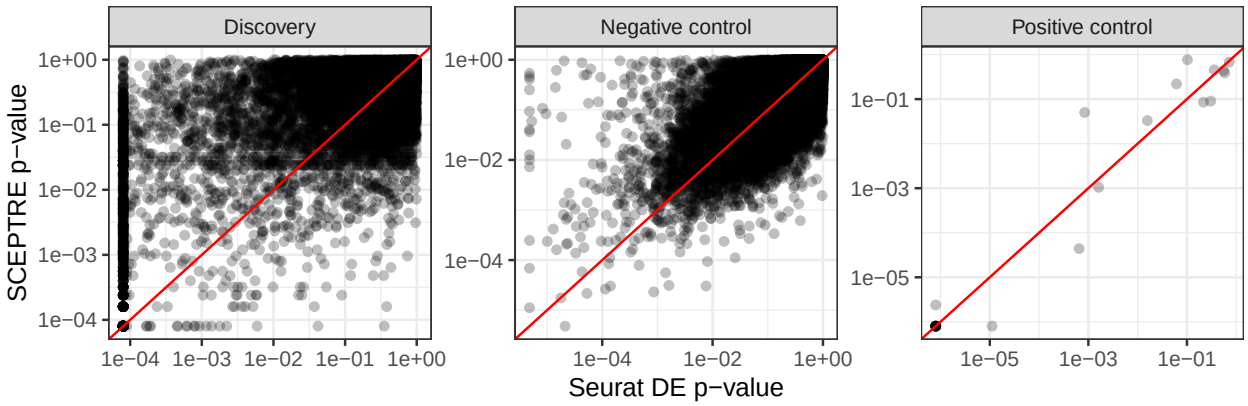
Comparing SCEPTRE and Seurat p-values on the Papalexi data

2023-03-10

If we are going to compare SCEPTRE and Seurat DE on the Papalexi discovery analysis, a first step is to get a sense for how these two sets of p-values compare to each other. To this end, I first restricted the Papalexi negative control, positive control, and CUL3 discovery pairs to those where the number of cells with treatment or negative control perturbations is at least 7. Then, I created scatter plots to compare the SCEPTRE and Seurat DE p-values. I plotted the p-values on a log-scale, truncating the Seurat DE p-values from below so they match the range of the SCEPTRE p-values. We find that the agreement on the positive and negative control pairs is very imperfect but still decent. On the other hand, on the CUL3 discovery pairs there seems to be very little agreement at all.



Next I applied a BH correction to the CUL3 discovery pairs at level $q = 0.1$. Below is the table of how many pairs are rejected based on the SCEPTRE and Seurat DE p-values. We find that 623 discoveries are common to both methods, whereas 1516 are unique to Seurat DE and 88 are unique to SCEPTRE. The divergence between the two methods is pretty large.

Table 1: Comparing number of BH rejections on CUL3 trans analysis between SCEPTRE and Seurat DE.

SCEPTRE	Seurat DE	Number of rejections
FALSE	FALSE	9288
FALSE	TRUE	1516
TRUE	FALSE	88
TRUE	TRUE	623

So the strong divergence between the SCEPTRE and Seurat DE CUL3 p-values found by Kaishu is not explained by contamination by pairs with low effective sample size. This divergence is somewhat suspicious, and perhaps should be investigated further.

Appendix: All code for this report

Read and process the results

```
library(tidyverse)
library(conflicted)
library(kableExtra)
conflicts_prefer(dplyr::filter)
# file paths
data_dir <- paste0(.get_config_path("LOCAL_CODE_DIR"),
                   "/sceptre2-manuscript/writeups/papalexi_analysis/")
result_dir <- paste0(.get_config_path("LOCAL_SCEPTRE2_DATA_DIR"), "results/")
sceptre_discovery_filename <- paste0(data_dir,
                    'sceptre_CUL3_and_PDL1_mrna_results_with_effect_size.rds')
seurat_CUL3_discovery_filename <- paste0(data_dir,'seurat_CUL3_results_no_filter.rds')
undercover_res_filename <- paste0(
  result_dir,
  "undercover_grna_analysis/undercover_result_grp_1_processed.rds"
)
pc_res_filename <- paste0(result_dir, "positive_control_analysis/pc_results_processed.rds")
sample_size_df_filename <- paste0(result_dir, "dataset_sample_sizes/n_nonzero_cells_per_grna.rds")
# undercover results
undercover_res <- readRDS(undercover_res_filename)

# discovery results (CUL3 trans analysis)
sceptre_res_discovery <- readRDS(sceptre_discovery_filename)
seurat_CUL3_res_discovery <- readRDS(seurat_CUL3_discovery_filename)

# positive control results
pc_res <- readRDS(pc_res_filename)

# get sample size information
sample_size_df <- readRDS(sample_size_df_filename)

# effective sample size thresholds
N_NONZERO_TREATMENT_CUTOFF <- 7
N_NONZERO_CONTROL_CUTOFF <- 7

### compute effective samples sizes for CUL3 analysis

# number of treatment cells with nonzero expression
num_trt_cells <- sample_size_df |>
  filter(dataset_concat == "papalexi/eccite_screen/gene",
         grna_id %in% c(paste0("CUL3g", 1:3))) |>
  group_by(feature_id) |>
  summarise(n_treatment = sum(n_nonzero_cells))

# number of control cells with nonzero expression
num_ctrl_cells <- sample_size_df |>
  filter(dataset_concat == "papalexi/eccite_screen/gene",
         grna_id %in% c(paste0("NTg", 1:10))) |>
  group_by(feature_id) |>
  summarise(n_control = sum(n_nonzero_cells))
```

```r
# joining the treatment and control cells with nonzeor expression
eff_sample_size_df <- inner_join(num_trt_cells, num_ctrl_cells, by = "feature_id")

### add effective sample size information to discovery data
cul3_results <- bind_rows(
  # sceptre results
  sceptre_res_discovery |>
    na.omit() |>
    unique() |>
    mutate(method = "sceptre") |>
    filter(grna_group == "CUL3") |>
    select(response_id, grna_group, p_value, method),
  # seurat results
  seurat_CUL3_res_discovery |>
    na.omit() |>
    rownames_to_column(var = "response_id") |>
    mutate(
      grna_group = "CUL3",
      method = "seurat_de"
    ) |>
    rename(p_value = p_val) |>
    select(response_id, grna_group, p_value, method)
) |>
  left_join(eff_sample_size_df |> rename(response_id = feature_id), by = "response_id") |>
  mutate(set = "Discovery", dataset = "papalexi_eccite_screen_gene") |>
  select(grna_group, response_id, method, dataset, p_value, n_treatment, n_control, set)

# join all of the results into a big data frame
results_joined <- bind_rows(
  # undercover results
  undercover_res |>
    filter(dataset == "papalexi_eccite_screen_gene") |>
    rename(
      grna_group = undercover_grna,
      n_treatment = n_nonzero_treatment,
      n_control = n_nonzero_control
    ) |>
    select(grna_group, response_id, method, dataset, p_value, n_treatment, n_control) |>
    mutate(set = "Negative control"),
  # positive control results
  pc_res |>
    filter(dataset == "papalexi_eccite_screen_gene") |>
    select(grna_group, response_id, method, dataset, p_value, n_treatment, n_control) |>
    mutate(set = "Positive control")
) |>
  # restrict attention to SCEPTRE and Seurat
  filter(
    method %in% c("sceptre", "seurat_de"),
  ) |>
  # CUL3 trans analysis results
  bind_rows(cul3_results) |>
  # apply pairwise QC thresholds
  filter(
```

```
    n_treatment >= N_NONZERO_TREATMENT_CUTOFF,
    n_control >= N_NONZERO_CONTROL_CUTOFF
  ) |>
  as_tibble()
```

Create the plot

```
results_joined |>
  pivot_wider(names_from = method, values_from = p_value) |>
  na.omit() |>
  group_by(set) |>
  mutate(seurat_de = pmax(seurat_de, min(sceptre))) |>
  ggplot(aes(x = seurat_de, y = sceptre)) +
  geom_point(alpha = 0.25) +
  geom_abline(color = "red") +
  facet_wrap(~set, scales = "free") +
  scale_x_log10() +
  scale_y_log10() +
  labs(x = "Seurat DE p-value",
       y = "SCEPTRE p-value")
```

Create the table

```
q <- 0.1
results_joined |>
  filter(set == "Discovery") |>
  group_by(method) |>
  mutate(discovered = p.adjust(p_value, "BH") <= q) |>
  ungroup() |>
  select(response_id, method, discovered) |>
  pivot_wider(names_from = method, values_from = discovered) |>
  na.omit() |>
  count(sceptre, seurat_de) |>
  kable(booktabs = TRUE,
        linesep = "",
        caption = "Comparing number of BH rejections on CUL3 trans analysis
        between SCEPTRE and Seurat DE.",
        col.names = c("SCEPTRE", "Seurat DE", "Number of rejections")) |>
  kable_styling(latex_options = "HOLD_position")
```