

Figure 2d, 2e, 2f — Subclusters volcano and scatter (alpha vs beta)

MAXOMOD_CSF

2026-02-16

Contents

Overview	1
Setup	2
Load packages	2
Project root and parameters	2
Volcano plot function (alpha vs beta)	3
Load subclusters DE results	4
Figure 2d — Discovery subclusters volcano (alpha vs beta)	4
Figure 2e — Validation subclusters volcano (alpha vs beta)	6
Figure 2f — Discovery vs Validation scatter (alpha vs beta, signed FDR)	6
Outputs	10

Overview

This document reproduces **Figure 2d** (Discovery subclusters: alpha vs beta volcano), **Figure 2e** (Validation subclusters: alpha vs beta volcano), and **Figure 2f** (Discovery vs Validation signed FDR scatter for alpha vs beta) from the subclusters pipeline.

Equivalent commands (run from project root):

```
Rscript Script/04_Vis_Differential_expression_analysis_subclusters.R -i
↪ Discovery/03_Differential_expression_analysis_subclusters/res.rds -o
↪ Discovery/04_Vis_Differential_expression_analysis_subclusters -d
↪ Discovery/02_Missing_Inspection_subclusters/norm_imp_MinProb.rds -c
↪ Discovery/08_Clustering_als/cluster_assignments_2.csv -e 9
Rscript Script/04_Vis_Differential_expression_analysis_subclusters.R -i
↪ Validation/03_Differential_expression_analysis_subclusters/res.rds -o
↪ Validation/04_Vis_Differential_expression_analysis_subclusters -d
↪ Validation/02_Missing_Inspection_subclusters/norm_imp_MinProb.rds -c
↪ Validation/08_Clustering_als/cluster_assignments_2.csv -e 9
```

```
Rscript Script/12_Scatterplot_FDR_subclusters.R -i  
→ 03_Differential_expression_analysis_subclusters/res.rds -o  
→ 12_Scatterplot_FDR_subclusters
```

Inputs are read from `demo/Discovery/03_Differential_expression_analysis_subclusters/` and `demo/Validation/03_Differential_expression_analysis_subclusters/` (already copied). Paths are relative to the **project root**. Run the “Project root and parameters” chunk first when running chunks interactively.

Setup

Load packages

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

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_res_path <- file.path(project_root, params$discovery_res)  
validation_res_path <- file.path(project_root, params$validation_res)  
output_dir <- file.path(project_root, params$output)  
set.seed(params$seed)  
fdr_cutoff <- params$fdr_cutoff  
cut_off <- -log10(fdr_cutoff)  
fdr_cutoff_scatter <- params$fdr_cutoff_scatter  
cut_off_scatter <- -log10(fdr_cutoff_scatter)
```

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

Volcano plot function (alpha vs beta)

Same logic as 04_Vis_Differential_expression_analysis_subclusters.R: alpha (case) vs beta (ctrl), FDR threshold, top proteins labeled.

```
# Same as 04_Vis_Differential_expression_analysis_subclusters.R: Set2 for alpha/beta
clustering_colours <- brewer.pal(8, "Set2")[c(2, 3, 5)]
names(clustering_colours) <- c("alpha", "beta", "theta")
# Strip names so scale_colour_manual gets plain hex values (up = alpha, down = beta)
case_color <- as.character(clustering_colours["alpha"]) # more abundant in alpha
ctrl_color <- as.character(clustering_colours["beta"]) # more abundant in beta

volcano_plot <- function(df, alpha_sig, name_title, labels, max_overlaps = Inf) {
  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 = max_overlaps, size = 3.5, segment.size = 0.3
    ) +
    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 subclusters DE results

```
if (!file.exists(discovery_res_path)) {
  stop("Discovery res not found: ", discovery_res_path,
       ". Populate demo/Discovery/03_Differential_expression_analysis_subclusters/"
       ↪ first.")
}
if (!file.exists(validation_res_path)) {
  stop("Validation res not found: ", validation_res_path,
       ". Populate demo/Validation/03_Differential_expression_analysis_subclusters/"
       ↪ first.")
}

res_discovery <- readRDS(discovery_res_path)
res_validation <- readRDS(validation_res_path)
# Use k2 (alpha vs beta) for volcanos and scatter
k2_disc <- res_discovery$k2
k2_val <- res_validation$k2
```

Figure 2d — Discovery subclusters volcano (alpha vs beta)

```
diff_col <- grep("diff", colnames(k2_disc), value = TRUE)[1]
fdr_col <- grep("fdr", colnames(k2_disc), value = TRUE)[1]
df_disc <- data.frame(
  x = k2_disc[[diff_col]],
  y = -log10(k2_disc[[fdr_col]]),
  name = k2_disc$name
)
# Genes to label on Discovery volcano (user-specified; typos corrected: PARK7, ATRN,
↪ SERPINA4)
genes_to_label_disc <- c(
  "CNTN1", "PCSK1N", "PTPRZ1", "PTPRS", "NPTX1", "SCG2", "PARK7", "MCAM", "SCG3",
  "KNG1", "ORM2", "ATRN", "GC", "C8A", "APOH", "APOL1", "A1BG", "C8G", "SERPINA4"
)
labels_disc <- intersect(genes_to_label_disc, df_disc$name)
```

```
p_d <- volcano_plot(df_disc, fdr_cutoff, "Discovery (alpha vs beta)", labels_disc,
↪ max_overlaps = Inf)
print(p_d)
```

```
ggsave(file.path(output_dir, "Fig2d_Discovery_subclusters_volcano.pdf"), p_d, width = 7,
↪ height = 6, dpi = 300)
ggsave(file.path(output_dir, "Fig2d_Discovery_subclusters_volcano.png"), p_d, width = 7,
↪ height = 6, dpi = 300)
```

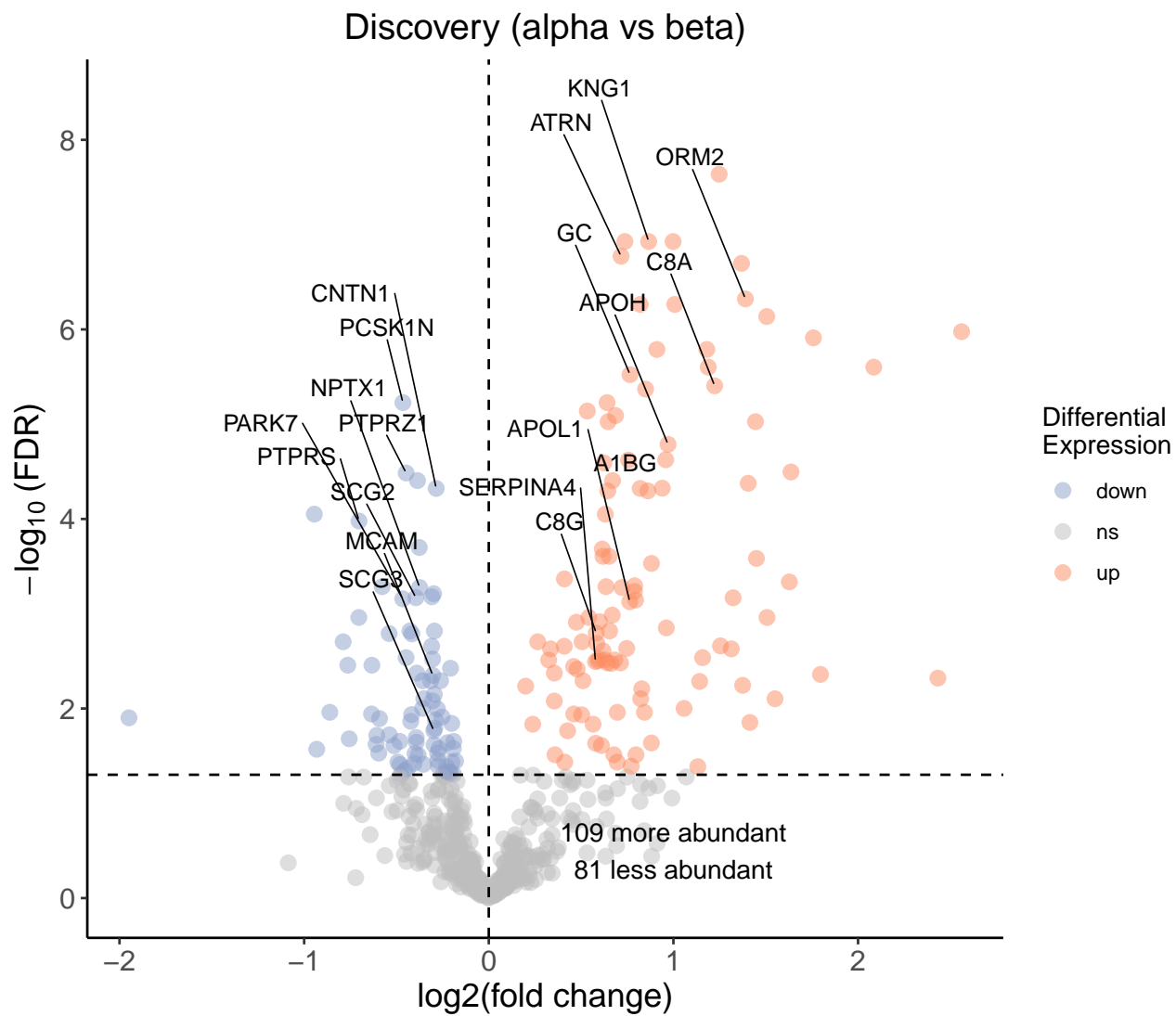


Figure 1: Figure 2d — Discovery subclusters: alpha vs beta (FDR 0.05).

Figure 2e — Validation subclusters volcano (alpha vs beta)

```
df_val <- data.frame(
  x = k2_val[[diff_col]],
  y = -log10(k2_val[[fdr_col]]),
  name = k2_val$name
)
# Genes to label on Validation volcano (user-specified)
genes_to_label_val <- c(
  "CNTN1", "APLP1", "SCG3", "PCSK1N", "PTPRZ1", "SCG2", "MCAM", "NPTX1",
  "A1BG", "KNG1", "GC", "C8A", "APOL1", "C8G", "APOH", "ATRN", "SERPINA4", "ORM2"
)
labels_val <- intersect(genes_to_label_val, df_val$name)

p_e <- volcano_plot(df_val, fdr_cutoff, "Validation (alpha vs beta)", labels_val,
  ↪ max_overlaps = Inf)
print(p_e)
```

```
ggsave(file.path(output_dir, "Fig2e_Validation_subclusters_volcano.pdf"), p_e, width = 7,
  ↪ height = 6, dpi = 300)
ggsave(file.path(output_dir, "Fig2e_Validation_subclusters_volcano.png"), p_e, width = 7,
  ↪ height = 6, dpi = 300)
```

Figure 2f — Discovery vs Validation scatter (alpha vs beta, signed FDR)

Same logic as 12_Scatterplot_FDR_subclusters.R: x = signed $-\log_{10}(\text{FDR})$ Discovery, y = signed $-\log_{10}(\text{FDR})$ Validation for alpha vs beta.

```
scatterplot_FDR <- 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)
```

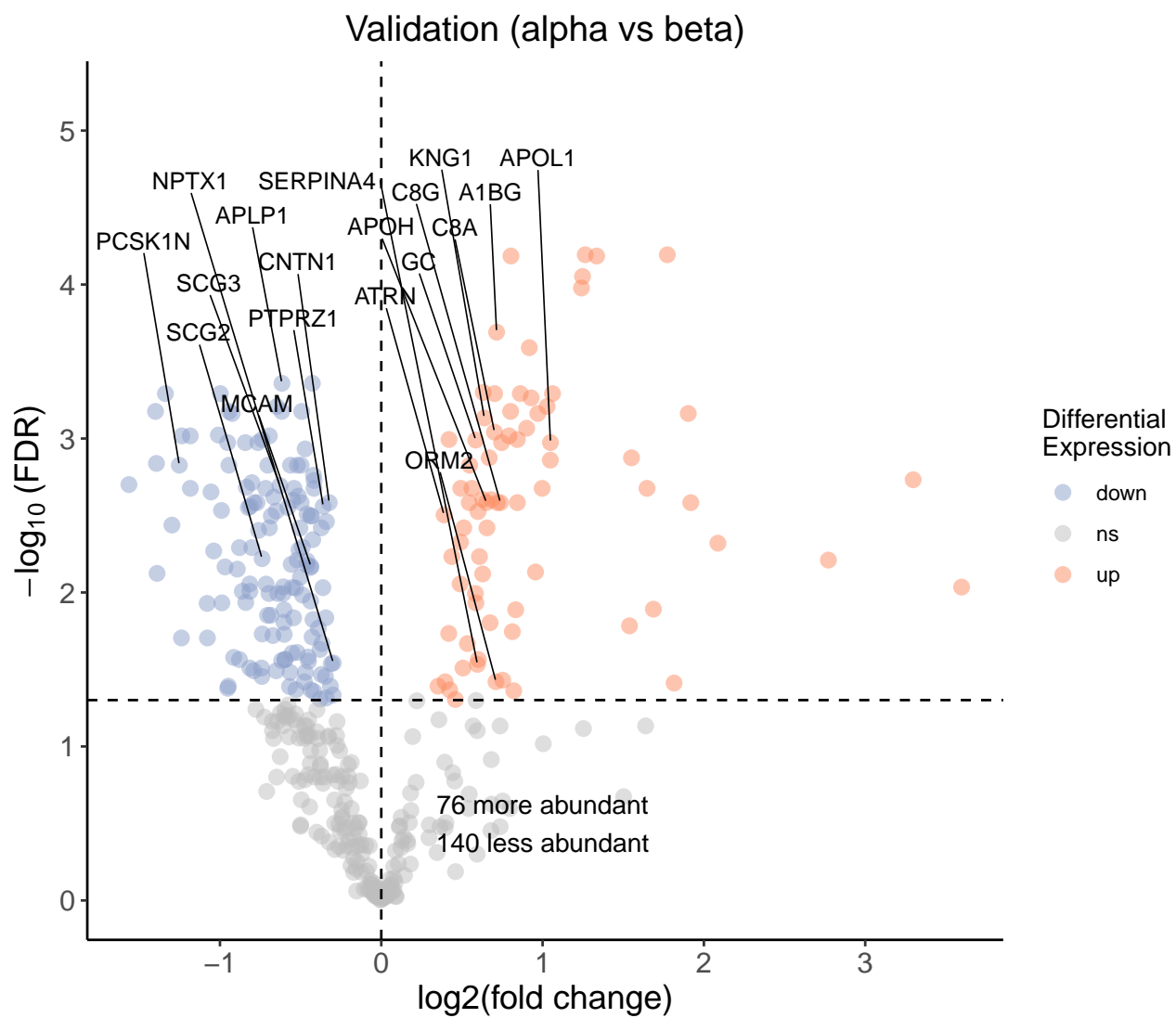


Figure 2: Figure 2e — Validation subclusters: alpha vs beta (FDR 0.05).

```

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(k2_disc$name, k2_val$name)
disc_inter <- k2_disc[match(inter, k2_disc$name), ]
val_inter <- k2_val[match(inter, k2_val$name), ]
diff_disc <- grep("diff", colnames(k2_disc), value = TRUE)[1]

data_scatter <- data.table(
  name = disc_inter$name,
  x = -log10(disc_inter[[fdr_col]]),
  y = -log10(val_inter[[fdr_col]])
)
data_scatter$x[disc_inter[[diff_disc]] < 0] <- -data_scatter$x[disc_inter[[diff_disc]] <
  ↪ 0]
data_scatter$y[val_inter[[diff_disc]] < 0] <- -data_scatter$y[val_inter[[diff_disc]] <
  ↪ 0]
# Genes to label on scatter (user-specified; typo corrected: SERPINA6)
genes_to_label_scatter <- c(
  "NRXN1", "NRXN2", "NPTX1", "SCG2", "PTPRZ1", "PCSK1N", "ATP6AP1", "CDH2", "APLP1",
  ↪ "MCAM",
  "PTPRG", "ICAM5", "NRCAM", "SCG3", "BASP1", "VGF", "NPTXR", "SERPINA6", "SERPINA4",
  ↪ "C8G",

```



```

"C3", "APOL1", "APOH", "GC", "C8A", "HPR", "A1BG", "FGA", "FGB", "KNG1", "ATRN", "ORM2"
)
labels_scatter <- intersect(genes_to_label_scatter, data_scatter$name)

```

```

p_f <- scatterplot_FDR(
  data_scatter, cut_off = cut_off_scatter, q = 0.95,
  main_title = paste0("Alpha vs beta (FDR ", fdr_cutoff_scatter, ")"),
  max.overlaps = Inf, labels = labels_scatter
)
print(p_f)

```

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

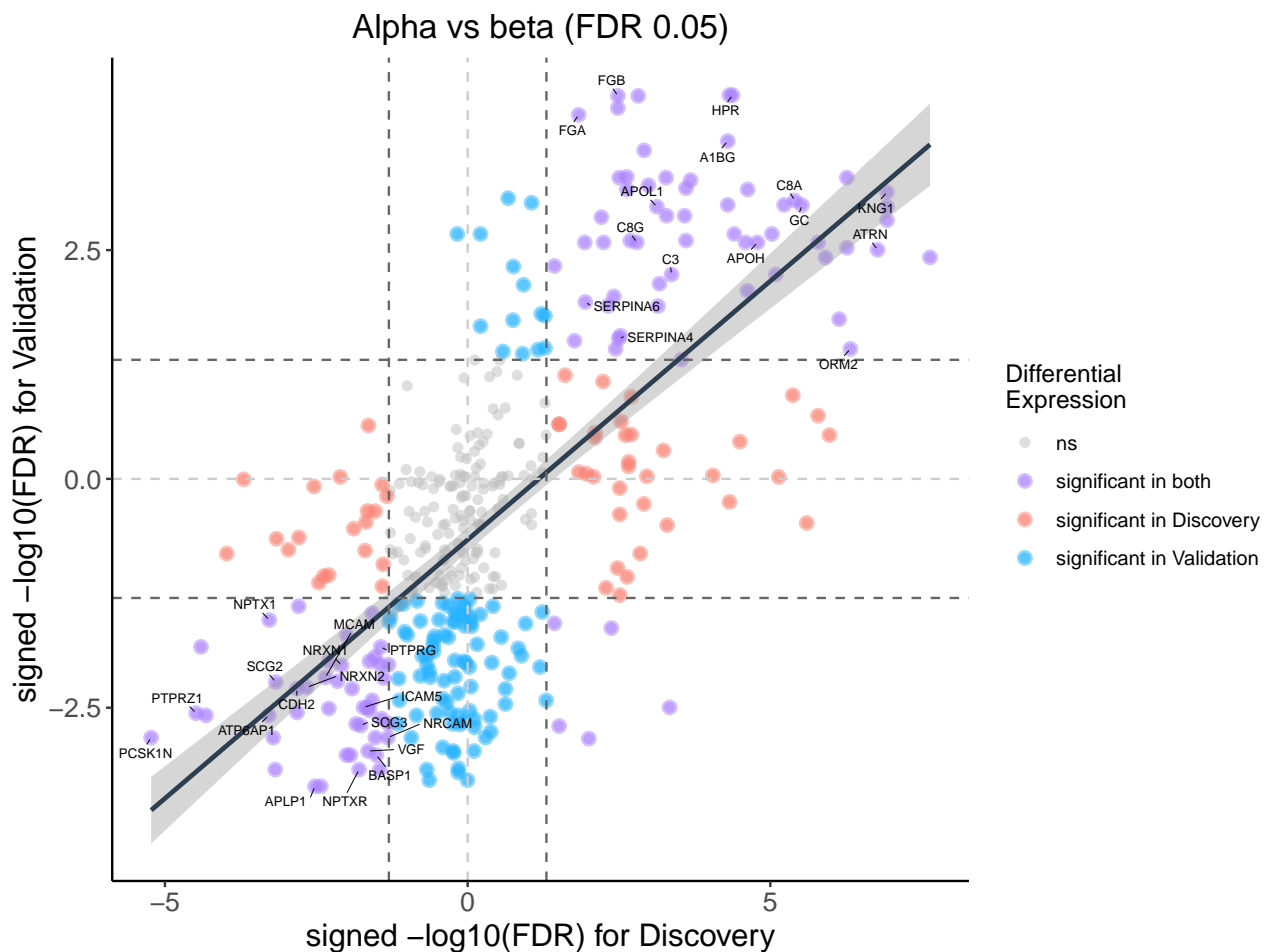


Figure 3: Figure 2f — Alpha vs beta: Discovery vs Validation (signed $-\log_{10}(\text{FDR})$).

```

ggsave(file.path(output_dir, "Fig2f_Discovery_vs_Validation_alpha_vs_beta_scatter.pdf"),
  p_f, width = 8, height = 6, dpi = 300)

```

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

```
ggsave(file.path(output_dir, "Fig2f_Discovery_vs_Validation_alpha_vs_beta_scatter.png"),
  ↪ p_f, 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, alpha vs beta): r = ",
  ↪ round(cor_test$estimate, 4), ", p = ", format.pval(cor_test$p.value, digits = 3))
```

```
## Pearson correlation (signed -log10 FDR, alpha vs beta): r = 0.6809, p = <2e-16
```

Outputs

- **Fig 2d:** Fig2d_Discovery_subclusters_volcano.pdf / .png in /Users/xliu2942/Documents/Projects/MAXO
- **Fig 2e:** Fig2e_Validation_subclusters_volcano.pdf / .png in /Users/xliu2942/Documents/Projects/MAXO
- **Fig 2f:** Fig2f_Discovery_vs_Validation_alpha_vs_beta_scatter.pdf / .png in /Users/xliu2942/Documents/

Ensure `demo/Discovery/03_Differential_expression_analysis_subclusters/res.rds` and `demo/Validation/03_Differential_expression_analysis_subclusters/res.rds` exist (already copied into demo folder).