

Aim 1 Differential Discovery Analysis

1. Setup

```
# libraries
library(diffcyt)
library(tidyverse)
library(tidymodels)
library(lme4)

# source utils
source('~/.GitHub/classes/BIOMEDIN_212/r_scripts/pap_utils.R', echo = FALSE)

# paths
data_path <-
  file.path("~", "GitHub", "classes", "BIOMEDIN_212", "data-raw", "cell_table.csv")

metadata_path <-
  file.path(
    "~", "GitHub", "classes",
    "BIOMEDIN_212", "data-raw", "fov_labels.csv"
  )

surfactant_path <-
  file.path(
    "~", "GitHub", "classes",
    "BIOMEDIN_212", "data-raw", "surfactant_masks"
  )

# globals
healthy_fovs_in_pap_patients <-
  c(9, 15, 20, 29, 40)

other_markers <-
  c(
    "cd11c", "cd14",
    "cd16", "cd163", "cd20",
    "cd206", "cd209", "cd3",
    "cd31", "cd4", "cd45",
    "cd57", "cd68",
    "cd8", "calprotectin",
    "epox", "foxp3",
    "hh3", "hla_dr",
    "na_kat_pase", "pan_ck",
    "sma", "si",
    "tryptase", "vim",
    "i_nos"
```

```
)  
  
functional_markers <-  
  c(  
    "cd45ro", "grz_b", "h3k27me3", "h3k9ac", "ho_1",  
    "ido", "if_ng", "ki67", "mmp9", "pd1",  
    "p_s6", "tim3"  
  )
```

2. Read in data

```
# metadata
metadata <-
  metadata_path %>%
  read_csv() %>%
  rename(fov_id = point) %>%
  janitor::clean_names()

# mibi data
mibi_data <-
  data_path %>%
  read_csv() %>%
  rename(
    fov_id = point,
    cell_id = label,
    cluster_id = pixelfreq_hclust_cap,
    cluster_name = name,
    centroid_x = `centroid-0`,
    centroid_y = `centroid-1`
  ) %>%
  janitor::clean_names()

# surfactant data
surf_data <-
  tibble(
    filenames =
      surfactant_path %>%
      list.files(),
    paths =
      surfactant_path %>%
      list.files(full.names = TRUE),
    data = map(.x = paths, .f = pap_read_tif)
  )

surf_data <-
  surf_data %>%
  unnest(cols = data) %>%
  transmute(
    fov_id = str_extract(filenames, pattern = "[:digit:]+"),
    x,
    y,
    values
  )
```

3. Pre-process data

```
# join the mibi single-cell data with the FOV metadata
mibi_data <-
  mibi_data %>%
  left_join(metadata)
```

```
# count how many cells there are in each FOV type
mibi_data %>%
  count(category)
```

```
## # A tibble: 4 x 2
##   category      n
##   <chr>      <int>
## 1 nonSJIA-PAP 10584
## 2 Normal      4987
## 3 Pneumonia   6161
## 4 SJIA-PAP    57324
```

Because of our limited sample size, we more or less have to combine the “Normal” and “Pneumonia” category FOVs into a single category (“Control”). However, we should acknowledge that these two controls are not created equal - in fact, if we perform a simple t-test between the Pneumonia and Normal patient FOVs’ proportion of each our immune cell clusters, we can see that there are significant differences (at the level of $p = 0.05$) even after Benjamini-Hochberg adjustment.

```
cancer_pneumonia_counts <-
  mibi_data %>%
  filter(category %in% c("Pneumonia", "Normal")) %>%
  mutate(cluster_name = as.factor(cluster_name)) %>%
  count(patient_id, fov_id, category, cluster_name, .drop = FALSE) %>%
  group_by(fov_id) %>%
  mutate(prop = n / sum(n))
```

```
t_tests <-
  cancer_pneumonia_counts %>%
  group_by(cluster_name) %>%
  nest() %>%
  mutate(
    p_value =
      map_dbl(
        .x = data,
        .f = ~
          t.test(
            x =
              .x %>%
              dplyr::filter(patient_id == 13) %>%
              pull(prop),
            y =
              .x %>%
              dplyr::filter(patient_id == 14) %>%
              pull(prop)
          ) %>%
          tidy() %>%
          pull(p.value)
      ) %>%
```

```

    p.adjust(method = "BH")
  )

t_tests %>%
  select(-data) %>%
  arrange(p_value) %>%
  mutate(significant = if_else(p_value < 0.05, "*", "")) %>%
  knitr::kable()

```

cluster_name	p_value	significant
Mast_cell	0.0024518	*
Lung_Epithelium	0.0042509	*
CD57+_CD8+_Tcell	0.0129929	*
M2_Mac	0.0230855	*
Eosinophil	0.0341076	*
Endothelial	0.0561127	
CD8+_Tcell	0.0632200	
CD209+_Mac	0.0735467	
iNOS+_Pneumocyte	0.1045573	
CD16+_ImmuneOther	0.1383749	
Fibroblast	0.1780802	
CD11c+_mDC	0.2475057	
Neutrophil	0.2509420	
Bcell	0.2793292	
iNOS+_Mac	0.5925798	
CD57+_ImmuneOther	0.6663467	
CD4+_Tcell	0.7789749	
Mesenchymal	0.9513561	
Treg	0.9617596	
CD14+_Mono	0.9904512	

```

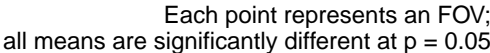
cluster_order <-
  t_tests %>%
  arrange(p_value) %>%
  pull(cluster_name) %>%
  as.character()

sig_clusters <-
  t_tests %>%
  filter(p_value < 0.05) %>%
  pull(cluster_name) %>%
  as.character()

cancer_pneumonia_counts %>%
  mutate(cluster_name = factor(cluster_name, levels = cluster_order)) %>%
  filter(cluster_name %in% sig_clusters) %>%
  ggplot(aes(y = prop, x = category, fill = category)) +
  geom_violin() +
  geom_jitter(shape = 21, width = 0.1) +
  facet_wrap(facets = vars(cluster_name), scales = "free") +
  labs(
    subtitle = "Differentially abundant clusters in our 2 control samples",

```

)



)

4. Basic data summary

Number of unique FOVs in each condition

```
mibi_data %>%
  distinct(outcome, fov_id) %>%
  count(outcome, name = "num_fovs") %>%
  arrange(-num_fovs)
```

```
## # A tibble: 3 x 2
##   outcome      num_fovs
##   <chr>         <int>
## 1 SJIA-PAP         48
## 2 Control          10
## 3 nonSJIA-PAP      9
```

Number of unique patients in each condition

```
mibi_data %>%
  distinct(outcome, patient_id) %>%
  count(outcome, name = "num_patients") %>%
  arrange(-num_patients)
```

```
## # A tibble: 3 x 2
##   outcome      num_patients
##   <chr>         <int>
## 1 SJIA-PAP         12
## 2 Control           2
## 3 nonSJIA-PAP      2
```

Number of unique cells in each condition

```
mibi_data %>%
  count(outcome, name = "num_cells") %>%
  arrange(-num_cells)
```

```
## # A tibble: 3 x 2
##   outcome      num_cells
##   <chr>         <int>
## 1 SJIA-PAP     57324
## 2 Control     11148
## 3 nonSJIA-PAP 10584
```

Number of cells for each patient

```
mibi_data %>%
  count(patient_id, name = "num_cells") %>%
  arrange(-num_cells)
```

```
## # A tibble: 16 x 2
##   patient_id num_cells
##   <dbl>         <int>
## 1         11      8710
## 2         16      7705
```

```
## 3      13      6161
## 4      10      5811
## 5       9      5294
## 6       8      5061
## 7      14      4987
## 8       4      4972
## 9      12      4935
## 10     2      4590
## 11     5      4140
## 12     7      4115
## 13     3      4025
## 14     6      3433
## 15    15      2879
## 16     1      2238
```

Number of cells in each FOV

```
mibi_data %>%
  count(fov_id, name = "num_cells") %>%
  arrange(-num_cells)
```

```
## # A tibble: 67 x 2
##   fov_id num_cells
##   <dbl>   <int>
## 1     14     2634
## 2     44     2341
## 3     12     2019
## 4     32     1924
## 5     43     1891
## 6      1     1885
## 7     25     1759
## 8     26     1754
## 9     49     1752
## 10    38     1726
## # ... with 57 more rows
```


5. Differential Abundance Analysis - between patients

In Aim 1, we proposed a differential abundance analysis of different immune cell subtypes (represented by the column `cluster_name` in `mibi_data`) across different types of MIBI images. The first of these analyses is to compare the abundance of each immune cell subtype between independent patients, each of which has either SJIA-PAP, PAP caused by something other than SJIA (nonSJIA-PAP), pneumonia, or lung cancer (which we code as “Normal” in `mibi_data`). We combine the last two conditions into the “control” category because neither of them have PAP, so they make as much sense as any sample we have access to to form our basis of comparison.

To perform our differential abundance analysis, we use the statistical framework proposed in the `{{diffcyt}}` framework.

Specifically, we use generalized linear mixed models (GLMMs) to test for differences in cluster abundance and cluster marker expression. The benefit of using mixed-models in this context is that, unlike more traditional differential abundance/expression testing tools commonly applied to cytometry data like `CITRUS`, GLMMs can account for complex experimental designs such as paired or block designs with known covariates representing batch effects or individual-to-individual variation. In the case of the present study, using random effects to model the variance in cluster abundance and marker expression that arises from random variation between individual patients (from whom we draw multiple FOVs), we can more reliably detect differences attributable solely to the effect of the `outcome` variable.

To do this, we can use the `{{diffcyt}}` R package to test for differential abundance of clusters across different levels of `outcome` using binomial regression. For each cluster, we can fit a binomial regression model in which we model the log-odds (and thus indirectly the proportion of cells in a given cluster) of each cluster in a given patient i and a given FOV j p_{ij} according to the following equation:

$$\text{logit}(p_{ij}) = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta_0 + \alpha_i + \beta_1 X_j$$

In the equation above, we use the following definitions:

- p_{ij} : The proportion of cells in a given cluster in patient i and FOV j
- α_i : A random intercept for each patient i in which $\alpha_i \sim N(0, \sigma)$, where σ is estimated during model fitting.
- X_j : an indicator variable representing whether or not an FOV j was taken from an SJIA-PAP patient (1 if yes, 0 otherwise). Depending on which comparisons we’re making, what X_j stands for can change (but it always represents which `outcome` FOV j has been annotated with).
- All β ’s are linear model parameters optimized during model fitting.

Using the above setup, we can apply null-hypothesis significance testing to β_1 (under the null hypothesis that $\beta_1 = 0$): if β_1 is significantly different from 0 in the model, we can state that the proportion of cells in our cluster differs significantly between the levels of `outcome` we’re investigating while controlling for individual-to-individual variation.

PAP vs. non-PAP

Using this framework, we can first compare the PAP samples (either SJIA-PAP or nonSJIA-PAP) to the control samples in our cohort. Note that we set a filter so that clusters that have fewer than 3 cells in 5 samples are removed from the analysis, as clusters with this few cells can’t be used to estimate reliable proportions for a cell subtype’s relative abundance in the sample it was collected from.

```
pap_daa <-
  mibi_data %>%
  select(fov_id, cluster_name, outcome, patient_id, any_of(functional_markers)) %>%
  mutate(outcome = if_else(outcome %in% c("SJIA-PAP", "nonSJIA-PAP"), "PAP", outcome)) %>%
  pap_perform_daa(
```

```

data_tibble = .,
sample_col = fov_id,
cluster_col = cluster_name,
fixed_effect_cols = outcome,
random_effect_cols = c(patient_id),
include_observation_level_random_effects = FALSE
)

pap_daa$da_results %>%
  topTable(all = TRUE) %>%
  as_tibble() %>%
  arrange(p_adj) %>%
  mutate(significance = if_else(p_adj < 0.05, "*", "")) %>%
  knitr::kable()

```

cluster_id	p_val	p_adj	significance
CD209+_Mac	0.0284889	0.2848894	
Neutrophil	0.0149365	0.2848894	
Bcell	0.3262681	0.7449216	
CD11c+_mDC	0.1519862	0.7449216	
CD14+_Mono	0.4472275	0.7449216	
CD57+_CD8+_Tcell	0.4812898	0.7449216	
CD8+_Tcell	0.2688716	0.7449216	
Endothelial	0.2369884	0.7449216	
Fibroblast	0.2432184	0.7449216	
iNOS+_Mac	0.1652710	0.7449216	
iNOS+_Pneumocyte	0.4402886	0.7449216	
Mesenchymal	0.4841990	0.7449216	
Treg	0.3633710	0.7449216	
CD16+_ImmuneOther	0.5616607	0.8023724	
CD4+_Tcell	0.6354191	0.8074845	
CD57+_ImmuneOther	0.6459876	0.8074845	
Eosinophil	0.8001735	0.9413806	
Lung_Epithelium	0.9541837	0.9541837	
M2_Mac	0.9149582	0.9541837	
Mast_cell	0.8741013	0.9541837	
Giant_cell	NA	NA	NA

From these results, we can see that, when taking individual random-effects into account, there are no statistically significant differentially abundant clusters between PAP and non-PAP samples (at least at the level of power we have available to us in this study).

SJIA-PAP vs. Controls

The second comparison we can run is between SJIA-PAP samples and control samples.

```

daa_sjia_pap_vs_controls <-
  mibi_data %>%
  select(fov_id, cluster_name, outcome, patient_id, any_of(functional_markers)) %>%
  filter(outcome != "nonSJIA-PAP") %>%
  pap_perform_daa(
    data_tibble = .,
    sample_col = fov_id,

```

```

cluster_col = cluster_name,
fixed_effect_cols = outcome,
random_effect_cols = c(patient_id),
include_observation_level_random_effects = FALSE
)

daa_sjia_pap_vs_controls$da_results %>%
  topTable(all = TRUE) %>%
  as_tibble() %>%
  arrange(p_adj) %>%
  mutate(significance = if_else(p_adj < 0.05, "*", "")) %>%
  knitr::kable()

```

cluster_id	p_val	p_adj	significance
Neutrophil	0.0000533	0.0010658	*
CD11c+_mDC	0.1296667	0.6439611	
CD14+_Mono	0.3199656	0.6439611	
CD209+_Mac	0.1100076	0.6439611	
CD8+_Tcell	0.2086889	0.6439611	
Endothelial	0.2859825	0.6439611	
Fibroblast	0.2608195	0.6439611	
iNOS+_Mac	0.1769249	0.6439611	
Mesenchymal	0.3219806	0.6439611	
Treg	0.2466686	0.6439611	
CD57+_CD8+_Tcell	0.3558237	0.6469522	
Bcell	0.4690123	0.7785946	
CD4+_Tcell	0.5060865	0.7785946	
CD16+_ImmuneOther	0.7383804	0.9488822	
CD57+_ImmuneOther	0.8849907	0.9488822	
Eosinophil	0.9014381	0.9488822	
Lung_Epithelium	0.6759856	0.9488822	
M2_Mac	0.8072448	0.9488822	
Mast_cell	0.7877623	0.9488822	
iNOS+_Pneumocyte	0.9523543	0.9523543	

In these results, we can see that neutrophils are differentially abundant in SJIA-PAP and control samples.

SJIA-PAP vs. nonSJIA-PAP

The third comparison we can run is between SJIA-PAP and nonSJIA-PAP samples.

```

daa_sjia_pap_vs_nonsjia_pap <-
  mibi_data %>%
  select(fov_id, cluster_name, outcome, patient_id, any_of(functional_markers)) %>%
  filter(outcome != "Control") %>%
  pap_perform_daa(
    data_tibble = .,
    sample_col = fov_id,
    cluster_col = cluster_name,
    fixed_effect_cols = outcome,
    random_effect_cols = c(patient_id),
    include_observation_level_random_effects = FALSE
  )

```

```

daa_sjia_pap_vs_nonsjia_pap$da_results %>%
  topTable(all = TRUE) %>%
  as_tibble() %>%
  arrange(p_adj) %>%
  mutate(significance = if_else(p_adj < 0.05, "*", "")) %>%
  knitr::kable()

```

cluster_id	p_val	p_adj	significance
CD209+_Mac	0.1262413	0.5950393	
CD4+_Tcell	0.0964778	0.5950393	
CD57+_ImmuneOther	0.0777235	0.5950393	
M2_Mac	0.1487598	0.5950393	
Mesenchymal	0.0569134	0.5950393	
Fibroblast	0.1843179	0.6143931	
CD16+_ImmuneOther	0.2473755	0.6184387	
Treg	0.2167738	0.6184387	
Neutrophil	0.3575358	0.7945239	
Bcell	0.4944349	0.8240582	
CD57+_CD8+_Tcell	0.4888039	0.8240582	
iNOS+_Pneumocyte	0.4535573	0.8240582	
CD11c+_mDC	0.6615094	0.8268868	
CD8+_Tcell	0.6364111	0.8268868	
Endothelial	0.5728799	0.8268868	
Eosinophil	0.6216371	0.8268868	
Lung_Epithelium	0.7370731	0.8671449	
CD14+_Mono	0.9862027	0.9869697	
iNOS+_Mac	0.8927271	0.9869697	
Mast_cell	0.9869697	0.9869697	
Giant_cell	NA	NA	NA

And once again we can see that there are no differentially abundant clusters between these sample types.

nonSJIA-PAP vs. Controls

The final between-patients comparison we can run is between nonSJIA-PAP samples and control samples.

```

daa_control_vs_nonsjia_pap <-
  mibi_data %>%
  select(fov_id, cluster_name, outcome, patient_id, any_of(functional_markers)) %>%
  filter(outcome != "SJIA-PAP") %>%
  pap_perform_daa(
    data_tibble = .,
    sample_col = fov_id,
    cluster_col = cluster_name,
    fixed_effect_cols = outcome,
    random_effect_cols = c(patient_id),
    include_observation_level_random_effects = FALSE
  )

daa_control_vs_nonsjia_pap$da_results %>%
  topTable(all = TRUE) %>%
  as_tibble() %>%

```

```

arrange(p_adj) %>%
mutate(significance = if_else(p_adj < 0.05, "*", "")) %>%
knitr::kable()

```

cluster_id	p_val	p_adj	significance
iNOS+_Pneumocyte	0.0002079	0.0041581	*
CD14+_Mono	0.0076753	0.0487124	*
Endothelial	0.0097425	0.0487124	*
iNOS+_Mac	0.0065791	0.0487124	*
CD57+_ImmuneOther	0.0125417	0.0501667	
CD16+_ImmuneOther	0.0346300	0.1154335	
Bcell	0.1194941	0.2830523	
Fibroblast	0.1273735	0.2830523	
Treg	0.1219037	0.2830523	
CD11c+_mDC	0.3118498	0.5669996	
Lung_Epithelium	0.3053443	0.5669996	
Neutrophil	0.3766638	0.6277730	
Eosinophil	0.4573792	0.6533988	
M2_Mac	0.4476823	0.6533988	
Mesenchymal	0.5003182	0.6670909	
CD57+_CD8+_Tcell	0.5506301	0.6882876	
CD4+_Tcell	0.6183913	0.7275192	
CD209+_Mac	0.6729049	0.7476721	
CD8+_Tcell	0.7636266	0.7636266	
Mast_cell	0.7446099	0.7636266	
Giant_cell	NA	NA	NA

And in this case we can see that there are several cell subtypes that are differentially abundant (pneumocytes, monocytes, endothelial cells, and iNOS+ Macrophages).

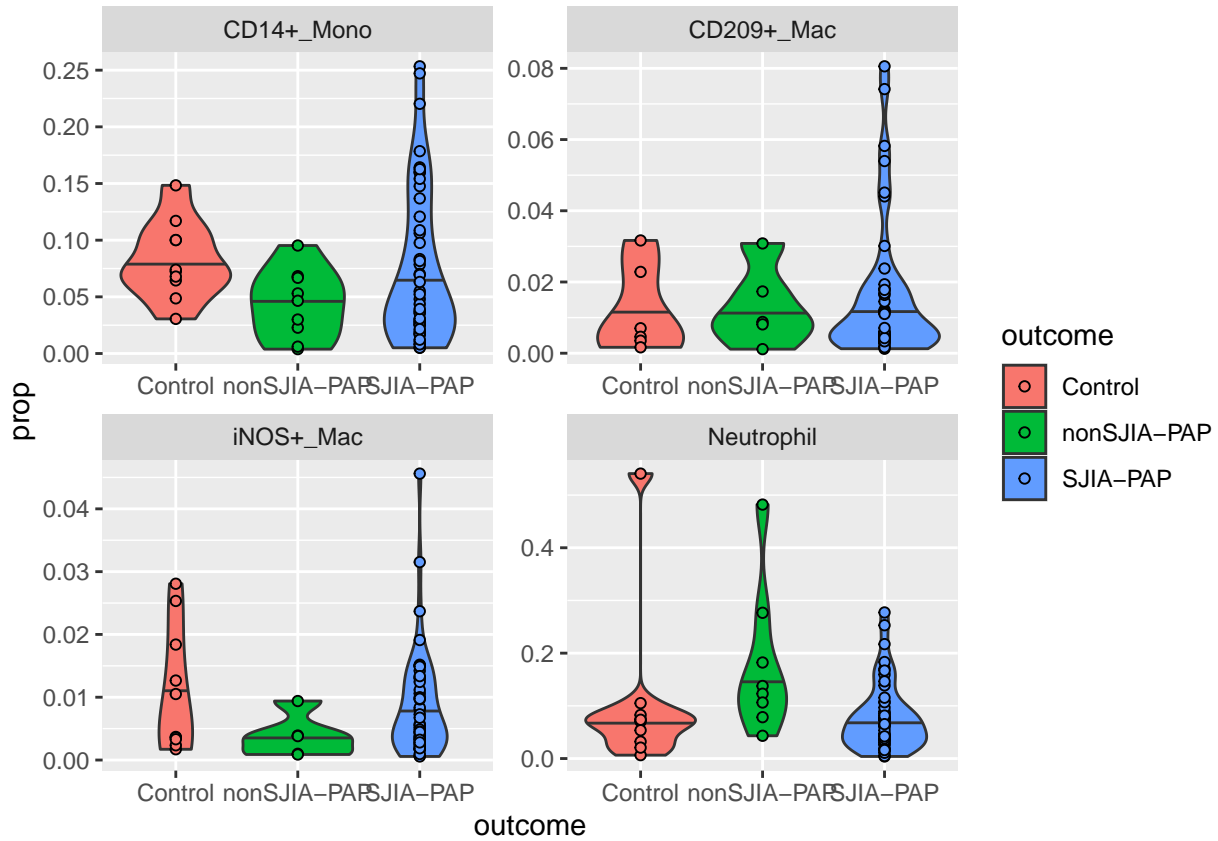
Visualization

```

interesting_clusters <-
  c("Neutrophil", "CD209+_Mac", "CD14+_Mono", "iNOS+_Mac")

mibi_data %>%
  count(fov_id, cluster_name, patient_id, outcome) %>%
  group_by(fov_id) %>%
  mutate(
    total_cells = sum(n),
    prop = n / total_cells
  ) %>%
  filter(cluster_name %in% interesting_clusters) %>%
  #group_by(patient_id, cluster_name, outcome) %>%
  #summarize(prop = mean(prop)) %>%
  ggplot(aes(x = outcome, y = prop, fill = outcome)) +
  geom_violin(draw_quantiles = 0.5) +
  geom_point(shape = 21, position = position_dodge(width = 0.3)) +
  facet_wrap(facets = vars(cluster_name), scales = "free")

```



From these plots, we can see that the differences between individual FOVS (even for these populations, which are the ones that came up in the analysis) are not very striking. In general, we might conclude that SJIA-PAP samples have slightly fewer neutrophils on average than control samples, but this seems largely driven by the outlier in the controls FOVs. To me, it looks like most of these differences are unreliable because of the low sample size. If you want to get into the realm of p-hacking, you could simply treat all the FOVs from a given patient as being entirely independent of one another (and I imagine that there are some studies that do this), but that's not an approach that I could endorse.

6. Differential Abundance Analysis - within patients

In addition to the between-patients comparisons, we can also run another set of comparisons that leverages a within-subjects design to increase statistical power. As it turns out, for several of our SJIA-PAP samples, one of the FOVs collected was annotated as a “healthy” section of tissue relative to the others (which had more of the hallmark histopathological features of SJIA-PAP). We can compare the abundance of each of our immune cell subpopulations within the same patients by comparing the “healthy” FOV to the other FOVs taken from the same patient. In this case the GLMM we’re using is the following:

$$\text{logit}(p_{ij}) = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta_0 + \alpha_i + \beta_1 X_{j_{diseased}},$$

where $X_{j_{diseased}}$ is an indicator variable representing if FOV j comes from a “diseased” area of the SJIA-PAP tissue or not (1 if yes; 0 otherwise). Thus, β_1 represents the difference in log-odds for a given cell subtype between the diseased and non-diseased FOVs within a given patients (controlling for patient-to-patient variability, which is represented by α_i).

Processing

```
# find patients who had at least one "healthy" FOV
interesting_patients <-
  mibi_data %>%
  filter(fov_id %in% healthy_fovs_in_pap_patients) %>%
  distinct(patient_id) %>%
  pull(patient_id)

interesting_patients

## [1] 16 8 9 11 1

# filter only the patients with "healthy-looking" FOVs from the full dataset
paired_patients <-
  mibi_data %>%
  #filter(patient_id %in% interesting_patients) %>%
  # annotate FOVs that are "healthy-looking" according to our pathologist
  mutate(
    fov_condition =
      if_else(fov_id %in% healthy_fovs_in_pap_patients, "healthy", "pap")
  )
```

Statistical analysis

```
paired_daa_results <-
  paired_patients %>%
  filter(outcome == "SJIA-PAP") %>%
  pap_perform_daa(
    data_tibble = .,
    sample_col = fov_id,
    cluster_col = cluster_name,
    fixed_effect_cols = fov_condition,
    random_effect_cols = patient_id,
    include_observation_level_random_effects = FALSE
  )

paired_daa_results %>%
```

```

pluck("da_results") %>%
topTable(all = TRUE) %>%
as_tibble() %>%
mutate(significant = if_else(p_adj < 0.05, "*", "")) %>%
arrange(p_adj) %>%
knitr::kable()

```

cluster_id	p_val	p_adj	significant
Bcell	0.0000000	0.0000000	*
Neutrophil	0.0000000	0.0000000	*
CD4+_Tcell	0.0000000	0.0000000	*
Fibroblast	0.0000000	0.0000000	*
M2_Mac	0.0000052	0.0000206	*
Treg	0.0000354	0.0001181	*
Eosinophil	0.0000445	0.0001246	*
Lung_Epithelium	0.0000498	0.0001246	*
Mast_cell	0.0007133	0.0015851	*
CD209+_Mac	0.0030874	0.0061748	*
CD57+_CD8+_Tcell	0.0218511	0.0397293	*
CD57+_ImmuneOther	0.0258245	0.0397300	*
iNOS+_Pneumocyte	0.0251787	0.0397300	*
iNOS+_Mac	0.0420420	0.0600600	
Endothelial	0.0630342	0.0840457	
Mesenchymal	0.0711934	0.0889917	
CD16+_ImmuneOther	0.3056815	0.3596253	
CD8+_Tcell	0.3353813	0.3726459	
CD11c+_mDC	0.9907574	0.9907574	
CD14+_Mono	0.9560004	0.9907574	

From these results, we can see that there are several immune cell subtypes that, when using a paired design, we find are enriched in parts of the SJIA-PAP lung that actually show histopathological signs of disease compared to paired parts of the SJIA-PAP lung that do not show histopathological signs of disease.

We can visualize these differences below.

Visualization

```

paired_p_values <-
  paired_daa_results %>%
  pluck("da_results") %>%
  topTable(all = TRUE) %>%
  as_tibble() %>%
  mutate(
    significant = if_else(p_adj < 0.05, "*", ""),
    new_cluster_name = if_else(significant == "*", str_c(cluster_id, "*"), as.character(cluster_id))
  ) %>%
  arrange(p_adj) %>%
  rename(cluster_name = cluster_id)

sig_clusters <-
  paired_p_values %>%
  filter(significant == "*") %>%

```



```

pull(cluster_name)

# calculate the number of fofs used for each patient in each "condition"
num_fov_tibble <-
  paired_patients %>%
    distinct(fov_id, patient_id, fov_condition) %>%
    count(patient_id, fov_condition, name = "num_fovs")

paired_plot_data <-
  paired_patients %>%
  mutate(
    cluster_name =
      factor(cluster_name, levels = pull(paired_p_values, cluster_name)) ,
  ) %>%
  count(cluster_name, patient_id, fov_id, fov_condition, .drop = FALSE) %>%
  group_by(fov_id) %>%
  mutate(
    total_fov_cells = sum(n),
    prop = n / total_fov_cells
  ) %>%
  # ungroup() %>%
  # group_by(fov_condition, patient_id, cluster_name) %>%
  # summarize(
  #   sd = sd(prop, na.rm = TRUE),
  #   prop = mean(prop, na.rm = TRUE),
  # ) %>%
  # drop_na(cluster_name) %>%
  # ungroup() %>%
  complete(patient_id, fov_id, fov_condition, cluster_name, fill = list(prop = 0))
  # left_join(num_fov_tibble) %>%
  # mutate(
  #   sem = sd / sqrt(num_fovs)
  # )

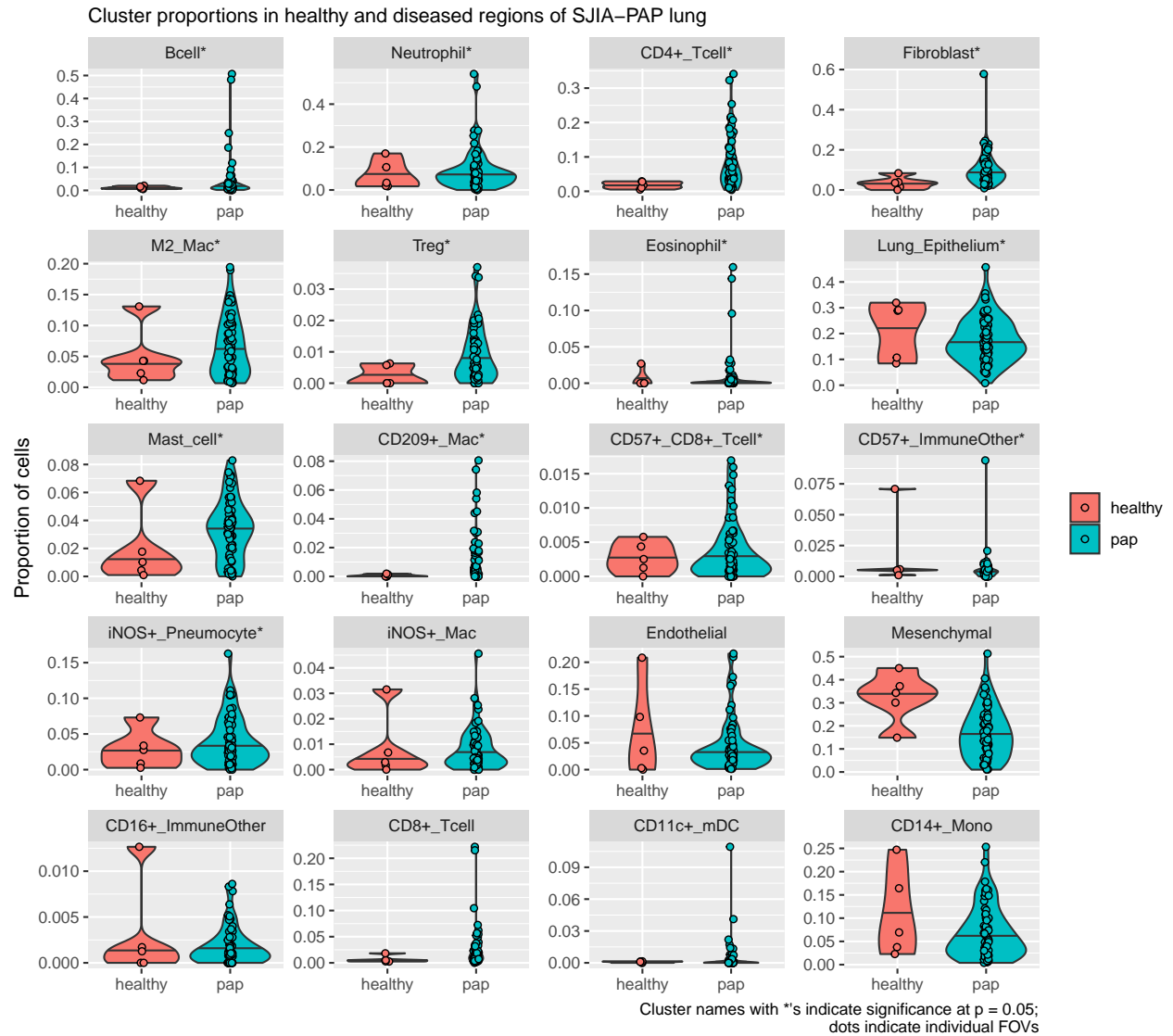
```

```

paired_plot_data %>%
  left_join(paired_p_values) %>%
  mutate(new_cluster_name = fct_reorder(new_cluster_name, p_adj)) %>%
  drop_na(new_cluster_name) %>%
  ggplot(aes(y = prop, x = fov_condition, fill = fov_condition)) +
  geom_violin(draw_quantiles = 0.5) +
  # geom_line(aes(group = patient_id), color = "black") +
  # geom_errorbar(
  #   aes(x = fov_condition, y = prop, ymin = prop - sem, ymax = prop + sem),
  #   width = 0.2,
  #   alpha = 0.7
  # ) +
  geom_jitter(shape = 21, size = 1.5, width = 0.03) +
  facet_wrap(facets = vars(new_cluster_name), scales = "free", ncol = 4) +
  labs(
    subtitle = "Cluster proportions in healthy and diseased regions of SJIA-PAP lung",
    x = NULL,
    y = "Proportion of cells",
    caption = "Cluster names with *'s indicate significance at p = 0.05;\n dots indicate individual FOVs",
    fill = NULL
  )

```

)



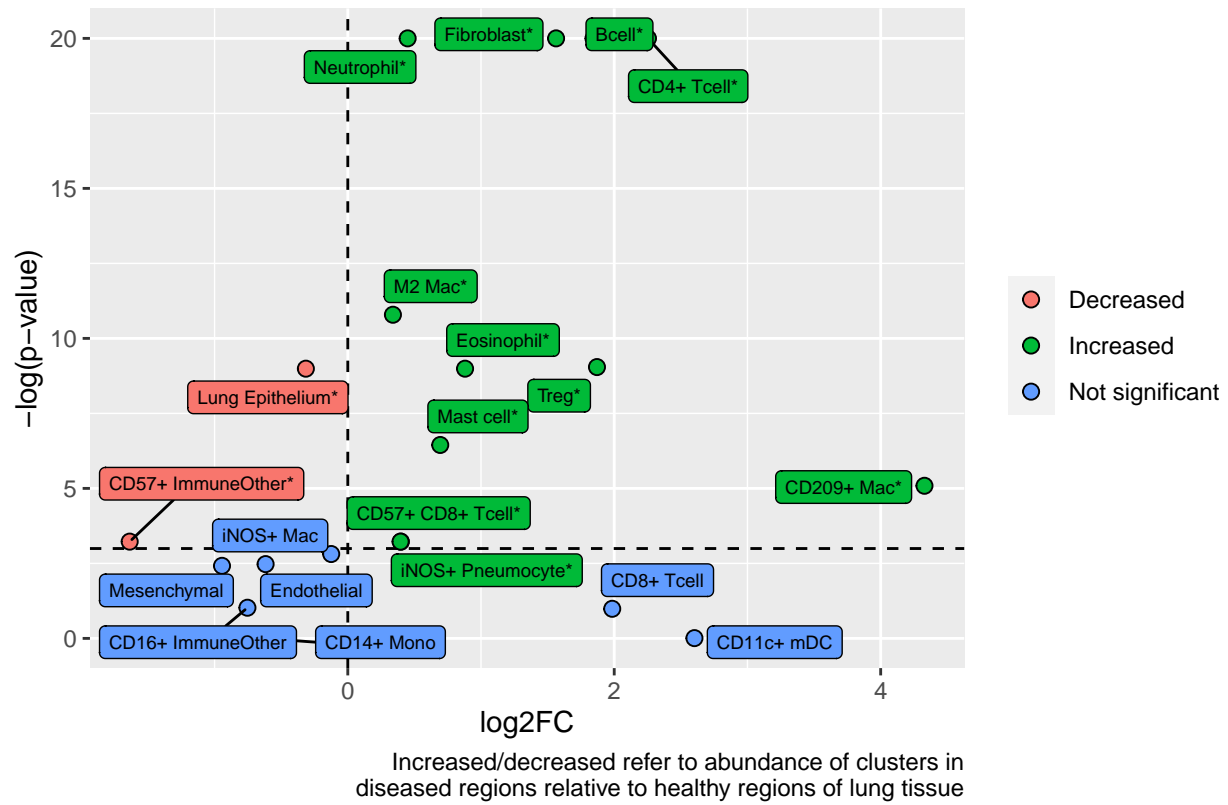
```
paired_plot_data %>%
  left_join(paired_p_values) %>%
  mutate(new_cluster_name = fct_reorder(new_cluster_name, p_adj)) %>%
  drop_na(new_cluster_name) %>%
  select(patient_id, fov_condition, new_cluster_name, prop, p_adj) %>%
  pivot_wider(
    names_from = fov_condition,
    values_from = prop
  ) %>%
  group_by(new_cluster_name) %>%
  summarize(across(c(healthy, pap), mean, na.rm = TRUE)) %>%
  mutate(
    fc = pap / healthy,
    log2_fc = log(fc, base = 2)
  ) %>%
  left_join(paired_p_values) %>%
```

```

mutate(
  neg_log_p_val = -log(p_adj),
  cluster_type =
    case_when(
      p_adj > 0.05 ~ "Not significant",
      fc > 1      ~ "Increased",
      fc < 1      ~ "Decreased"
    )
) %>%
ggplot(aes(x = log2_fc, y = neg_log_p_val, fill = cluster_type)) +
geom_hline(yintercept = -log(0.05), color = "black", linetype = "dashed") +
geom_vline(xintercept = 0, color = "black", linetype = "dashed") +
geom_point(shape = 21, size = 2.5) +
ggrepel::geom_label_repel(
  aes(label = str_replace_all(new_cluster_name, "_", " ")),
  size = 2.5,
  color = "black",
  show.legend = FALSE
) +
scale_y_continuous(limits = c(NA, 20), oob = scales::oob_squish_any) +
labs(
  subtitle = "Differentially abundant clusters in diseased vs. non-diseased regions of SJIA-PAP lung"
  x = "log2FC",
  y = "-log(p-value)",
  fill = NULL,
  caption = "Increased/decreased refer to abundance of clusters in\ndiseased regions relative to heal
)

```

Differentially abundant clusters in diseased vs. non-diseased regions of SJIA-PAP lung



7. Differential Expression Analysis - between patients

We can use a similar procedure as above to test for differences in mean marker expression across clusters by fitting a linear regression model (thus, an LMM, not a GLMM) for each cluster/marker pair. For this, we implicitly assume the that mean marker expression values in each cluster are distributed normally among patients (which is a large assumption even though this method is state-of-the-art in the cytometry community). Thus, for differential marker expression, we use the following equation to predict the mean expression value y_{ij} for a given cluster/marker pair in patient i and sample j :

$$y_{ij} = \beta_0 + \alpha_i + \beta_1 X_j,$$

where each of the variables on the right-hand side of the equation are defined as in section 5. Due to our small sample size, we use the `{{limma}}` package's parametric empirical Bayes method of estimating the variance for each marker being analyzed - this method allows for the sharing of variability information across all markers (and thereby an increase in statistical power in rarer cell types and less highly-expressed markers).

Thus, a linear model can be fit for each marker-cluster pair across all FOVs, and the significance of β_1 can be tested to indicate the effect of the `outcome` variable on marker expression in a given cluster.

PAP vs. non-PAP

We perform the same comparisons in the same order as in section 5 starting with all PAP vs. non-PAP samples:

```
pap_dea <-
  mibi_data %>%
  select(fov_id, cluster_name, outcome, patient_id, any_of(functional_markers)) %>%
  mutate(outcome = if_else(outcome %in% c("SJIA-PAP", "nonSJIA-PAP"), "PAP", outcome)) %>%
  pap_perform_dea(
    data_tibble = .,
    sample_col = fov_id,
    cluster_col = cluster_name,
    fixed_effect_cols = outcome,
    random_effect_col = c(patient_id),
    min_cells = 5,
    min_samples = 5
  )

# only show the top 25 most significant results
pap_dea$de_results %>%
  topTable(top_n = 25) %>%
  as_tibble() %>%
  arrange(p_adj) %>%
  mutate(significance = if_else(p_adj < 0.1, "*", "")) %>%
  knitr::kable()
```

cluster_id	marker_id	p_val	p_adj	significance
CD4+_Tcell	ho_1	0.0000000	0.0000002	*
CD8+_Tcell	ido	0.0000000	0.0000042	*
CD8+_Tcell	cd45ro	0.0000358	0.0028621	*
Neutrophil	cd45ro	0.0011362	0.0681716	*
CD4+_Tcell	cd45ro	0.0017471	0.0838621	*
Eosinophil	ido	0.0027084	0.1039222	
CD8+_Tcell	p_s6	0.0030311	0.1039222	

cluster_id	marker_id	p_val	p_adj	significance
Bcell	ho_1	0.0042130	0.1263909	
CD57+_CD8+_Tcell	ho_1	0.0050861	0.1356296	
CD11c+_mDC	h3k27me3	0.0061983	0.1487587	
iNOS+_Pneumocyte	tim3	0.0076825	0.1676173	
Bcell	ki67	0.0088984	0.1779685	
CD57+_ImmuneOther	ki67	0.0146243	0.2601873	
iNOS+_Pneumocyte	pd1	0.0151776	0.2601873	
CD16+_ImmuneOther	tim3	0.0187963	0.3007403	
CD209+_Mac	cd45ro	0.0217638	0.3264572	
Fibroblast	cd45ro	0.0246028	0.3473342	
CD8+_Tcell	pd1	0.0285025	0.3800328	
CD11c+_mDC	h3k9ac	0.0337778	0.4266667	
Endothelial	cd45ro	0.0357044	0.4284526	
Neutrophil	if_ng	0.0447195	0.5110802	
CD57+_CD8+_Tcell	p_s6	0.0604396	0.6593407	
CD57+_CD8+_Tcell	cd45ro	0.0639189	0.6669802	
CD14+_Mono	cd45ro	0.0884679	0.7473783	
CD16+_ImmuneOther	cd45ro	0.1261774	0.7473783	

From these results, we can see that most markers that are differentially expressed are in different kinds of T-cells and Neutrophils (with mast cells and M2 macrophages having 1 differentially expressed marker each).

SJIA-PAP vs. Controls

Now we compare SJIA-PAP samples to control samples:

```
dea_sjia_pap_vs_controls <-
  mibi_data %>%
  select(fov_id, cluster_name, outcome, patient_id, any_of(functional_markers)) %>%
  filter(outcome != "nonSJIA-PAP") %>%
  pap_perform_dea(
    data_tibble = .,
    sample_col = fov_id,
    cluster_col = cluster_name,
    fixed_effect_cols = outcome,
    random_effect_col = c(patient_id),
    min_cells = 5,
    min_samples = 5
  )

dea_sjia_pap_vs_controls$de_results %>%
  topTable(top_n = 25) %>%
  as_tibble() %>%
  arrange(p_adj) %>%
  mutate(significance = if_else(p_adj < 0.1, "*", "")) %>%
  knitr::kable()
```

cluster_id	marker_id	p_val	p_adj	significance
CD4+_Tcell	ho_1	0.0000000	0.0000026	*
CD8+_Tcell	ido	0.0000008	0.0000958	*
CD8+_Tcell	cd45ro	0.0000096	0.0007653	*
CD4+_Tcell	cd45ro	0.0012904	0.0774250	*

cluster_id	marker_id	p_val	p_adj	significance
Neutrophil	cd45ro	0.0040967	0.1966427	
CD11c+_mDC	h3k27me3	0.0060649	0.2425975	
iNOS+_Pneumocyte	tim3	0.0073126	0.2507194	
CD8+_Tcell	p_s6	0.0109152	0.3274561	
CD209+_Mac	cd45ro	0.0207195	0.3300516	
Endothelial	cd45ro	0.0189489	0.3300516	
Fibroblast	cd45ro	0.0215328	0.3300516	
CD57+_CD8+_Tcell	ho_1	0.0184302	0.3300516	
Eosinophil	ido	0.0220034	0.3300516	
Neutrophil	if_ng	0.0142605	0.3300516	
Bcell	ki67	0.0171273	0.3300516	
CD8+_Tcell	pd1	0.0200131	0.3300516	
iNOS+_Pneumocyte	pd1	0.0307803	0.4345460	
CD57+_CD8+_Tcell	cd45ro	0.0460641	0.5357992	
M2_Mac	cd45ro	0.0490651	0.5357992	
CD11c+_mDC	h3k9ac	0.0491149	0.5357992	
Bcell	ho_1	0.0437684	0.5357992	
CD16+_ImmuneOther	tim3	0.0438271	0.5357992	
CD8+_Tcell	h3k27me3	0.0551139	0.5751013	
CD14+_Mono	cd45ro	0.0765926	0.6287272	
Lung_Epithelium	cd45ro	0.0819103	0.6287272	

As above, we can see that most marker differences occur in different kinds of T-cells and neutrophils.

SJIA-PAP vs. nonSJIA-PAP

Next, we compare SJIA-PAP and nonSJIA-PAP samples:

```
dea_sjia_pap_vs_nonsjia_pap <-
  mibi_data %>%
  select(fov_id, cluster_name, outcome, patient_id, any_of(functional_markers)) %>%
  filter(outcome != "Control") %>%
  pap_perform_dea(
    data_tibble = .,
    sample_col = fov_id,
    cluster_col = cluster_name,
    fixed_effect_cols = outcome,
    random_effect_col = c(patient_id),
    min_cells = 20,
    min_samples = 10
  )

dea_sjia_pap_vs_nonsjia_pap$de_results %>%
  topTable(top_n = 25) %>%
  as_tibble() %>%
  arrange(p_adj) %>%
  mutate(significance = if_else(p_adj < 0.1, "*", "")) %>%
  knitr::kable()
```

cluster_id	marker_id	p_val	p_adj	significance
Endothelial	ido	0.0022663	0.3807362	
CD14+_Mono	cd45ro	0.1840452	0.9338707	

cluster_id	marker_id	p_val	p_adj	significance
CD209+_Mac	cd45ro	0.0865775	0.9338707	
CD4+_Tcell	cd45ro	0.1168509	0.9338707	
CD8+_Tcell	cd45ro	0.0169816	0.9338707	
Endothelial	cd45ro	0.0145036	0.9338707	
Fibroblast	cd45ro	0.0765620	0.9338707	
iNOS+_Pneumocyte	cd45ro	0.2110685	0.9338707	
Lung_Epithelium	cd45ro	0.1931939	0.9338707	
M2_Mac	cd45ro	0.0368404	0.9338707	
Mast_cell	cd45ro	0.1200942	0.9338707	
Mesenchymal	cd45ro	0.0781259	0.9338707	
Neutrophil	cd45ro	0.1918380	0.9338707	
Treg	cd45ro	0.1132751	0.9338707	
CD209+_Mac	h3k27me3	0.1989831	0.9338707	
CD8+_Tcell	h3k27me3	0.1088028	0.9338707	
iNOS+_Pneumocyte	h3k27me3	0.0370002	0.9338707	
Mesenchymal	h3k27me3	0.1248957	0.9338707	
Neutrophil	h3k27me3	0.1814009	0.9338707	
CD209+_Mac	h3k9ac	0.0769374	0.9338707	
CD4+_Tcell	h3k9ac	0.2185563	0.9338707	
Endothelial	h3k9ac	0.2352985	0.9338707	
Mast_cell	h3k9ac	0.1112206	0.9338707	
Mesenchymal	ido	0.2361879	0.9338707	
Treg	if_ng	0.2639630	0.9338707	

nonSJIA-PAP vs. Controls

Next, we compare nonSJIA-PAP vs. control samples:

```
dea_control_vs_nonsjia_pap <-
  mibi_data %>%
  select(fov_id, cluster_name, outcome, patient_id, any_of(functional_markers)) %>%
  filter(outcome != "SJIA-PAP") %>%
  pap_perform_dea(
    data_tibble = .,
    sample_col = fov_id,
    cluster_col = cluster_name,
    fixed_effect_cols = outcome,
    random_effect_col = c(patient_id),
    min_cells = 20,
    min_samples = 10
  )

dea_control_vs_nonsjia_pap$de_results %>%
  topTable(top_n = 25) %>%
  as_tibble() %>%
  arrange(p_adj) %>%
  mutate(significance = if_else(p_adj < 0.1, "*", "")) %>%
  knitr::kable()
```

cluster_id	marker_id	p_val	p_adj	significance
CD8+_Tcell	cd45ro	0.0045181	0.2144709	
Neutrophil	cd45ro	0.0045826	0.2144709	

cluster_id	marker_id	p_val	p_adj	significance
CD4+_Tcell	pd1	0.0053618	0.2144709	
Mast_cell	ido	0.0085479	0.2564375	
M2_Mac	ho_1	0.0124323	0.2983745	
CD4+_Tcell	cd45ro	0.0210810	0.4216193	
M2_Mac	ido	0.0278035	0.4676386	
CD4+_Tcell	tim3	0.0311759	0.4676386	
CD8+_Tcell	h3k27me3	0.0409855	0.4690515	
Neutrophil	h3k27me3	0.0469052	0.4690515	
CD4+_Tcell	p_s6	0.0459375	0.4690515	
Neutrophil	p_s6	0.0384854	0.4690515	
CD8+_Tcell	p_s6	0.0785973	0.7255132	
Neutrophil	ido	0.0894153	0.7664164	
M2_Mac	h3k9ac	0.1151328	0.8935096	
CD4+_Tcell	ki67	0.1191346	0.8935096	
CD8+_Tcell	h3k9ac	0.1495883	0.9606100	
Neutrophil	h3k9ac	0.1398555	0.9606100	
CD8+_Tcell	ido	0.1594351	0.9606100	
CD8+_Tcell	pd1	0.1601017	0.9606100	
Mast_cell	p_s6	0.1690468	0.9659818	
CD14+_Mono	cd45ro	0.9524603	1.0000000	
Fibroblast	cd45ro	0.2988803	1.0000000	
iNOS+_Pneumocyte	cd45ro	0.8601321	1.0000000	
Lung_Epithelium	cd45ro	0.8190682	1.0000000	

And from these results, we can see that there are no differentially expressed functional markers between nonSJIA-PAP vs. control samples.

Visualization

In our visualizations, we focus on the differences that we saw between SJIA-PAP and control samples.

```
dea_sjia_pap_vs_control_results <-
  dea_sjia_pap_vs_controls$de_results %>%
  topTable(all = TRUE) %>%
  as_tibble() %>%
  arrange(p_adj) %>%
  mutate(significance = if_else(p_adj < 0.1, "*", "")) %>%
  rename(cluster_name = cluster_id)

# number of significantly different markers in each cluster
dea_sjia_pap_vs_control_results %>%
  filter(significance == "*") %>%
  count(cluster_name) %>%
  knitr::kable()
```

cluster_name	n
CD4+_Tcell	2
CD8+_Tcell	2

As noted above, we can see that T-cells have the significant functional marker differences.

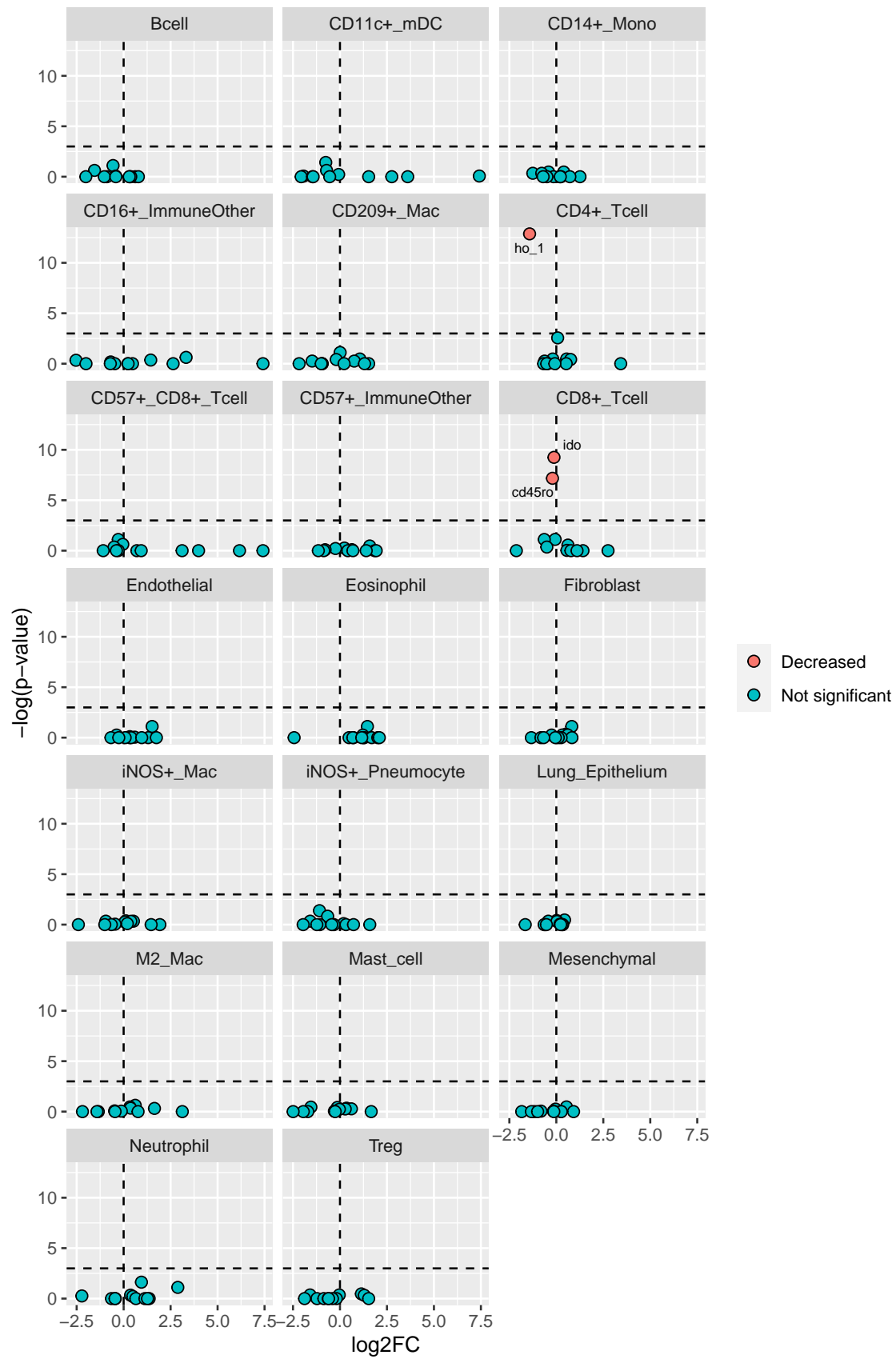
We can also make some volcano plots to summarize the different cluster-marker differences and their statistical significance:

```
feature_volcano_tibble <-
  mibi_data %>%
  filter(outcome != "nonSJIA-PAP") %>%
  group_by(cluster_name, outcome, patient_id) %>%
  summarize(across(any_of(functional_markers), .f = mean, na.rm = TRUE)) %>%
  ungroup() %>%
  pivot_longer(
    cols = any_of(functional_markers),
    names_to = "marker_id",
    values_to = "median"
  ) %>%
  group_by(outcome, marker_id, cluster_name) %>%
  summarize(median = mean(median, na.rm = TRUE)) %>%
  ungroup() %>%
  pivot_wider(
    names_from = outcome,
    values_from = median
  ) %>%
  mutate(
    fc = `SJIA-PAP` / Control,
    log2_fc = log(fc, base = 2)
  ) %>%
  #filter(!is.nan(fc)) %>%
  left_join(
    dea_sjia_pap_vs_control_results %>%
      drop_na()
  ) %>%
  arrange(p_adj) %>%
  mutate(
    neg_log_p_val = -log(p_adj),
    feature_type =
      case_when(
        p_adj > 0.05 ~ "Not significant",
        fc < 1 ~ "Decreased",
        fc > 1 ~ "Increased"
      ),
    feature = str_c(marker_id, cluster_name, sep = "@"),
  ) %>%
  drop_na(feature_type, significance)

feature_volcano_tibble %>%
  ggplot(aes(x = log2_fc, y = neg_log_p_val, fill = feature_type)) +
  geom_hline(yintercept = -log(0.05), color = "black", linetype = "dashed") +
  geom_vline(xintercept = 0, color = "black", linetype = "dashed") +
  geom_point(shape = 21, size = 2.5) +
  ggrepel::geom_text_repel(
    aes(label = marker_id),
    data = filter(feature_volcano_tibble, feature_type != "Not significant"),
    size = 2.5,
    color = "black"
  ) +
```

```
scale_x_continuous(oob = scales::oob_squish_infinite) +  
scale_y_continuous(oob = scales::oob_squish_infinite) +  
facet_wrap(facets = vars(cluster_name), ncol = 3) +  
labs(  
  subtitle = "Differentially expressed markers in SJIA-PAP vs. control samples",  
  x = "log2FC",  
  y = "-log(p-value)",  
  fill = NULL  
)
```

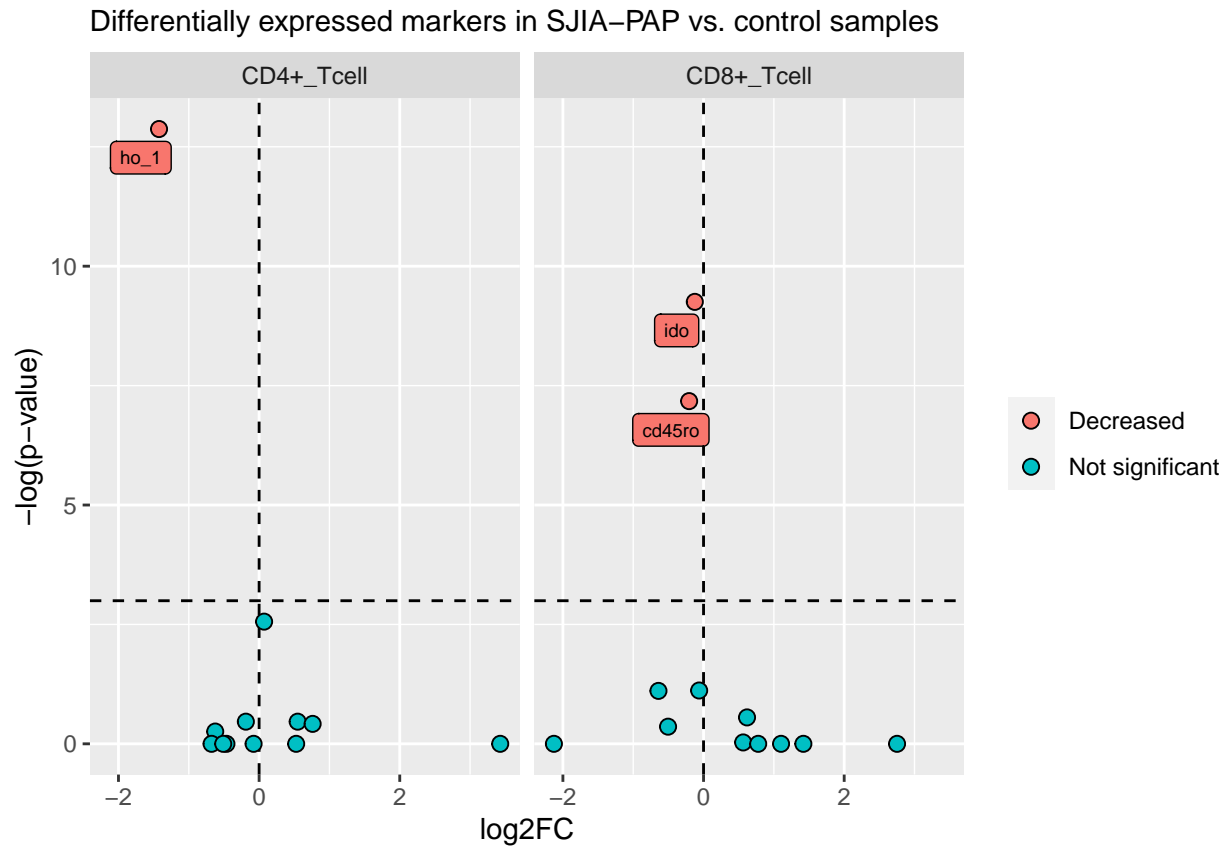
Differentially expressed markers in SJIA-PAP vs. control samples



Or the same plot, but only in clusters where there were significant results:

```
important_clusters <-
  dea_sjia_pap_vs_control_results %>%
  filter(significance == "*") %>%
  count(cluster_name) %>%
  pull(cluster_name) %>%
  as.character()

feature_volcano_tibble %>%
  filter(cluster_name %in% important_clusters) %>%
  ggplot(aes(x = log2_fc, y = neg_log_p_val, fill = feature_type)) +
  geom_hline(yintercept = -log(0.05), color = "black", linetype = "dashed") +
  geom_vline(xintercept = 0, color = "black", linetype = "dashed") +
  geom_point(shape = 21, size = 2.5) +
  ggrepel::geom_label_repel(
    aes(label = marker_id),
    data = filter(feature_volcano_tibble, feature_type != "Not significant"),
    size = 2.5,
    color = "black",
    show.legend = FALSE
  ) +
  scale_x_continuous(oob = scales::oob_squish_infinite) +
  scale_y_continuous(oob = scales::oob_squish_infinite) +
  facet_wrap(facets = vars(cluster_name), ncol = 2) +
  labs(
    subtitle = "Differentially expressed markers in SJIA-PAP vs. control samples",
    x = "log2FC",
    y = "-log(p-value)",
    fill = NULL
  )
```



8. Differential Expression Analysis - within patients

As above, we can compare healthy-looking and diseased-looking sections of the same SJIA-PAP patient samples using a linear model similar to that used in section 7:

$$y_{ij} = \beta_0 + \alpha_i + \beta_1 X_{j_{diseased}},$$

where all the parameters on the right-hand side of the equation are defined as in section 6.

```
paired_dea_results <-
  paired_patients %>%
  filter(outcome != "nonSJIA-PAP") %>%
  select(
    fov_id,
    cluster_name,
    fov_condition,
    patient_id,
    any_of(functional_markers)
  ) %>%
  pap_perform_dea(
    data_tibble = .,
    sample_col = fov_id,
    cluster_col = cluster_name,
    fixed_effect_cols = fov_condition,
    random_effect_col = patient_id,
    min_cells = 5,
    min_samples = 5
  )

paired_dea_results %>%
  pluck("de_results") %>%
  topTable(top_n = 25) %>%
  as_tibble() %>%
  mutate(significant = if_else(p_adj < 0.10, "*", "")) %>%
  arrange(p_adj) %>%
  knitr::kable()
```

cluster_id	marker_id	p_val	p_adj	significant
Endothelial	grz_b	0.0004056	0.0924748	*
Bcell	cd45ro	0.0031363	0.3575354	
CD14+_Mono	cd45ro	0.0128996	0.9803704	
CD16+_ImmuneOther	cd45ro	0.5196961	1.0000000	
CD209+_Mac	cd45ro	0.8096830	1.0000000	
CD4+_Tcell	cd45ro	0.1204974	1.0000000	
CD57+_CD8+_Tcell	cd45ro	0.5775643	1.0000000	
CD57+_ImmuneOther	cd45ro	0.4191940	1.0000000	
CD8+_Tcell	cd45ro	0.4382790	1.0000000	
Endothelial	cd45ro	0.1850688	1.0000000	
Eosinophil	cd45ro	0.6576018	1.0000000	
Fibroblast	cd45ro	0.1129570	1.0000000	
iNOS+_Mac	cd45ro	0.4454689	1.0000000	
iNOS+_Pneumocyte	cd45ro	0.1136502	1.0000000	
Lung_Epithelium	cd45ro	0.0905363	1.0000000	
M2_Mac	cd45ro	0.0869136	1.0000000	

cluster_id	marker_id	p_val	p_adj	significant
Mast_cell	cd45ro	0.3258158	1.0000000	
Mesenchymal	cd45ro	0.1030600	1.0000000	
Neutrophil	cd45ro	0.4846902	1.0000000	
Treg	cd45ro	0.1836610	1.0000000	
Bcell	grz_b	0.9955328	1.0000000	
CD14+_Mono	grz_b	1.0000000	1.0000000	
CD16+_ImmuneOther	grz_b	0.9174188	1.0000000	
CD209+_Mac	grz_b	1.0000000	1.0000000	
CD4+_Tcell	grz_b	1.0000000	1.0000000	

9. Spatial Analysis

Within FOVs, we can also annotate individual cells depending on whether they overlap with surfactant/lipid plaques and use these annotations to compare cell types that interact with surfactant and those that don't. To do so, we can use a similar GLMM as above:

[Write out GLMM]

Annotate cells

```
mibi_data <-
  mibi_data %>%
  mutate(fov_id = as.character(fov_id)) %>%
  left_join(surf_data, by = c("fov_id", "centroid_x" = "x", "centroid_y" = "y")) %>%
  rename(surfactant = values) %>%
  mutate(surfactant = replace_na(surfactant, "no surfactant"))
```

```
# count how many cells are associated with surfactant in the entire dataset
```

```
mibi_data %>%
  count(surfactant)
```

```
## # A tibble: 2 x 2
##   surfactant      n
##   <chr>        <int>
## 1 no surfactant 65161
## 2 surfactant   13895
```

DAA

```
surf_daa_results <-
  mibi_data %>%
  filter(outcome == "SJIA-PAP") %>%
  mutate(sample_id = str_c(fov_id, surfactant)) %>%
  pap_perform_daa(
    sample_col = sample_id,
    cluster_col = cluster_name,
    fixed_effect_cols = surfactant,
    random_effect_cols = patient_id,
    include_observation_level_random_effects = FALSE,
    min_cells = 3,
    min_samples = 5
  )

surf_daa_results <-
  surf_daa_results %>%
  pluck("da_results") %>%
  topTable(all = TRUE) %>%
  as_tibble() %>%
  mutate(significant = if_else(p_adj < 0.05, "*", "")) %>%
  arrange(p_adj)

surf_daa_results %>%
  knitr::kable()
```

cluster_id	p_val	p_adj	significant
Bcell	0.0000000	0.0000000	*
CD4+_Tcell	0.0000000	0.0000000	*
CD14+_Mono	0.0000000	0.0000000	*
Eosinophil	0.0000000	0.0000000	*
Endothelial	0.0000438	0.0001753	*
CD209+_Mac	0.0001663	0.0005543	*
Neutrophil	0.0011513	0.0032894	*
M2_Mac	0.0093387	0.0233466	*
Fibroblast	0.0300237	0.0604102	
Treg	0.0302051	0.0604102	
CD57+_ImmuneOther	0.0435328	0.0791506	
CD8+_Tcell	0.0847449	0.1412415	
CD57+_CD8+_Tcell	0.1249054	0.1921621	
iNOS+_Mac	0.1543959	0.2205656	
Mesenchymal	0.2176455	0.2901939	
Lung_Epithelium	0.2439831	0.3049788	
iNOS+_Pneumocyte	0.3890947	0.4577584	
Mast_cell	0.5190599	0.5767332	
CD11c+_mDC	0.9718087	0.9815762	
CD16+_ImmuneOther	0.9815762	0.9815762	

```

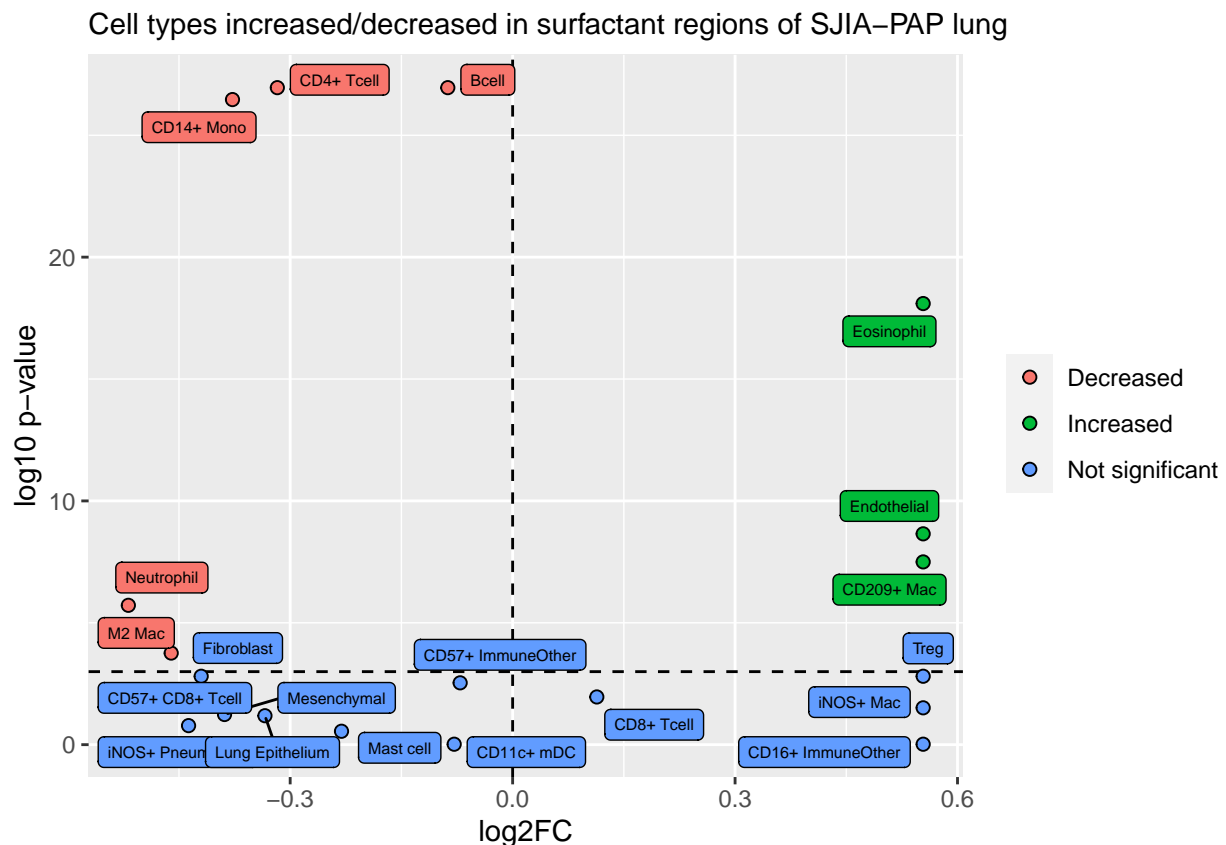
mibi_data %>%
  filter(outcome == "SJIA-PAP") %>%
  mutate(sample_id = str_c(fov_id, surfactant)) %>%
  count(patient_id, fov_id, cluster_name, surfactant) %>%
  group_by(patient_id, fov_id, surfactant) %>%
  mutate(prop = n / sum(n)) %>%
  ungroup() %>%
  select(-n) %>%
  pivot_wider(
    names_from = surfactant,
    values_from = prop,
    values_fill = 0
  ) %>%
  left_join(surf_daa_results, by = c("cluster_name" = "cluster_id")) %>%
  mutate(
    fc = surfactant / `no surfactant`,
  ) %>%
  group_by(cluster_name) %>%
  summarize(
    fc = mean(fc, na.rm = TRUE),
    p_adj = mean(p_adj, na.rm = TRUE),
    cluster_type =
      case_when(
        p_adj > 0.05 ~ "Not significant",
        fc < 1 ~ "Decreased",
        fc > 1 ~ "Increased"
      )
  ) %>%
  mutate(
    log2_fc = log(fc, base = 2),

```

```

    neg_log_10_p = -log(p_adj)
  ) %>%
  #mutate(cluster_name = fct_reorder(cluster_name, neg_log_10_p)) %>%
  ggplot(aes(y = neg_log_10_p, x = log2_fc, fill = cluster_type)) +
  geom_hline(yintercept = -log(0.05), color = "black", linetype = "dashed") +
  geom_vline(xintercept = 0, color = "black", linetype = "dashed") +
  #geom_violin() +
  geom_point(shape = 21, size = 2) +
  ggrepel::geom_label_repel(
    aes(label = str_replace_all(cluster_name, "_", " ")),
    size = 2,
    show.legend = FALSE
  ) +
  scale_x_continuous(oob = scales::oob_squish_any) +
  scale_y_continuous(oob = scales::oob_squish_any) +
  labs(
    subtitle = "Cell types increased/decreased in surfactant regions of SJIA-PAP lung",
    x = "log2FC",
    y = "log10 p-value",
    fill = NULL
  )
)

```



DEA

```

surf_dea_results <-
  mibi_data %>%

```

```

filter(outcome == "SJIA-PAP") %>%
mutate(sample_id = as.factor(str_c(fov_id, surfactant))) %>%
select(
  sample_id,
  cluster_name,
  surfactant,
  patient_id,
  any_of(functional_markers)
) %>%
pap_perform_dea(
  sample_col = sample_id,
  cluster_col = cluster_name,
  fixed_effect_cols = surfactant,
  random_effect_col = patient_id,
  min_cells = 10,
  min_samples = 5
)

surf_dea_results %>%
  pluck("de_results") %>%
  topTable(top_n = 25) %>%
  as_tibble() %>%
  mutate(significant = if_else(p_adj < 0.05, "*", "")) %>%
  arrange(p_adj) %>%
  knitr::kable()

```

cluster_id	marker_id	p_val	p_adj	significant
Bcell	cd45ro	0.0187763	1	
CD11c+_mDC	cd45ro	0.4947505	1	
CD14+_Mono	cd45ro	0.6168207	1	
CD209+_Mac	cd45ro	0.4528022	1	
CD4+_Tcell	cd45ro	0.3620357	1	
CD57+_CD8+_Tcell	cd45ro	0.9669379	1	
CD8+_Tcell	cd45ro	0.8879444	1	
Endothelial	cd45ro	0.2332030	1	
Eosinophil	cd45ro	0.3619556	1	
Fibroblast	cd45ro	0.7300307	1	
iNOS+_Mac	cd45ro	0.5407737	1	
iNOS+_Pneumocyte	cd45ro	0.7939156	1	
Lung_Epithelium	cd45ro	0.8236510	1	
M2_Mac	cd45ro	0.5732392	1	
Mast_cell	cd45ro	0.7655893	1	
Mesenchymal	cd45ro	0.6473117	1	
Neutrophil	cd45ro	0.5851816	1	
Treg	cd45ro	0.5137352	1	
Bcell	grz_b	0.7974925	1	
CD11c+_mDC	grz_b	0.9716647	1	
CD14+_Mono	grz_b	0.8856596	1	
CD209+_Mac	grz_b	0.9601187	1	
CD4+_Tcell	grz_b	1.0000000	1	
CD57+_CD8+_Tcell	grz_b	0.6993405	1	
CD8+_Tcell	grz_b	0.6946353	1	

We can see from this analysis that cells that overlap with surfactant and those that do not don't differ in any of our functional markers.