Figure 1c, 1d, 1e — Discovery & Validation volcano plots and correlation

MAXOMOD_CSF

2026-02-16

Contents

Overview

This document reproduces **Figure 1c** (Discovery cohort volcano), **Figure 1d** (Validation cohort volcano), and **Figure 1e** (Discovery vs Validation signed FDR scatter) from the MAXOMOD CSF pipeline.

Inputs are read from **demo/Discovery** and **demo/Validation** (subfolder `03_Differential_expression_analysis`). Ensure the demo folder has been populated before running. Paths are relative to the **project root**. **Run the "Project root and parameters" chunk first** when running chunks interactively.

Upstream pipeline (run from project root):

- Discovery: `03_Differential_expression_analysis.R` → `04_Vis_Differential_expression_analysis.R`
- Validation: same for Validation cohort
- Scatter: `12_Scatterplot_FDR_0102.R` (uses both DE results)

Setup

Load packages

```
library(dplyr)
library(ggplot2)
library(ggrepel)
library(ggpubr)
library(data.table)
```

Project root and parameters

```
project_root <- if (basename(getwd()) == "vignettes") {
  normalizePath("..", winslash = "/")
} else {
  getwd()
}
setwd(project_root)
message("Project root: ", project_root)
```

```
## Project root: /Users/xliu2942/Documents/Projects/MAXOMOD/MAXOMOD_CSF
```

```
discovery_de_path <- file.path(project_root, params$discovery_de)
validation_de_path <- file.path(project_root, params$validation_de)
output_dir <- file.path(project_root, params$output)
set.seed(params$seed)
fdr_cutoff <- params$fdr_cutoff
cut_off <- -log10(fdr_cutoff)
# Figure e (scatter) uses a separate cutoff (default 0.2)
fdr_cutoff_scatter <- params$fdr_cutoff_scatter
cut_off_scatter <- -log10(fdr_cutoff_scatter)
```

```
if (!dir.exists(output_dir)) dir.create(output_dir, recursive = TRUE)
```

Volcano plot function

Same logic as `04_Vis_Differential_expression_analysis.R`: FDR threshold, up/down/ns coloring, top proteins labeled.

```
volcano_plot <- function(df, alpha_sig, name_title, labels, case_color, ctrl_color) {
  df <- df %>%
    mutate(omic_type = case_when(
      x >= 0 & y >= -log10(alpha_sig) ~ "up",
      x <= 0 & y >= -log10(alpha_sig) ~ "down",
```

```
      TRUE ~ "ns"
    ))
  cols <- c("up" = case_color, "down" = ctrl_color, "ns" = "grey")
  ggplot(data = df, aes(x, y)) +
    geom_point(aes(colour = omic_type), alpha = 0.5, shape = 16, size = 3) +
    geom_hline(yintercept = -log10(alpha_sig), linetype = "dashed") +
    geom_text_repel(
      data = filter(df, name %in% labels),
      aes(label = name),
      force = 1, nudge_x = -0.3, nudge_y = 1.5, direction = "both",
      max.overlaps = 20, size = 4
    ) +
    geom_vline(xintercept = 0, linetype = "dashed") +
    scale_colour_manual(values = cols) +
    labs(
      title = name_title,
      x = "log2(fold change)",
      y = expression(-log[10] ~ "(FDR)"),
      colour = "Differential \nExpression"
    ) +
    theme_classic() +
    theme(
      axis.title = element_text(size = 14),
      axis.text = element_text(size = 12),
      plot.title = element_text(size = 15, hjust = 0.5)
    ) +
    annotate("text", x = 1, y = 0.5,
             label = paste0(sum(df$omic_type == "up"), " more abundant\n",
                            sum(df$omic_type == "down"), " less abundant"))
}
```

Load differential expression results

We use the **all-patients** model: `norm_imp_MinProb_age_sex_cov_all_patients`.

```
if (!file.exists(discovery_de_path)) {
  stop("Discovery DE results not found: ", discovery_de_path,
       ". Populate demo/Discovery/03_Differential_expression_analysis/ or run
       ↪  03_Differential_expression_analysis.R for Discovery first.")
}
if (!file.exists(validation_de_path)) {
  stop("Validation DE results not found: ", validation_de_path,
       ". Populate demo/Validation/03_Differential_expression_analysis/ or run
       ↪  03_Differential_expression_analysis.R for Validation first.")
}

res_discovery   <- readRDS(discovery_de_path)
res_validation  <- readRDS(validation_de_path)

model_name <- "norm_imp_MinProb_age_sex_cov_all_patients"
```

```r
data_discovery  <- res_discovery[[model_name]]
data_validation <- res_validation[[model_name]]

# Fold change column (name contains "diff")
diff_col <- colnames(data_discovery)[grep("diff", colnames(data_discovery))][1]
```

Figure 1c — Discovery cohort volcano

```r
df_disc <- data.frame(
  x = data_discovery[[diff_col]],
  y = -log10(data_discovery$fdr),
  name = data_discovery$name
)
top10_up_disc   <- df_disc %>% filter(x >= 0, y >= cut_off) %>% arrange(desc(x)) %>%
↪  slice_head(n = 10)
top10_down_disc <- df_disc %>% filter(x <= 0, y >= cut_off) %>% arrange(x) %>%
↪  slice_head(n = 10)
labels_disc <- c(top10_up_disc$name, top10_down_disc$name)
```

```r
case_color <- "#D73027"
ctrl_color <- "#4575B4"
p_c <- volcano_plot(df_disc, fdr_cutoff,
                    "Discovery cohort",
                    labels_disc, case_color, ctrl_color)
print(p_c)
```

```r
ggsave(file.path(output_dir, "Fig1c_Discovery_volcano.pdf"), p_c, width = 7, height = 6,
↪  dpi = 300)
ggsave(file.path(output_dir, "Fig1c_Discovery_volcano.png"), p_c, width = 7, height = 6,
↪  dpi = 300)
```

Figure 1d — Validation cohort volcano

```r
df_val <- data.frame(
  x = data_validation[[diff_col]],
  y = -log10(data_validation$fdr),
  name = data_validation$name
)
top10_up_val   <- df_val %>% filter(x >= 0, y >= cut_off) %>% arrange(desc(x)) %>%
↪  slice_head(n = 10)
top10_down_val <- df_val %>% filter(x <= 0, y >= cut_off) %>% arrange(x) %>% slice_head(n
↪  = 10)
labels_val <- c(top10_up_val$name, top10_down_val$name)
```

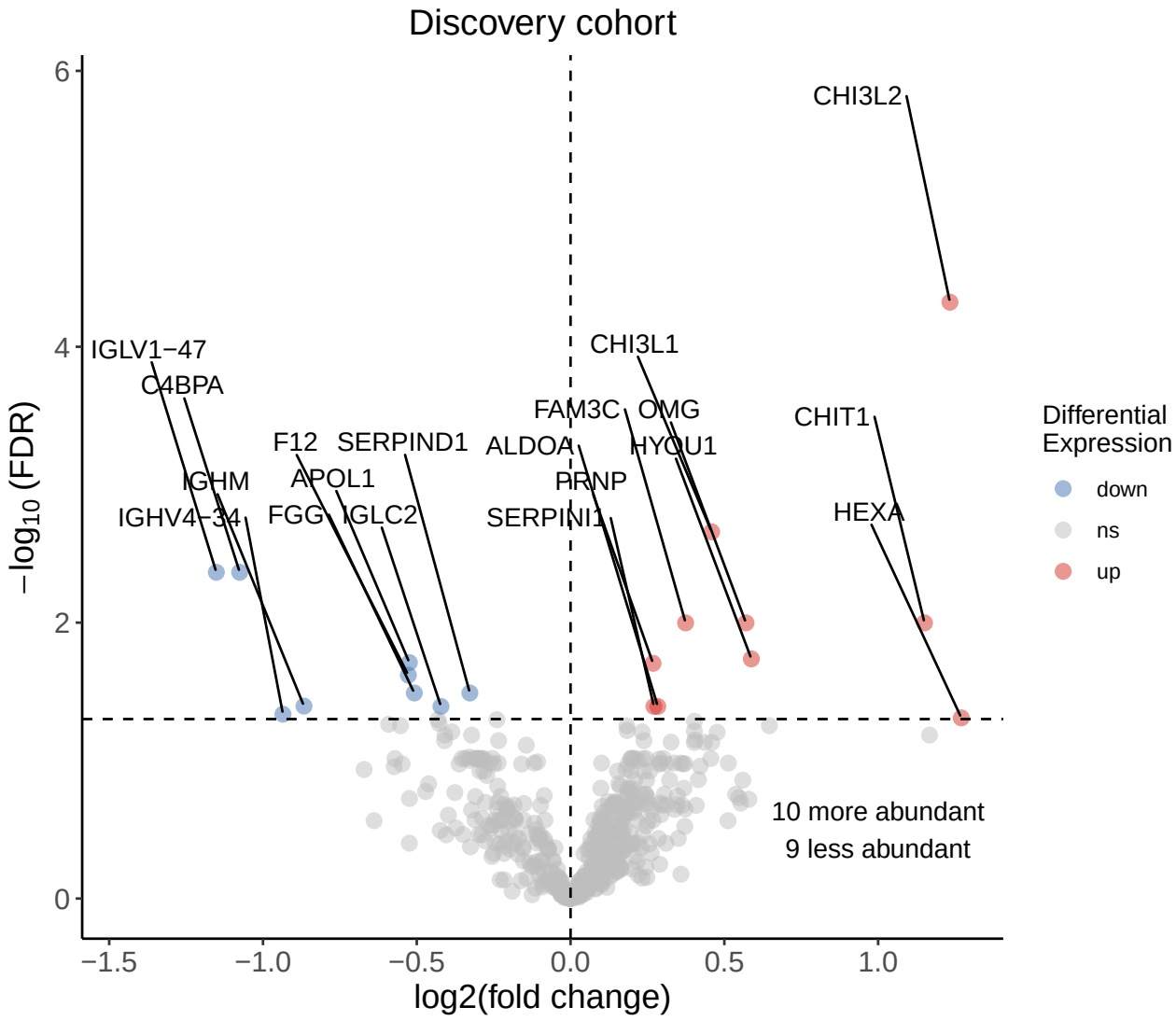

Figure 1: Figure 1c — Discovery cohort: proteins more/less abundant in ALS (FDR 0.05).

```
p_d <- volcano_plot(df_val, fdr_cutoff,
                    "Validation cohort",
                    labels_val, case_color, ctrl_color)
print(p_d)
```

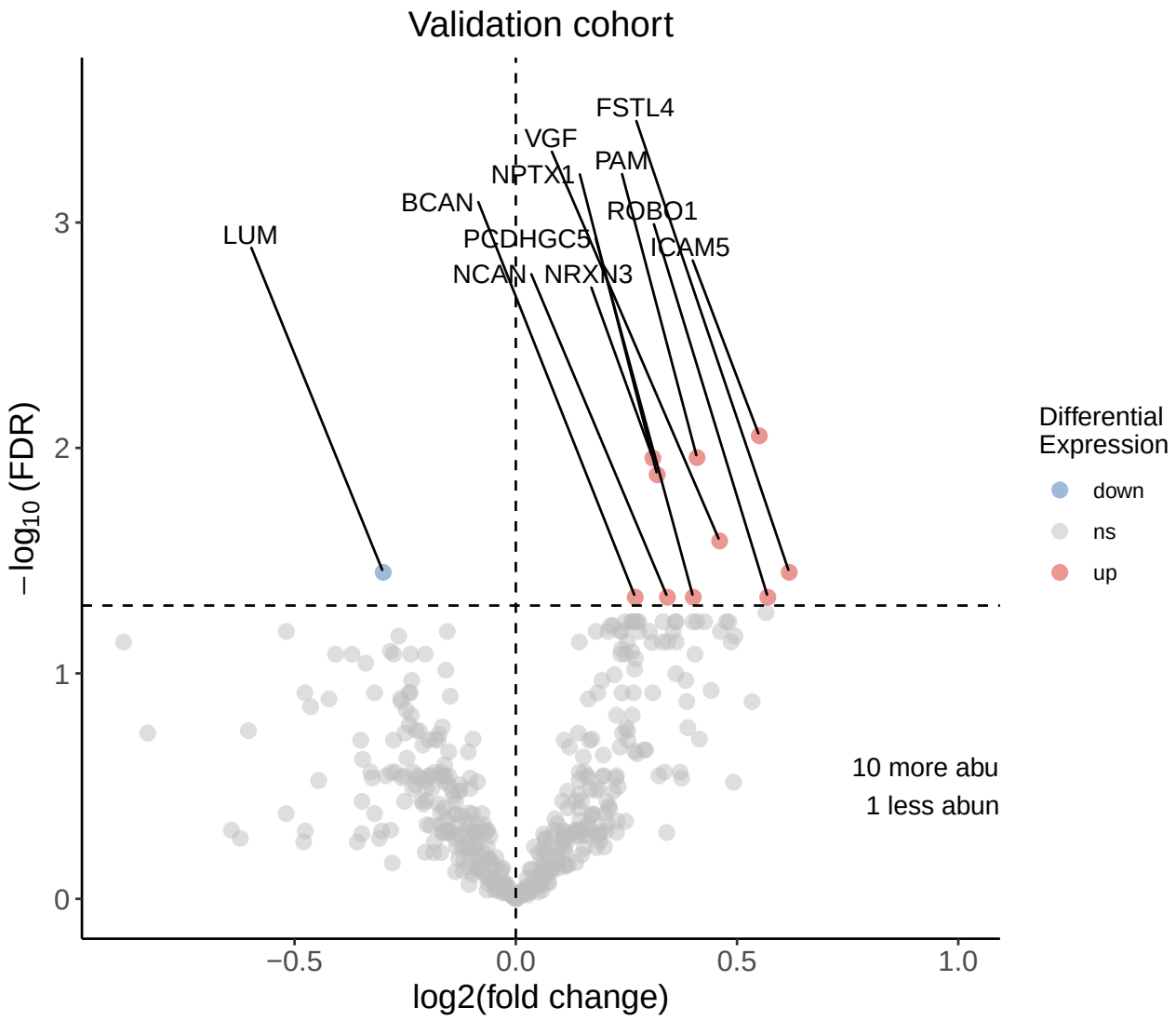

Figure 2: Figure 1d — Validation cohort: proteins more/less abundant in ALS (FDR 0.05).

```
ggsave(file.path(output_dir, "Fig1d_Validation_volcano.pdf"), p_d, width = 7, height = 6,
↪  dpi = 300)
ggsave(file.path(output_dir, "Fig1d_Validation_volcano.png"), p_d, width = 7, height = 6,
↪  dpi = 300)
```

Figure 1e — Discovery vs Validation scatter (signed -log10(FDR))

Same logic as `12_Scatterplot_FDR_0102.R`: x = signed -log10(FDR) Discovery, y = signed -log10(FDR) Validation; color by significant in Discovery only, Validation only, or both. Uses FDR cutoff **0.2** for this figure.

```
scatterplot_FDR_discovery_validation <- function(data, cut_off, q = 0.95, main_title,
↪  max.overlaps = Inf, labels = NULL) {
  data$omic_type <- "ns"
  data$omic_type[abs(data$y) >= cut_off] <- "significant in Validation"
  data$omic_type[abs(data$x) >= cut_off] <- "significant in Discovery"
  data$omic_type[(abs(data$x) >= cut_off) & (abs(data$y) >= cut_off)] <- "significant in
↪  both"
  cols <- c(
    "significant in Discovery" = "salmon",
    "significant in Validation" = "#26b3ff",
    "ns" = "grey",
    "significant in both" = "mediumpurple1"
  )
  if (!is.null(labels)) {
    label_data <- filter(data, name %in% labels)
  } else {
    quantile_y <- quantile(abs(data$y), na.rm = TRUE, probs = q)
    quantile_x <- quantile(abs(data$x), na.rm = TRUE, probs = q)
    label_data <- filter(data, abs(y) >= quantile_y | abs(x) >= quantile_x)
  }
  ggplot(data, aes(x, y)) +
    geom_point(aes(colour = omic_type), alpha = 0.5, shape = 16, size = 2) +
    geom_point(data = filter(data, abs(y) >= cut_off | abs(x) >= cut_off),
               aes(colour = omic_type), alpha = 0.5, shape = 16, size = 3) +
    geom_smooth(method = "lm", color = "#2C3E50", se = TRUE) +
    geom_hline(yintercept = c(cut_off, -cut_off), linetype = "dashed", colour = "grey40")
    ↪  +
    geom_vline(xintercept = c(cut_off, -cut_off), linetype = "dashed", colour = "grey40")
    ↪  +
    geom_hline(yintercept = 0, linetype = "dashed", colour = "grey80") +
    geom_vline(xintercept = 0, linetype = "dashed", colour = "grey80") +
    geom_text_repel(
      data = label_data,
      aes(label = name), force = 1, hjust = 1, max.overlaps = max.overlaps,
      segment.size = 0.2, min.segment.length = 0, size = 2
    ) +
    scale_colour_manual(values = cols) +
    labs(
      title = main_title,
      x = "signed -log10(FDR) for Discovery",
      y = "signed -log10(FDR) for Validation",
      colour = "Differential \nExpression"
    ) +
    theme_classic() +
    theme(
      axis.title = element_text(size = 14),
      axis.text = element_text(size = 12),
      plot.title = element_text(size = 15, hjust = 0.5)
```

```
    )
}
```

```
inter <- intersect(data_discovery$name, data_validation$name)
disc_inter <- data_discovery[match(inter, data_discovery$name), ]
val_inter  <- data_validation[match(inter, data_validation$name), ]

# Signed -log10(FDR): positive = more abundant in ALS
data_scatter <- data.table(
  name = disc_inter$name,
  x = -log10(disc_inter$fdr),
  y = -log10(val_inter$fdr)
)
data_scatter$x[disc_inter$als_vs_ctrl_diff < 0] <-
↪  -data_scatter$x[disc_inter$als_vs_ctrl_diff < 0]
data_scatter$y[val_inter$als_vs_ctrl_diff < 0]   <-
↪  -data_scatter$y[val_inter$als_vs_ctrl_diff < 0]
# Genes to label on scatter (user-specified)
genes_to_label_scatter <- c(
  "GC", "KNG1", "HRG", "C8G", "CFH", "IGL1", "C4BPA", "IGLC2", "APOH", "VTN",
  "C1S", "C6", "APOA2", "IGG1", "IGKV4-1", "CFB", "NPTX1", "VGF", "ATP6AP2", "BCAN",
  "NPTX2", "NPTXR", "APP", "CNTN1", "SEMA7A", "SCG2", "APLP2", "PTPRG", "FAM3C", "CHI3L1"
)
labels_scatter <- intersect(genes_to_label_scatter, data_scatter$name)
```

```
p_e <- scatterplot_FDR_discovery_validation(
  data_scatter, cut_off = cut_off_scatter, q = 0.95,
  main_title = paste0("ALS vs Control (FDR ", fdr_cutoff_scatter, ")"),
  max.overlaps = Inf, labels = labels_scatter
)
print(p_e)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
ggsave(file.path(output_dir, "Fig1e_Discovery_vs_Validation_scatter.pdf"), p_e, width =
↪  8, height = 6, dpi = 300)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
ggsave(file.path(output_dir, "Fig1e_Discovery_vs_Validation_scatter.png"), p_e, width =
↪  8, height = 6, dpi = 300)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
cor_test <- cor.test(data_scatter$x, data_scatter$y, method = "pearson")
message("Pearson correlation (signed -log10 FDR): r = ", round(cor_test$estimate, 4), ",
↪  p = ", format.pval(cor_test$p.value, digits = 3))
```

```
## Pearson correlation (signed -log10 FDR): r = 0.6085, p = <2e-16
```

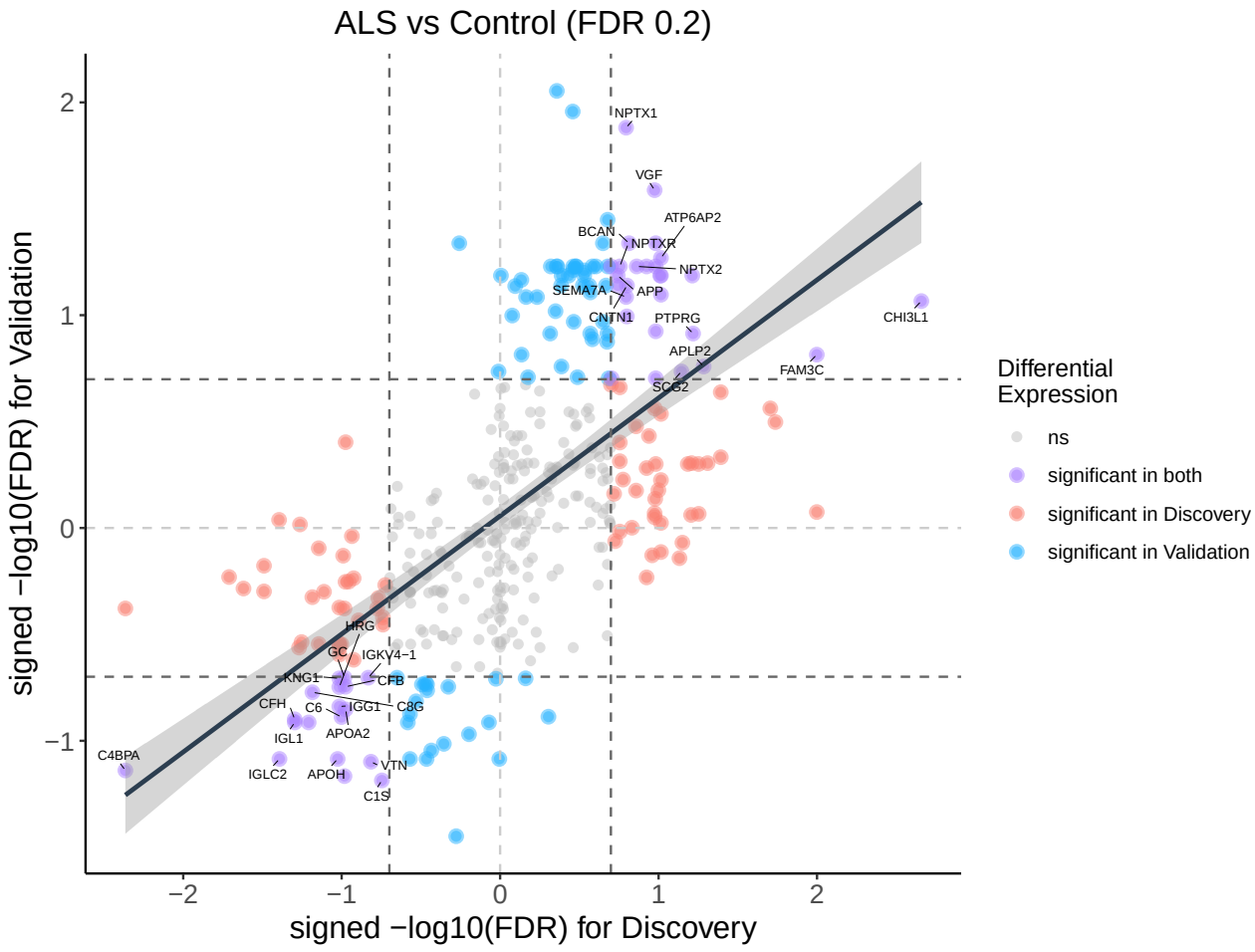

Figure 3: Figure 1e — Correlation of Discovery and Validation (signed -log10(FDR), cutoff 0.2).

Outputs

Figures are saved in **/Users/xliu2942/Documents/Projects/MAXOMOD/MAXOMOD_CSF/Plots/Fig1cde_vo**

- **Fig1c:** `Fig1c_Discovery_volcano.pdf` / `.png`
- **Fig1d:** `Fig1d_Validation_volcano.pdf` / `.png`
- **Fig1e:** `Fig1e_Discovery_vs_Validation_scatter.pdf` / `.png`

Ensure **demo/Discovery/03_Differential_expression_analysis/** and **demo/Validation/03_Differential_express** contain `Differential_Expression_Results.rds` before knitting (copy from pipeline output if needed).