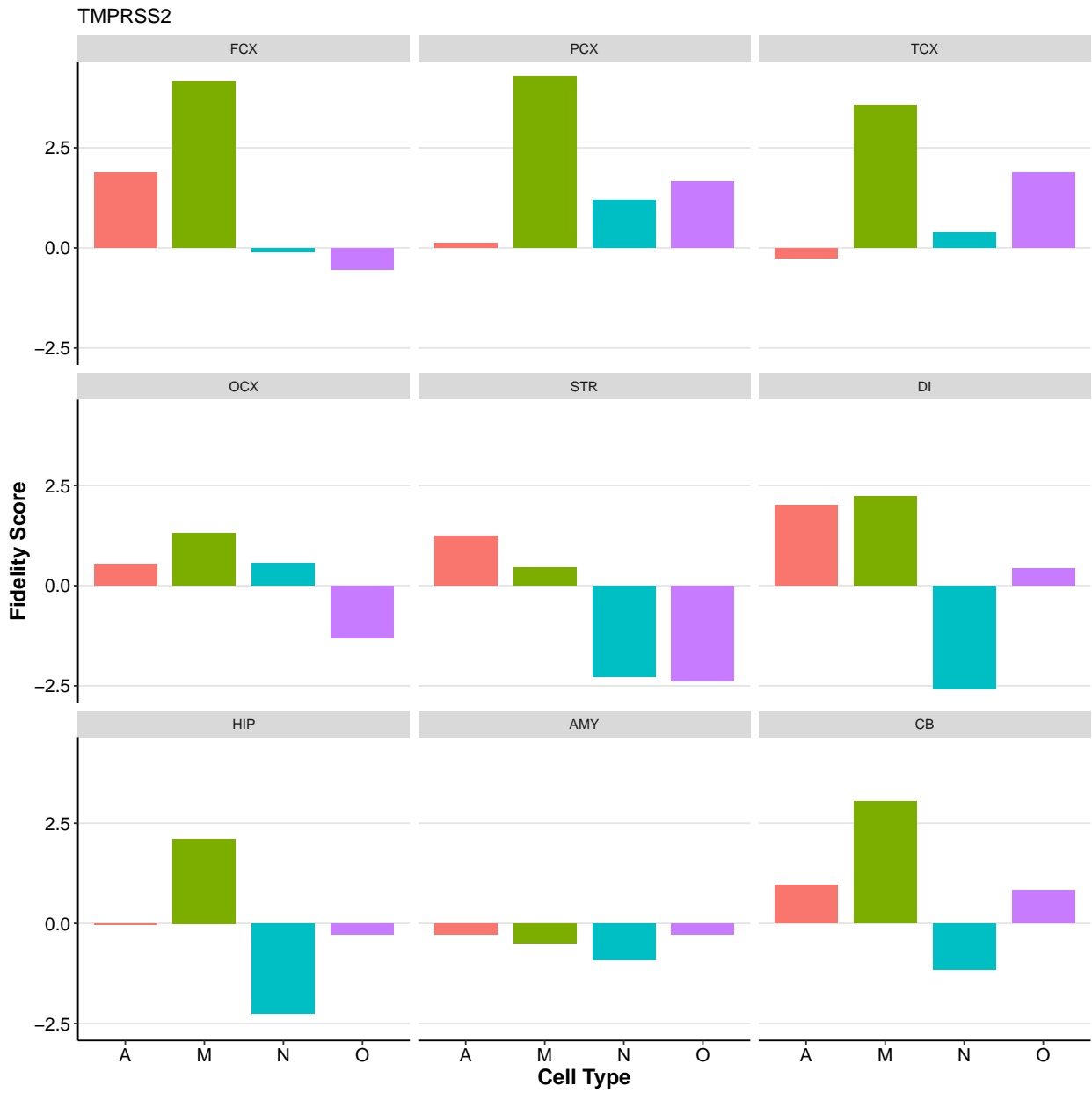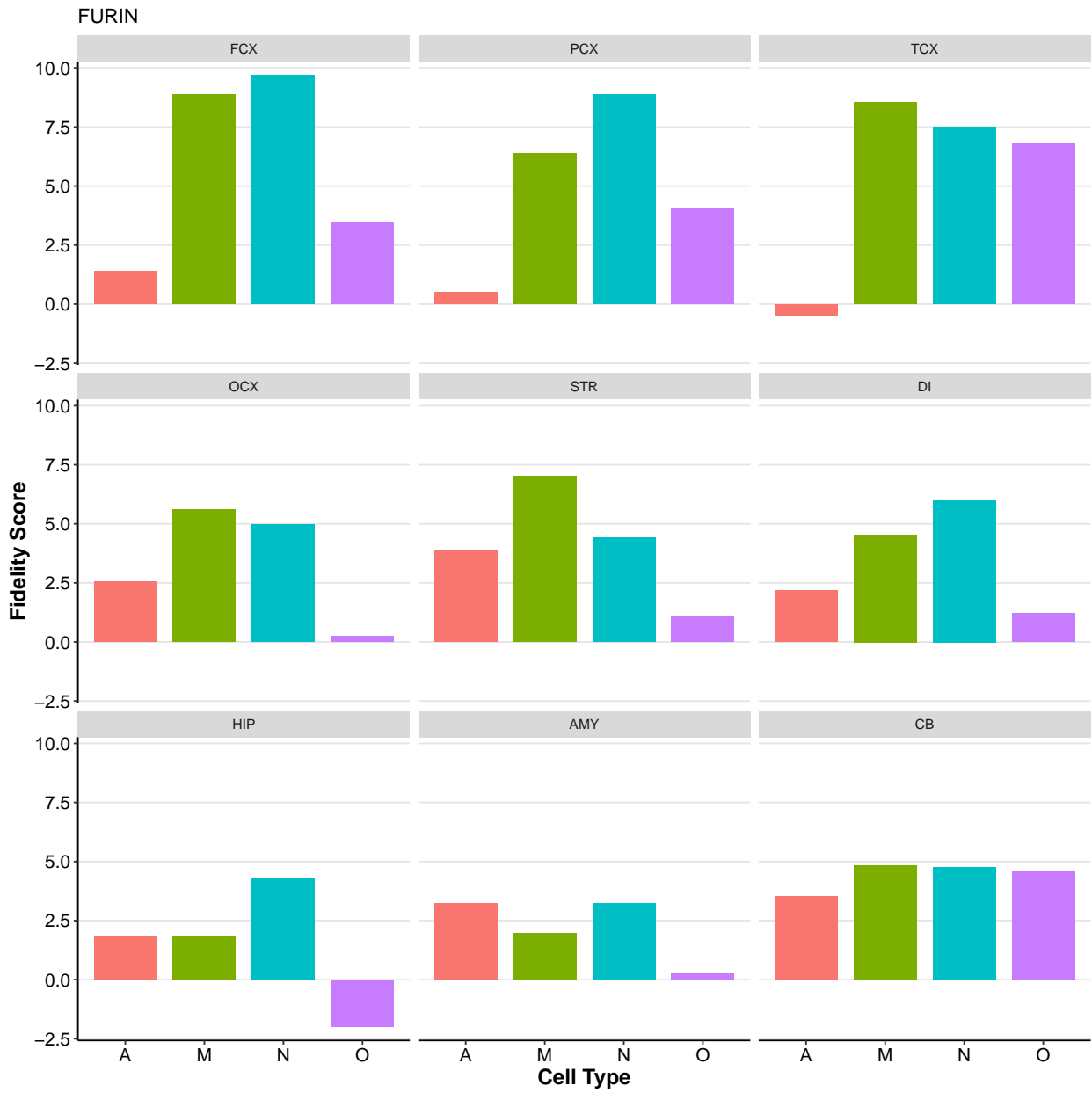**Bar Plots for Visualizing Cell Type Fidelity Scores of Each Gene by Brain Region**

ACE2

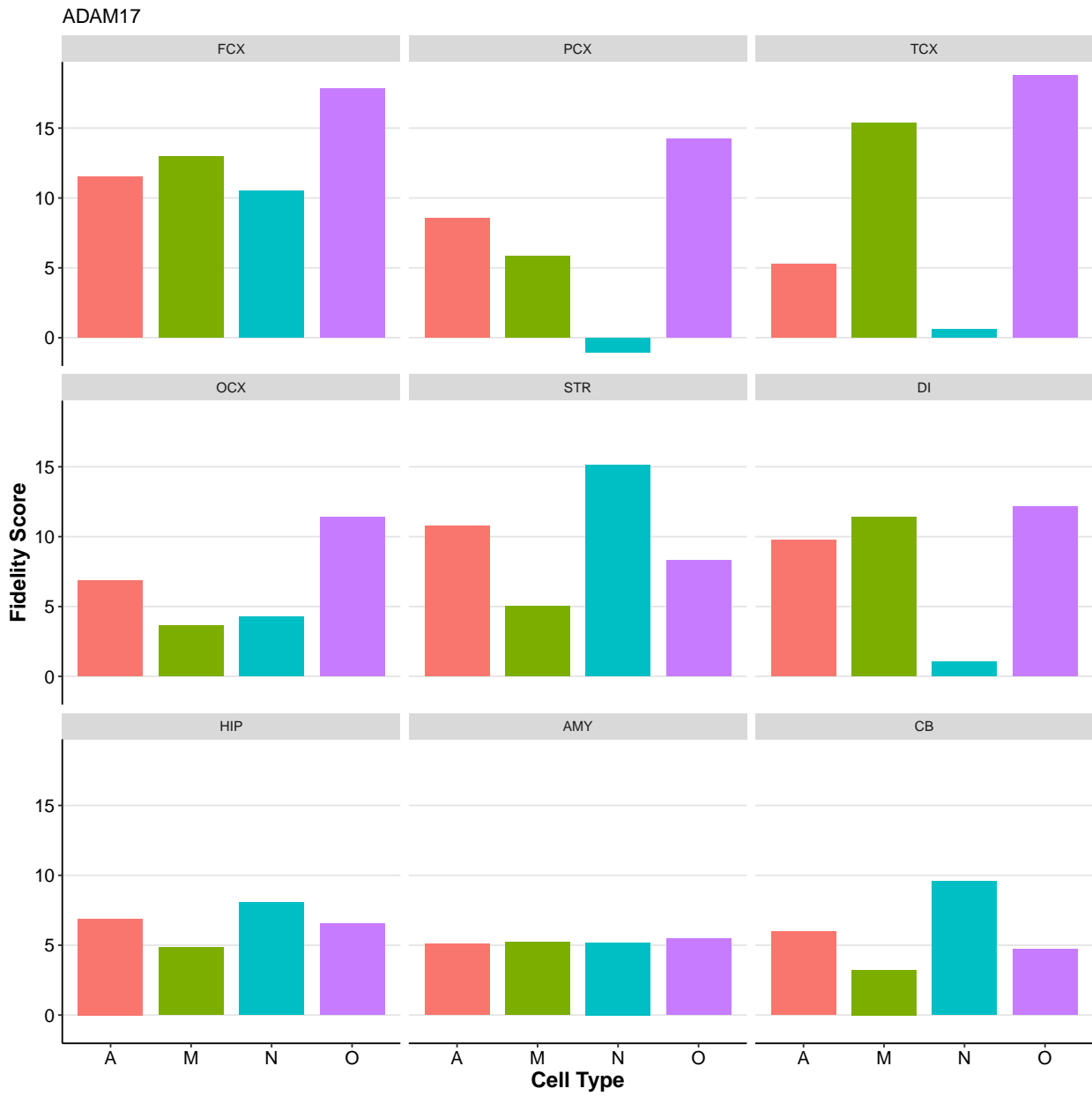ACE2

DPP4

DPP4

NRP1

TMPRSS2

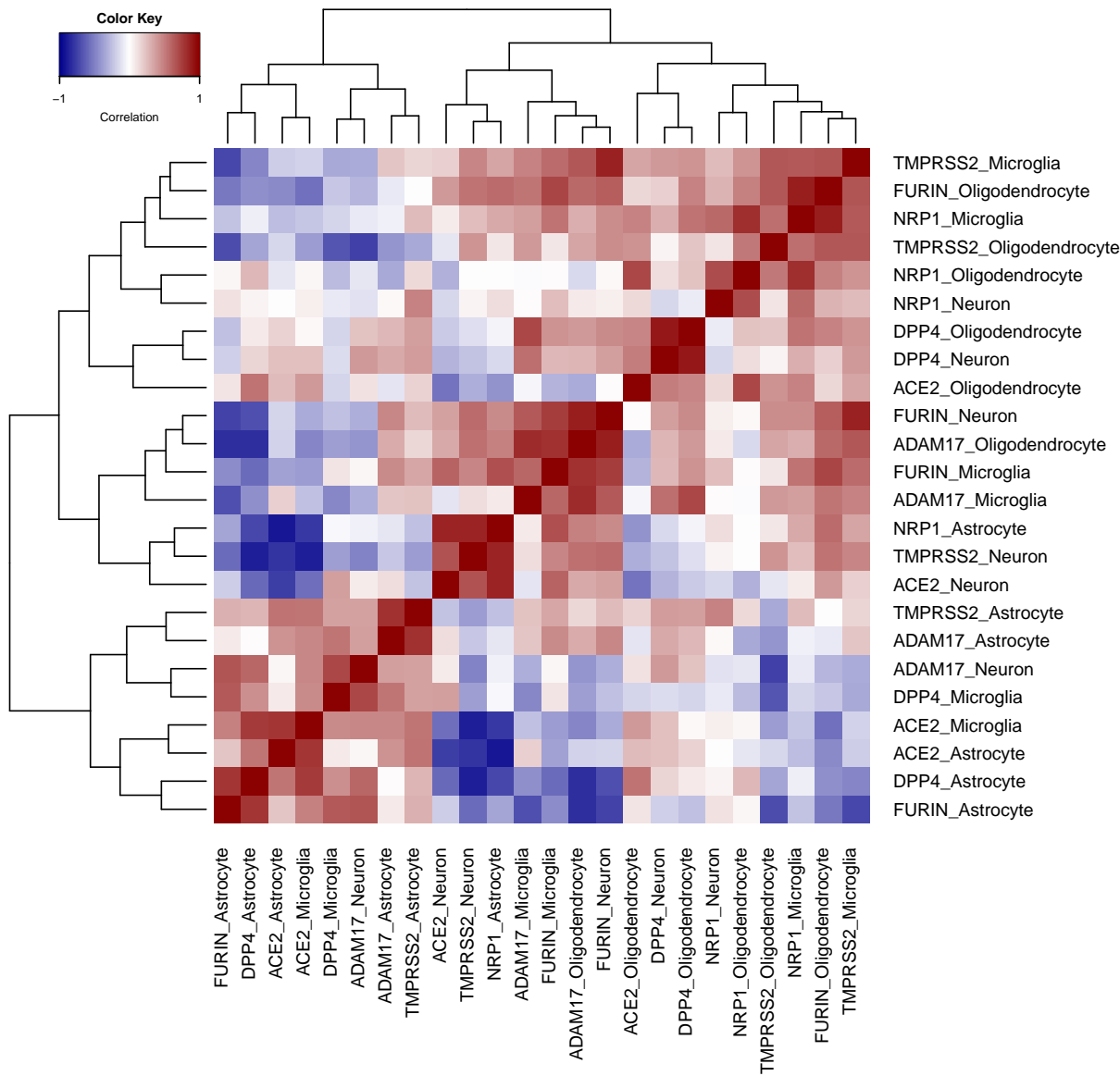

TMPRSS2

FURIN

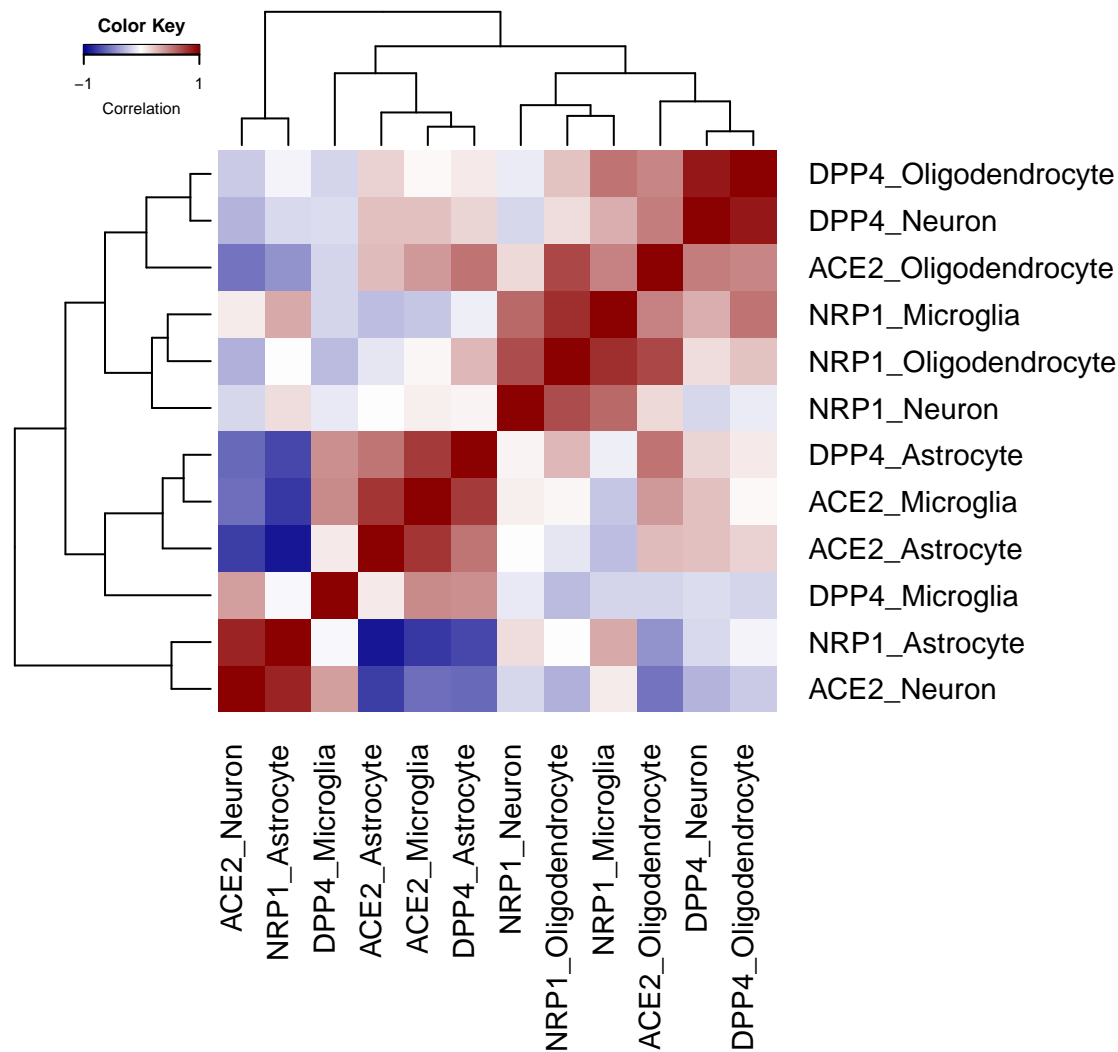

FURIN

ADAM17
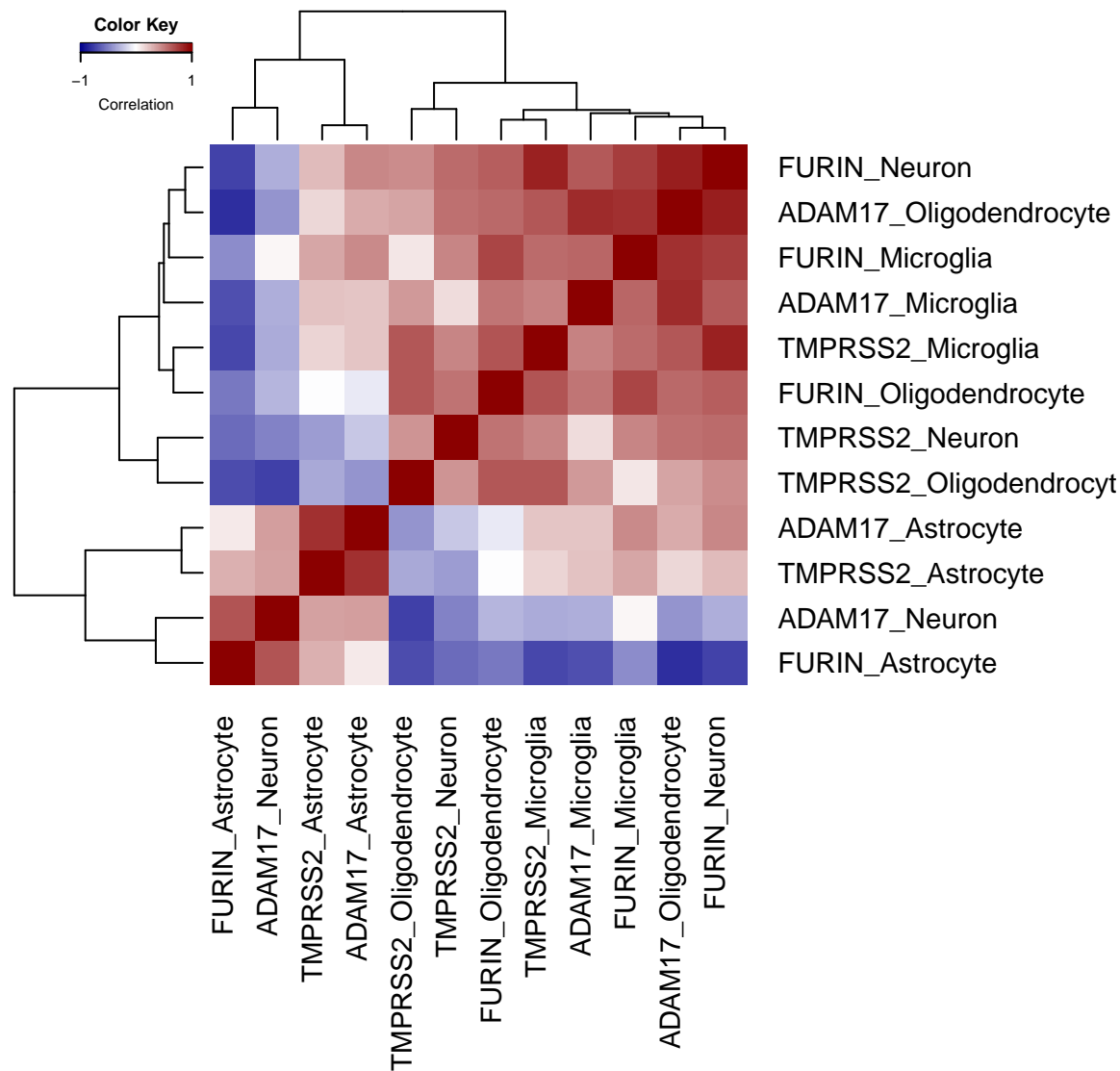

ADAM17

Correlation Matrices

All Gene and Cell Subtypes

Receptor Genes and Cell Subtypes for All Brain Regions

**Protease Genes and Cell Subtypes for All Brain Regions**

Visualizing Estimation Statistics

Cummings Estimation Plot

```
# Prepare data for generating estimation plots.
estim_plot_data <- dabest(fidelity_celltype, x = cluster_labels, y = Fidelity,
                          idx = list(c(1, 2, 3)),
                          paired = FALSE)

# Calculate mean differences for the estimation data.
cluster_mean_diff <- mean_diff(estim_plot_data)

# Create Cummings estimation plot.
estim_plot <- cluster_mean_diff %>% plot()
estim_plot
```
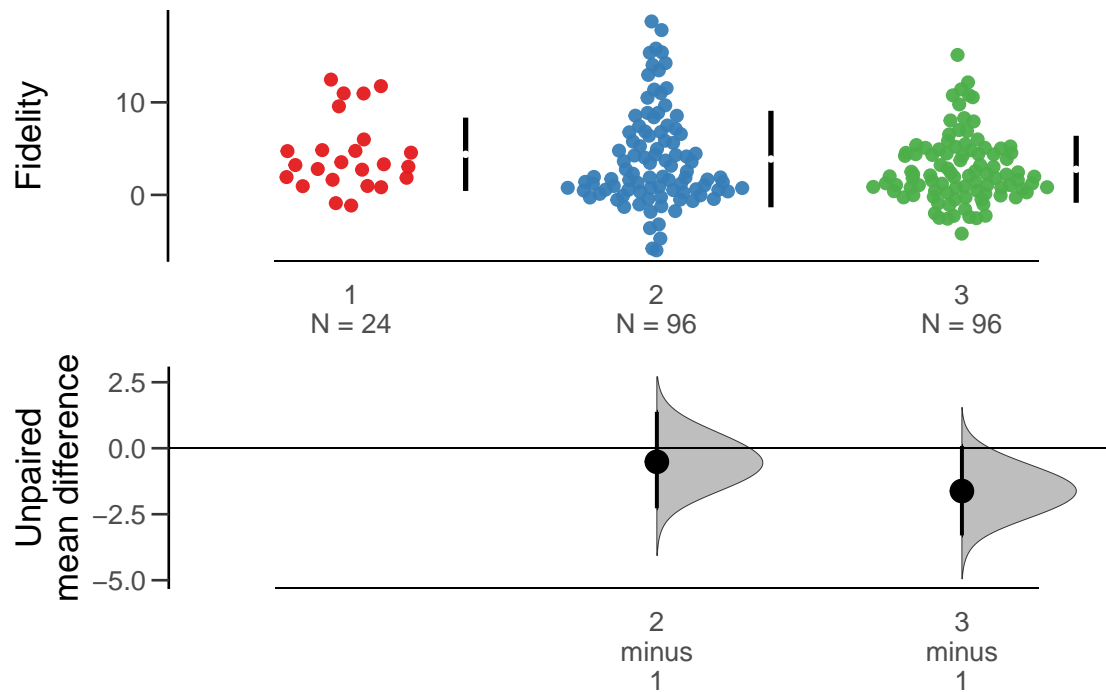
Session Information

```
## R version 4.0.5 (2021-03-31)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_CA.UTF-8/en_CA.UTF-8/en_CA.UTF-8/C/en_CA.UTF-8/en_CA.UTF-8
##
## attached base packages:
## [1] stats4    grid      stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] dabestr_0.3.0     see_0.6.4          ggdendro_0.1.22    dendextend_1.15.1
##  [5] gplots_3.1.1      viridis_0.6.1      viridisLite_0.4.0  RColorBrewer_1.1-2
##  [9] reshape2_1.4.4    factoextra_1.0.7   Rtsne_0.15         RSKC_2.4.2
## [13] flexclust_1.4-0   modeltools_0.2-23  lattice_0.20-44    ggpubr_0.4.0
## [17] magrittr_2.0.1    forcats_0.5.1      stringr_1.4.0      dplyr_1.0.7
## [21] purrr_0.3.4       readr_1.4.0        tidyr_1.1.3        tibble_3.1.2
## [25] ggplot2_3.3.5     tidyverse_1.3.1    here_1.0.1
##
## loaded via a namespace (and not attached):
##  [1] ggbeeswarm_0.6.0   colorspace_2.0-2   ggsignif_0.6.2     ellipsis_0.3.2
##  [5] class_7.3-19       rio_0.5.27         ggridges_0.5.3     rprojroot_2.0.2
##  [9] parameters_0.14.0  fs_1.5.0           rstudioapi_0.13    farver_2.1.0
## [13] ggrepel_0.9.1      fansi_0.5.0        lubridate_1.7.10   xml2_1.3.2
## [17] knitr_1.33         jsonlite_1.7.2     broom_0.7.8        dbplyr_2.1.1
## [21] effectsize_0.4.5   compiler_4.0.5     httr_1.4.2         backports_1.2.1
## [25] assertthat_0.2.1   cli_3.0.0          htmltools_0.5.1.1  tools_4.0.5
## [29] gtable_0.3.0       glue_1.4.2         Rcpp_1.0.6         carData_3.0-4
## [33] cellranger_1.1.0   vctrs_0.3.8        insight_0.14.2     xfun_0.24
## [37] openxlsx_4.2.4     rvest_1.0.0        lifecycle_1.0.0    gtools_3.9.2
## [41] rstatix_0.7.0      MASS_7.3-54        scales_1.1.1       hms_1.1.0
## [45] parallel_4.0.5     yaml_2.2.1         curl_4.3.2         gridExtra_2.3
## [49] stringi_1.6.2      highr_0.9          bayestestR_0.10.0  caTools_1.18.2
## [53] boot_1.3-28        zip_2.2.0          rlang_0.4.11       pkgconfig_2.0.3
## [57] bitops_1.0-7       evaluate_0.14      labeling_0.4.2     cowplot_1.1.1
## [61] tidyselect_1.1.1   plyr_1.8.6         R6_2.5.0           generics_0.1.0
## [65] DBI_1.1.1          pillar_1.6.1       haven_2.4.1        foreign_0.8-81
## [69] withr_2.4.2        abind_1.4-5        modelr_0.1.8       crayon_1.4.1
## [73] car_3.0-11         KernSmooth_2.23-20 utf8_1.2.1         rmarkdown_2.9
## [77] readxl_1.3.1       data.table_1.14.0  reprex_2.0.0       digest_0.6.27
## [81] munsell_0.5.0      beeswarm_0.4.0     vipor_0.4.5
```