Classification

Max J. Karlsson

2020 M03 3

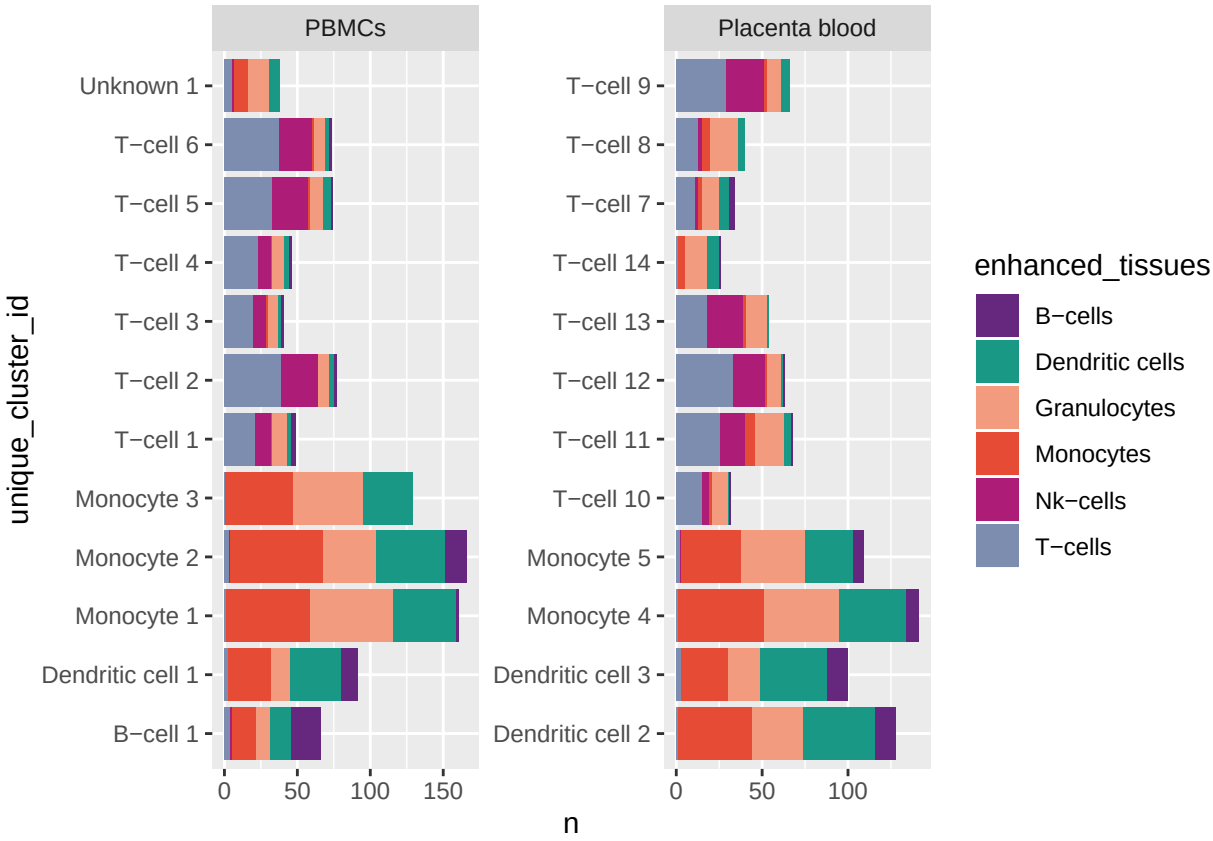
Basic plots

```
cluster_top_genes %>%
  left_join(cluster_annotation) %>%
  left_join(gene_info92 %>%
              select(1, 2),
            by = c("gene" = "gene_name")) %>%
  left_join(blood_cell_category) %>%
  separate_rows(enhanced_tissues, sep = ",") %>%
  mutate(enhanced_tissues = ifelse(is.na(enhanced_tissues),
                                   specificity_category,
                                   enhanced_tissues)) %>%
  filter(specificity_category %in% c("Tissue enriched", "Group enriched")) %>%
  group_by(dataset, cluster, cluster_id, unique_cluster_id, enhanced_tissues) %>%
  summarise(n = n()) %>%
  ungroup() %>%
  mutate(enhanced_tissues = str_to_sentence(enhanced_tissues),
         cluster_id = factor(cluster_id,
                             levels = paste0("Cluster-", 0:100))) %>%
  ggplot(aes(unique_cluster_id, n, fill = enhanced_tissues)) +
  geom_col() +
  coord_flip() +
  scale_fill_manual(values = c(tissue_colors, gene_category_pal)) +
  facet_wrap(~dataset, scales = "free")
```

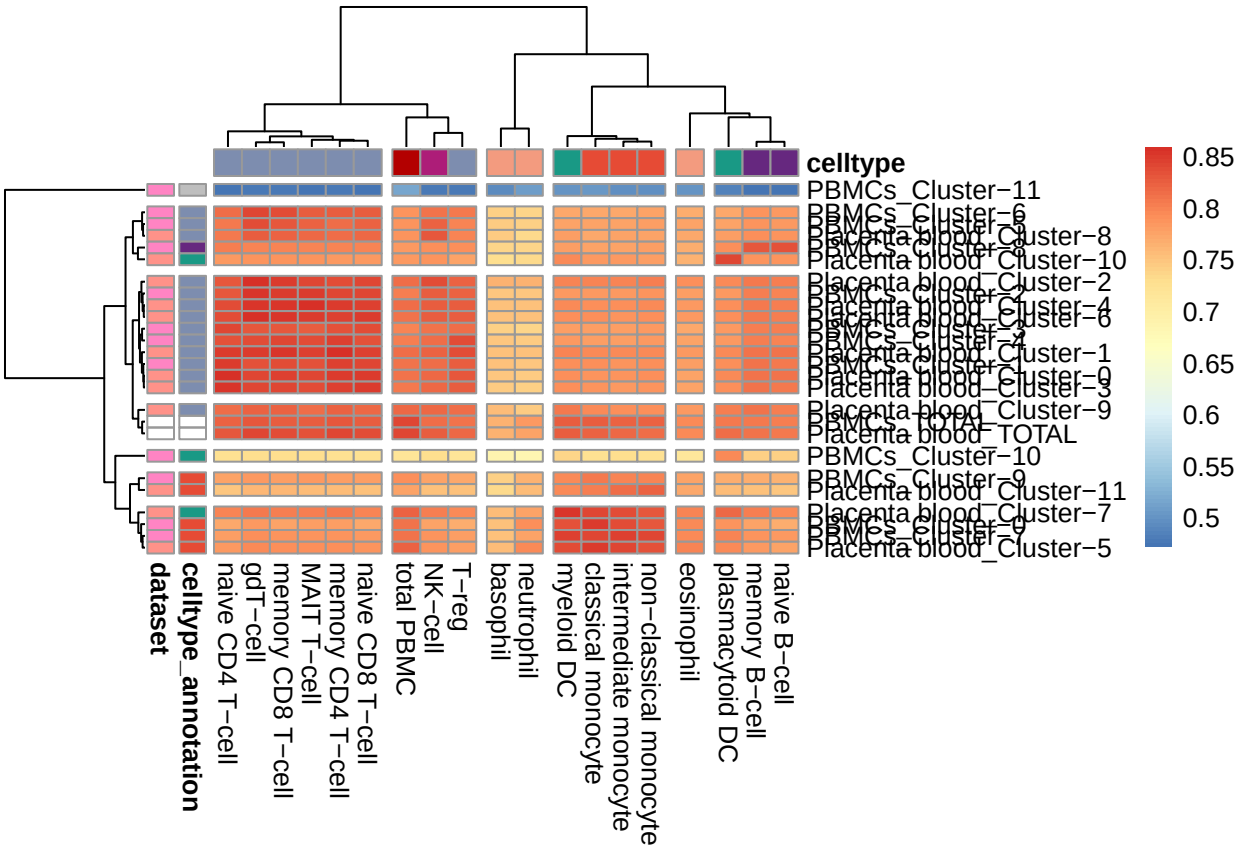
```
## Joining, by = c("dataset", "cluster_id")
```

```
## Warning: Column `cluster_id` joining factor and character vector, coercing
## into character vector
```

```
## Joining, by = "ensg_id"
```



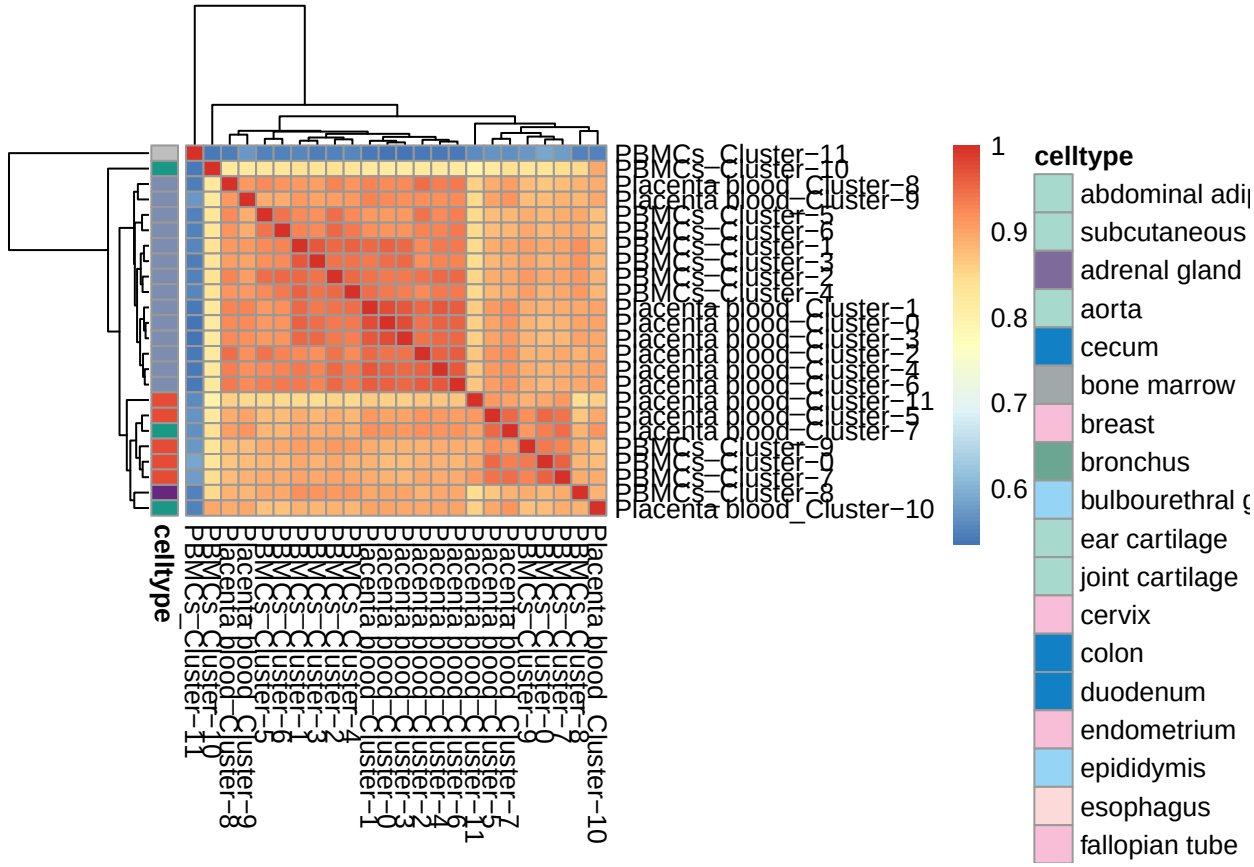

```
cluster_blood_cor %>%
  mutate(dataset_cluster = paste(dataset, cluster_id, sep = "_")) %>%
  select(dataset_cluster, cell_type, correlation) %>%
  spread(cell_type, correlation) %>%
  column_to_rownames("dataset_cluster") %>%
  pheatmap(clustering_method = "ward.D2",
           cutree_cols = 6,
           cutree_rows = 7,
           annotation_row = cluster_annotation %>%
             select(cluster, celltype_annotation = celltype, dataset) %>%
             column_to_rownames("cluster"),
           annotation_col = blood_cell_hierarchy %>%
             select(content, celltype = content_l1) %>%
             column_to_rownames("content"),
           annotation_colors = list(celltype = tissue_colors,
                                    celltype_annotation = tissue_colors),
           annotation_legend = F)
```



CD marker PCA

```
CD_pca_data <-
  cluster_norm_count %>%
  filter(ensg_id %in% CD_marker_list$Ensembl) %>%
  left_join(cluster_annotation,
            by = c("dataset", "cluster_id")) %>%
  select(unique_cluster_id, ensg_id, norm_count) %>%
  spread(unique_cluster_id, norm_count) %>%
  column_to_rownames("ensg_id") %>%
  {log10(. + 1)} %>%
  t()


cluster_norm_spearman <-
  cluster_norm_count %>%
  left_join(cluster_annotation,
            by = c("dataset", "cluster_id"))  %>%
  select(cluster, ensg_id, norm_count) %>%
  spread(cluster, norm_count) %>%
  column_to_rownames("ensg_id") %>%
  {log10(. + 1)} %>%
  cor(method = "spearman")
```

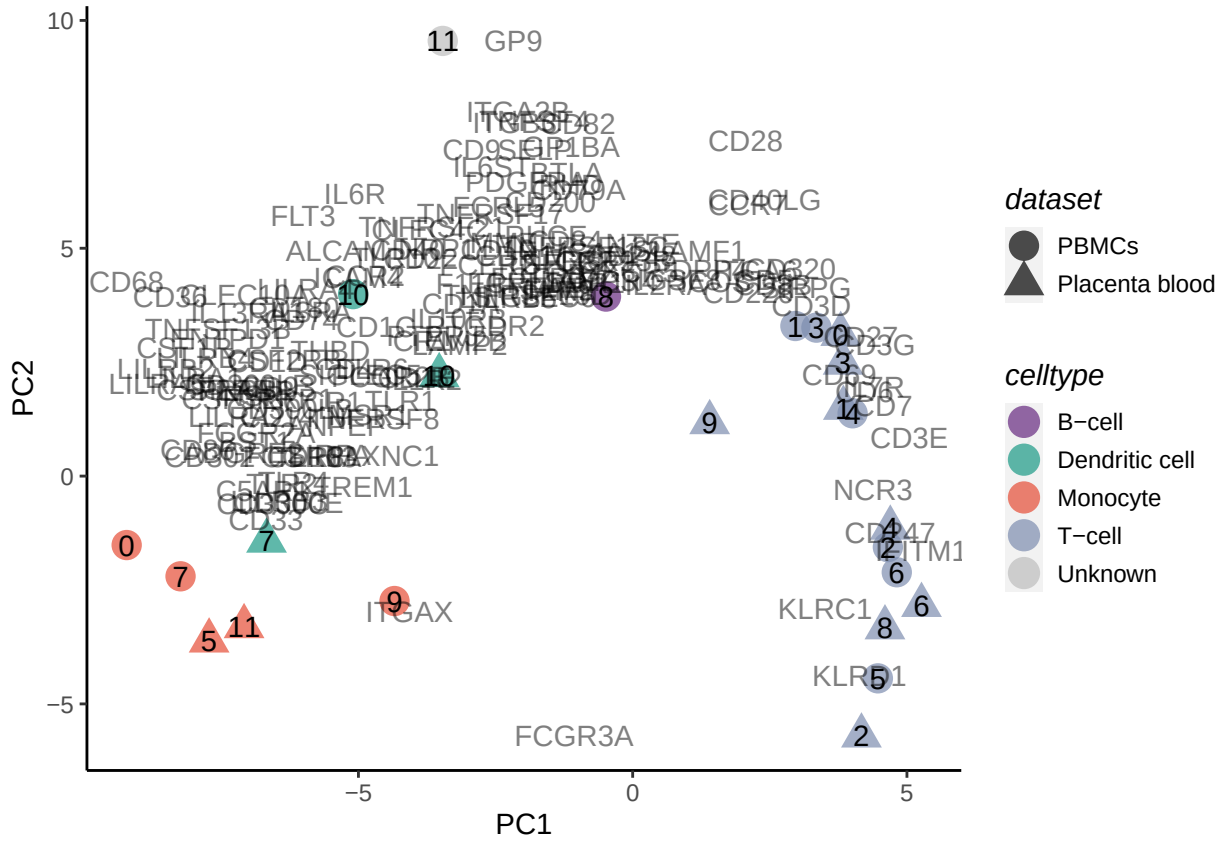
```
cluster_norm_spearman %>%
  pheatmap(annotation_row = cluster_annotation %>%
             select(cluster, celltype) %>%
             column_to_rownames("cluster"),
           annotation_colors = list(celltype = tissue_colors))
```




```
cluster_CD_pca <-
  CD_pca_data %>%
  pca_calc(npcs = 10)

PC1_lims <-
  cluster_CD_pca$scores[,"PC1"] %>%
  {c(min(.), max(.))}

PC2_lims <-
  cluster_CD_pca$scores[,"PC2"] %>%
  {c(min(.), max(.))}

cluster_CD_pca$scores %>%
  as_tibble(rownames = "unique_cluster_id") %>%
  left_join(cluster_annotation) %>%
  ggplot(aes(PC1, PC2, color = celltype, shape = dataset)) +
  geom_point(size = 5,
             alpha = 0.7) +
  geom_text(aes(label = str_extract(cluster_id, "\\d*$")),
            color = "black") +
```

```
  scale_color_manual(values = tissue_colors) +

  geom_text(data = cluster_CD_pca$loadings %>%
              as_tibble(rownames = "ensg_id") %>%
              left_join(gene_info92) %>%
              select(1:3, gene_name) %>%
              mutate(PC1 = scales::rescale(PC1, PC1_lims),
                     PC2 = scales::rescale(PC2, PC2_lims),
                     len = sqrt(PC1^2 + PC2^2)) %>%
              arrange(-len) %>%
              head(150),
            aes(PC1, PC2, label = gene_name ),
            inherit.aes = F,
            alpha = 0.5) +
  stripped_theme
```

```
## Joining, by = "unique_cluster_id"Joining, by = "ensg_id"
```




```
ggsave
```

```
## function (filename, plot = last_plot(), device = NULL, path = NULL,
##     scale = 1, width = NA, height = NA, units = c("in", "cm",
##         "mm"), dpi = 300, limitsize = TRUE, ...)
## {
##     dpi <- parse_dpi(dpi)
##     dev <- plot_dev(device, filename, dpi = dpi)
##     dim <- plot_dim(c(width, height), scale = scale, units = units,
```

```
##         limitsize = limitsize)
##     if (!is.null(path)) {
##         filename <- file.path(path, filename)
##     }
##     old_dev <- grDevices::dev.cur()
##     dev(filename = filename, width = dim[1], height = dim[2],
##         ...)
##     on.exit(utils::capture.output({
##         grDevices::dev.off()
##         if (old_dev > 1) grDevices::dev.set(old_dev)
##     }))
##     grid.draw(plot)
##     invisible()
## }
## <bytecode: 0x000000005aca8cd0>
## <environment: namespace:ggplot2>
```

Classification

```
celltype_max_norm_count <-
  cluster_norm_count %>%
  left_join(cluster_annotation) %>%
  group_by(celltype, ensg_id) %>%
  summarise(norm_count = max(norm_count)) %>%
  ungroup()
```

```
## Joining, by = c("dataset", "cluster_id")
```

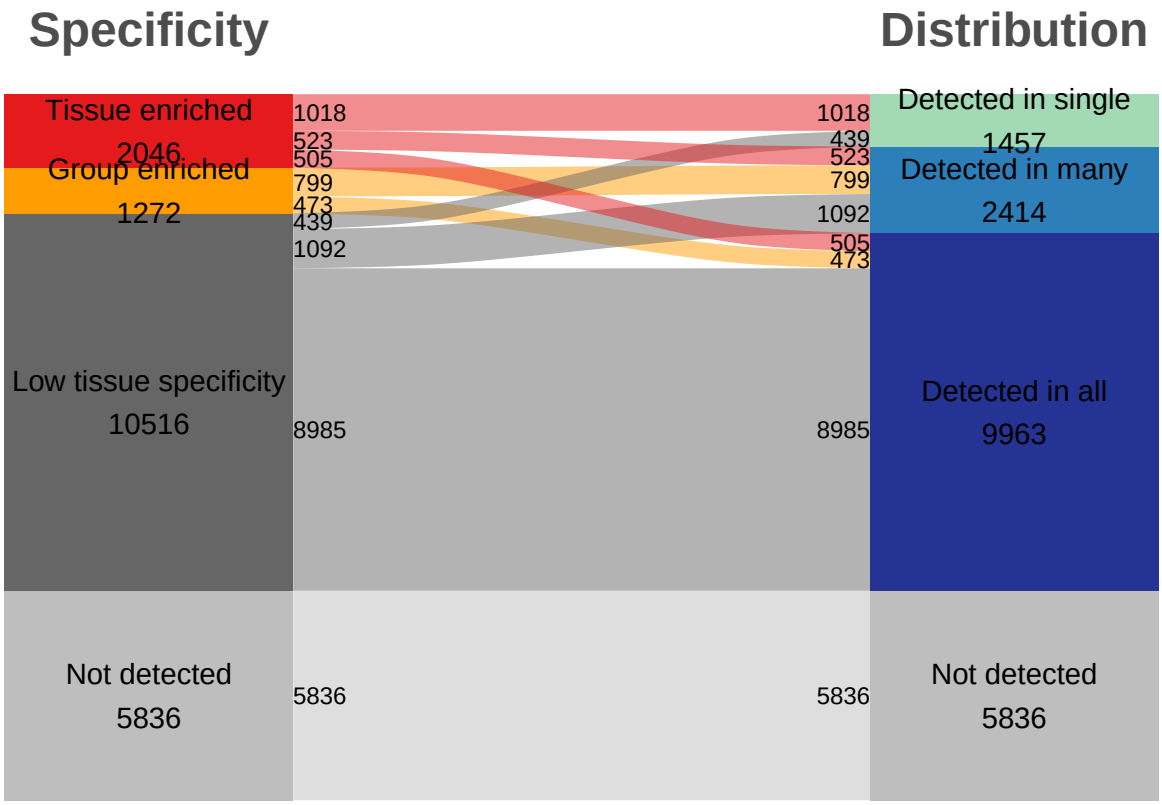
```
classification_max_norm_count <-
  celltype_max_norm_count %>%
  filter(celltype != "Unknown") %>%
  hpa_gene_classification(expression_col = "norm_count",
                          tissue_col = "celltype",
                          gene_col = "ensg_id",

                          enr_fold =  4,
                          max_group_n = 2,
                          det_lim = 1)

classification_sep_norm_count <-
  cluster_norm_count %>%
  left_join(cluster_annotation) %>%
  # filter(celltype != "Unknown") %>%
  hpa_gene_classification_multi_sample(expression_col = "norm_count",
                                       tissue_col = "celltype",
                                       gene_col = "ensg_id",
                                       sample_col = "unique_cluster_id",
                                       enr_fold =  4,
                                       max_group_n = 2,
                                       det_lim = 1)
```

```
## Joining, by = c("dataset", "cluster_id")
```

Classification plots

```
classification_max_norm_count %>%
  mutate(spec_category = str_to_sentence(spec_category),
         dist_category = str_to_sentence(dist_category)) %>%

  multi_alluvial_plot(vars = c("Specificity" = "spec_category",
                               "Distribution" = "dist_category"),
                      chunk_levels = c('Tissue enriched', 'Group enriched',
                                       'Tissue enhanced', 'Low tissue specificity',
                                       'Detected in single',
                                       'Detected in some',
                                       'Detected in many',
                                       'Detected in all',
                                       'Not detected'),
                      pal = c(gene_category_pal, elevation_identity_pal),
                      color_by = c(1, 1))
```

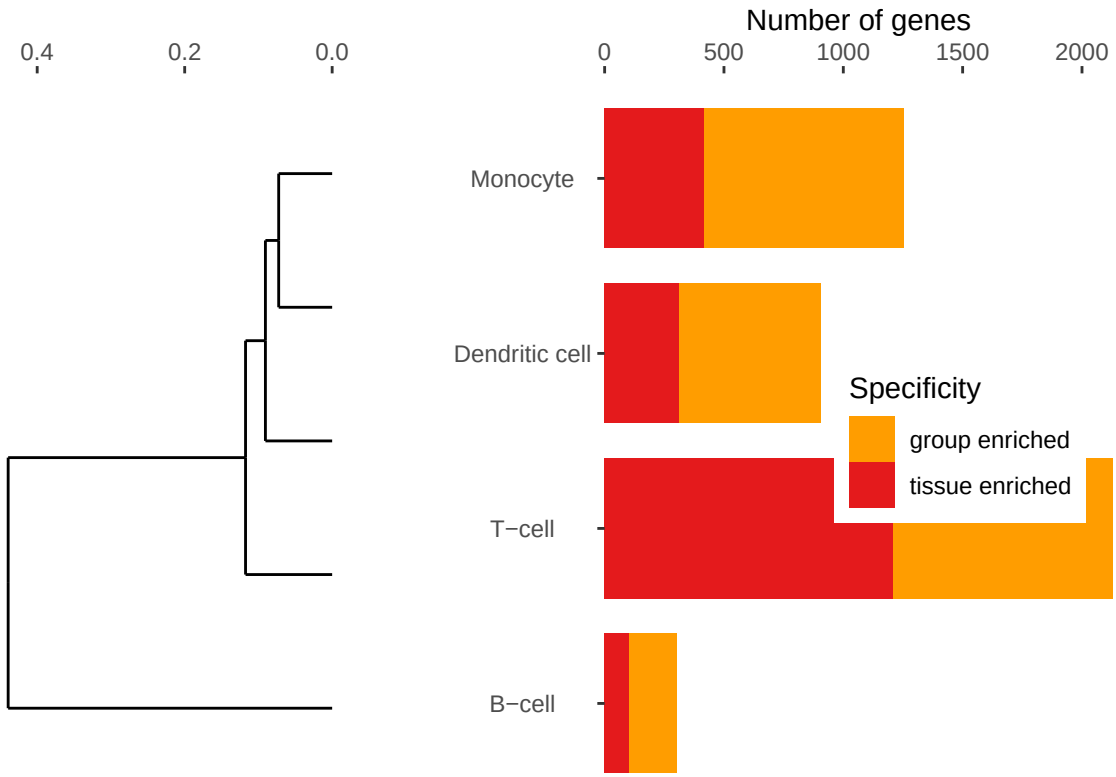
```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```




```
ggsave(savepath("single class comb n_genes alluvial.pdf"), width = 4, height = 6, useDingbats = F)


class_table_temp <-
  classification_max_norm_count %>%
```

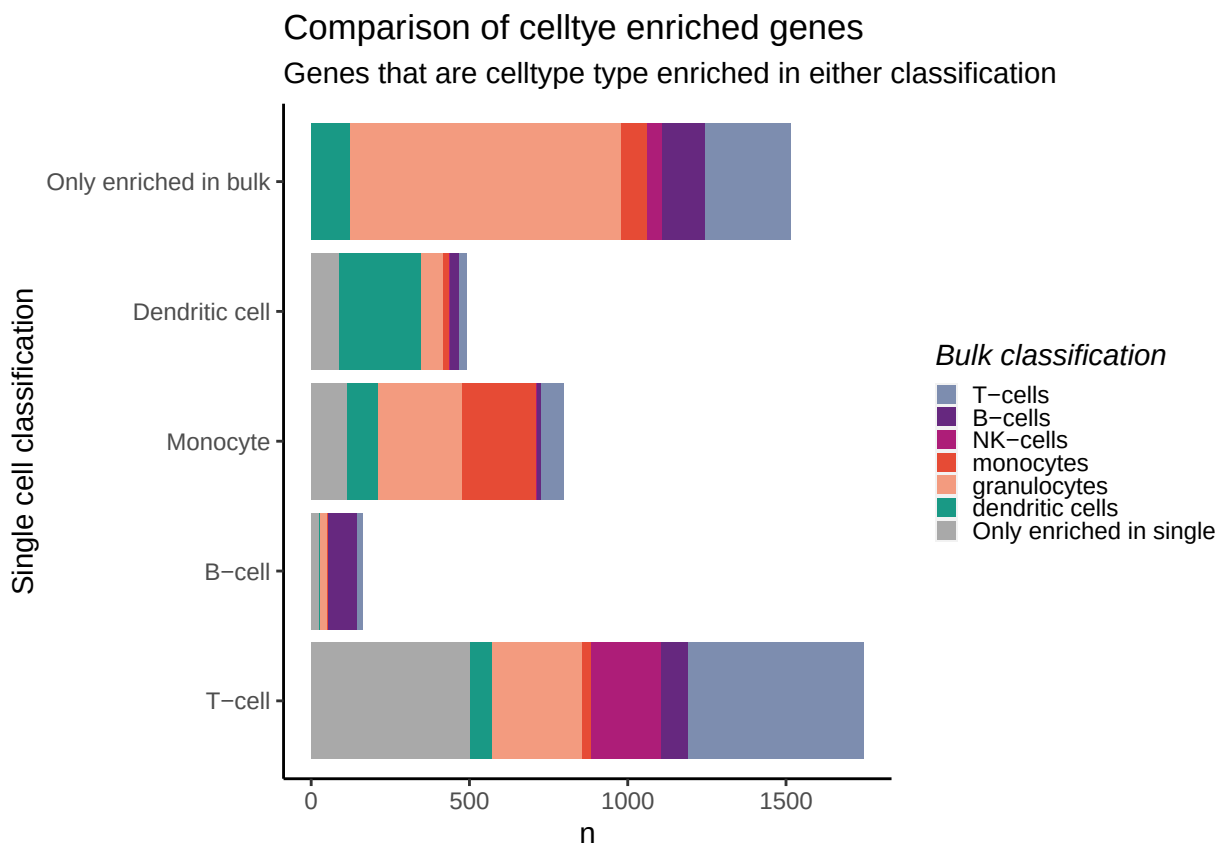
```
  select(gene, spec_category, enriched_tissues) %>%
  separate_rows(enriched_tissues, sep = ";") %>%
  mutate(spec_category = factor(spec_category, levels = rev(spec_category_levels)),
         enriched_tissues = str_to_sentence(enriched_tissues))

plot_dendro <-
  celltype_max_norm_count %>%
  spread(celltype, norm_count) %>%
  column_to_rownames("ensg_id") %>%
  cor(method = "spearman") %>%
  {1 - .} %>%
  as.dist() %>%
  hclust(method = "average") %>%
  dendro_data()


dendro_plot_data <-
  left_join(plot_dendro$segments,
            plot_dendro$labels,
            by = c("x" = "x", "yend" = "y"))

left_plot <-
  dendro_plot_data %>%
  ggplot() +
  geom_segment(aes(x=y, y=x, xend=yend, yend=xend, group = label))+
  geom_rect(aes(xmin=0, ymin=x + 0.5,
                xmax=-0.02, ymax=xend - 0.5,
                fill = label),
            show.legend = F) +
  scale_color_manual(values = celltype_pal)+
  scale_fill_manual(values = celltype_pal)+
  scale_x_reverse(expand = expand_scale(mult = 0.25), position = "top")+

  theme(axis.text.y = element_blank(),
        axis.title = element_blank(),
        axis.ticks.y = element_blank(),
        plot.margin = unit(c(1,1,1,1), units = "mm"),
        panel.background = element_blank())

right_plot <-
  class_table_temp %>%
  filter(!is.na(enriched_tissues)) %>%
  group_by(enriched_tissues, spec_category) %>%
  summarise(n_genes = n()) %>%
  ungroup() %>%
  mutate(enriched_tissues = factor(enriched_tissues, levels = plot_dendro$labels$label)) %>%
  ggplot(aes(enriched_tissues, n_genes, fill = spec_category)) +
  geom_col(width = 0.8, size = 0.1) +
  simple_theme +
  scale_fill_manual(values = gene_category_pal, name = "Specificity") +
  coord_flip() +
  xlab("Tissue") +
  ylab("Number of genes") +
```

```
  scale_y_continuous(position = "bottom", expand = c(0,0)) +

  theme(axis.text.y = element_text(hjust = 0.5),
        legend.position = c(0.7, 0.5),
        axis.title.y = element_blank(),
        panel.border = element_blank())

left_plot + right_plot
```



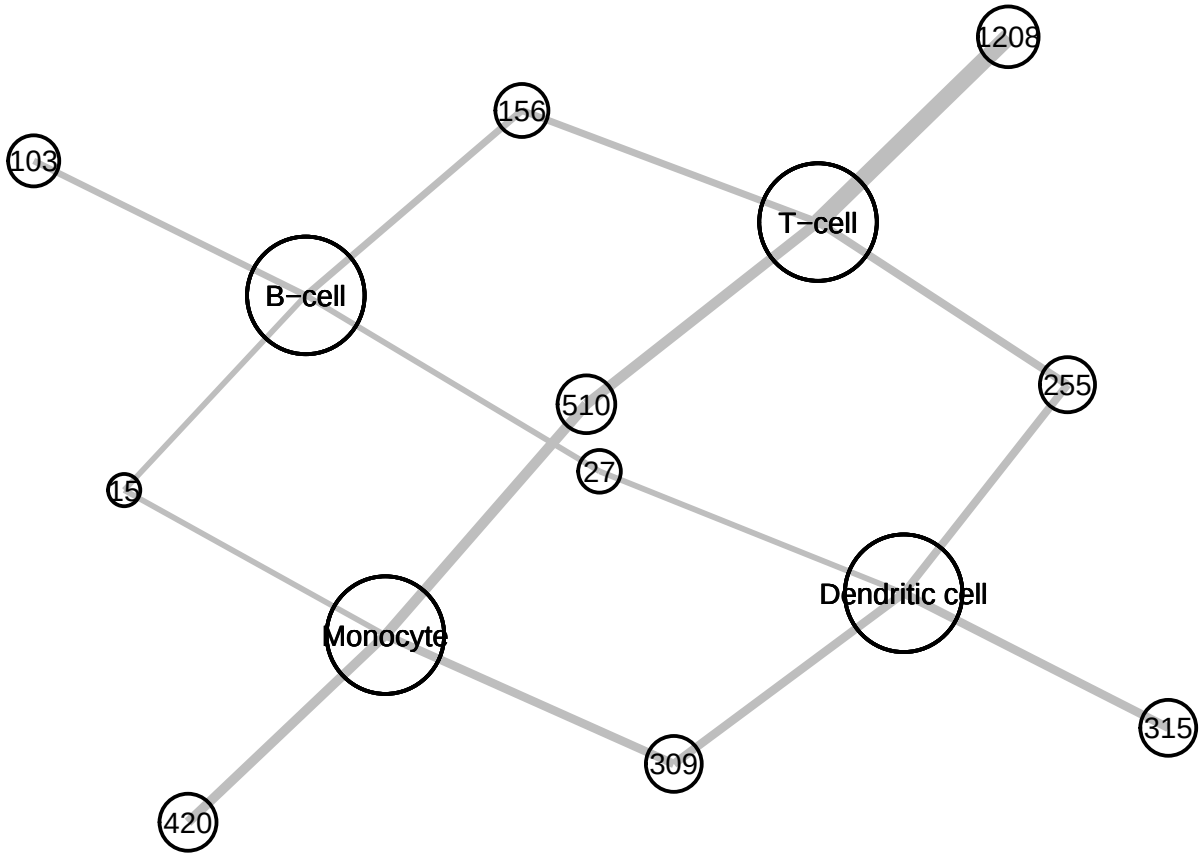

```
ggsave(savepath("N enr genes per tissue + dendro 2.pdf"), width = 5, height = 3)

classification_max_norm_count %>%
  left_join(gene_info92, by = c("gene" = "ensg_id")) %>%
  select(1, gene_name, 2, 3, enriched_tissues) %>%
  separate_rows(enriched_tissues, sep = ";") %>%

  left_join(blood_cell_category %>%
              separate_rows(enhanced_tissues, sep = ",") %>%
              select(1:3, enriched_tissues = enhanced_tissues),
            by = c("gene" = "ensg_id"),
            suffix = c("_singe_cell", "_blood_class")) %>%
  filter(spec_category == "tissue enriched" | specificity_category == "Tissue enriched") %>%
  group_by(enriched_tissues_singe_cell, enriched_tissues_blood_class) %>%
  summarise(n = n()) %>%
  ungroup() %>%
  mutate(enriched_tissues_singe_cell = ifelse(is.na(enriched_tissues_singe_cell),
```

```
                                          "Only enriched in bulk",
                                          enriched_tissues_singe_cell),
       enriched_tissues_blood_class = ifelse(is.na(enriched_tissues_blood_class),
                                          "Only enriched in single",
                                          enriched_tissues_blood_class)) %>%
mutate(enriched_tissues_blood_class = factor(enriched_tissues_blood_class,
                                           levels = c("T-cells",
                                                      "B-cells",
                                                      "NK-cells",
                                                      "monocytes",
                                                      "granulocytes",
                                                      "dendritic cells",
                                                      "Only enriched in single")),
       enriched_tissues_singe_cell = factor(enriched_tissues_singe_cell,
                                           levels = c("T-cell",
                                                      "B-cell",
                                                      "Monocyte",
                                                      "Dendritic cell",
                                                      "Only enriched in bulk"))) %>%
ggplot(aes(enriched_tissues_singe_cell, n, fill = enriched_tissues_blood_class)) +
geom_col() +
scale_fill_manual(values = c(tissue_colors, "Only enriched in single" = "darkgray"), name = "Bulk clas
ggtitle("Comparison of celltye enriched genes", "Genes that are celltype type enriched in either class
xlab("Single cell classification") +
stripped_theme +
coord_flip()
```



```
ggsave(savepath("N enriched genes bulk - single comparison.pdf"), width = 5, height = 3)

classification_max_norm_count %>%
  left_join(gene_info92, by = c("gene" = "ensg_id")) %>%
  select(1, gene_name, 2, 3, 4, enriched_tissues, tissues_detected) %>%
  filter(gene_name %in% c("CD3D",
                          "CD3E",
                          "CD19"))
```

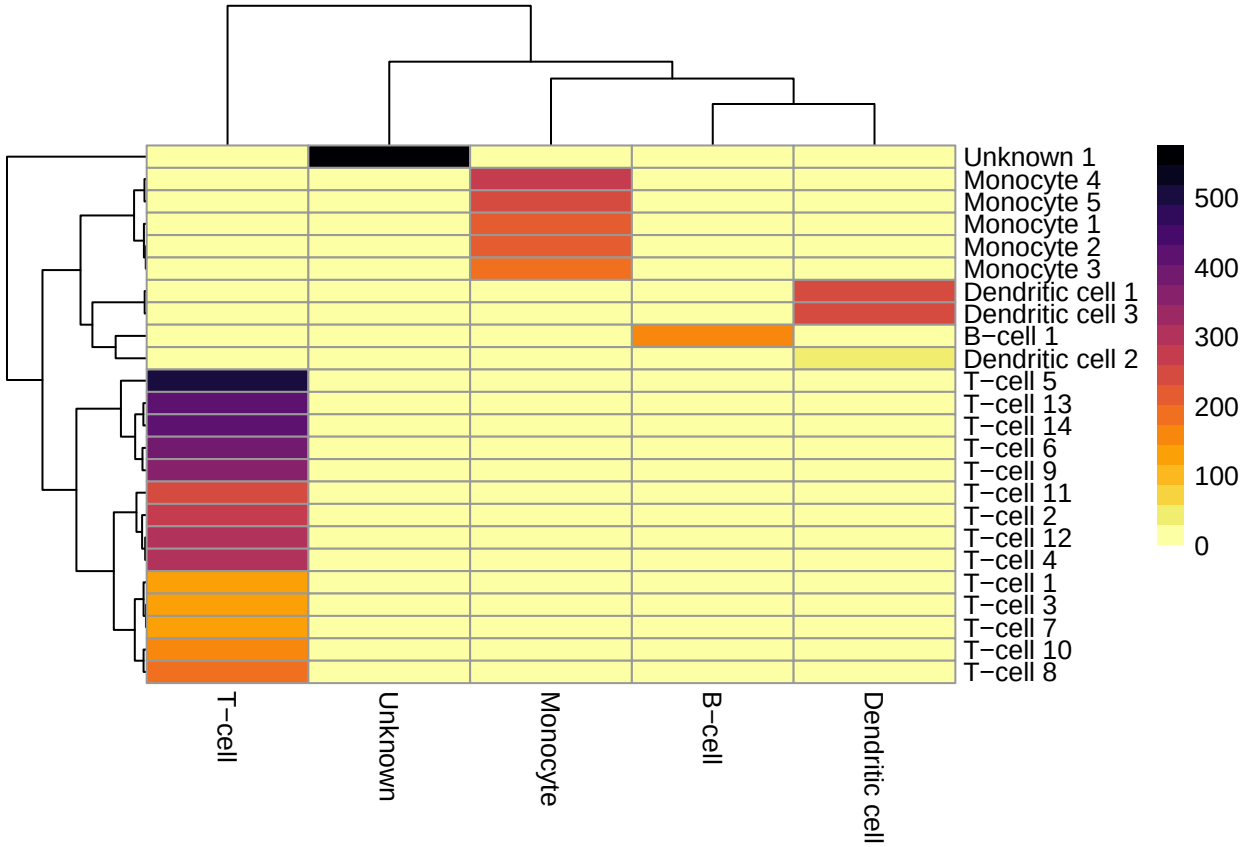
```
## # A tibble: 3 x 7
##   gene  gene_name spec_category dist_category spec_score enriched_tissues
##   <chr> <chr>     <chr>         <chr>              <dbl> <chr>
## 1 ENSG~ CD3D      tissue enric~ detected in ~       7.58 T-cell
## 2 ENSG~ CD19      tissue enric~ detected in ~       9.50 B-cell
## 3 ENSG~ CD3E      tissue enric~ detected in ~       6.81 T-cell
## # ... with 1 more variable: tissues_detected <chr>
```



Multisample classification

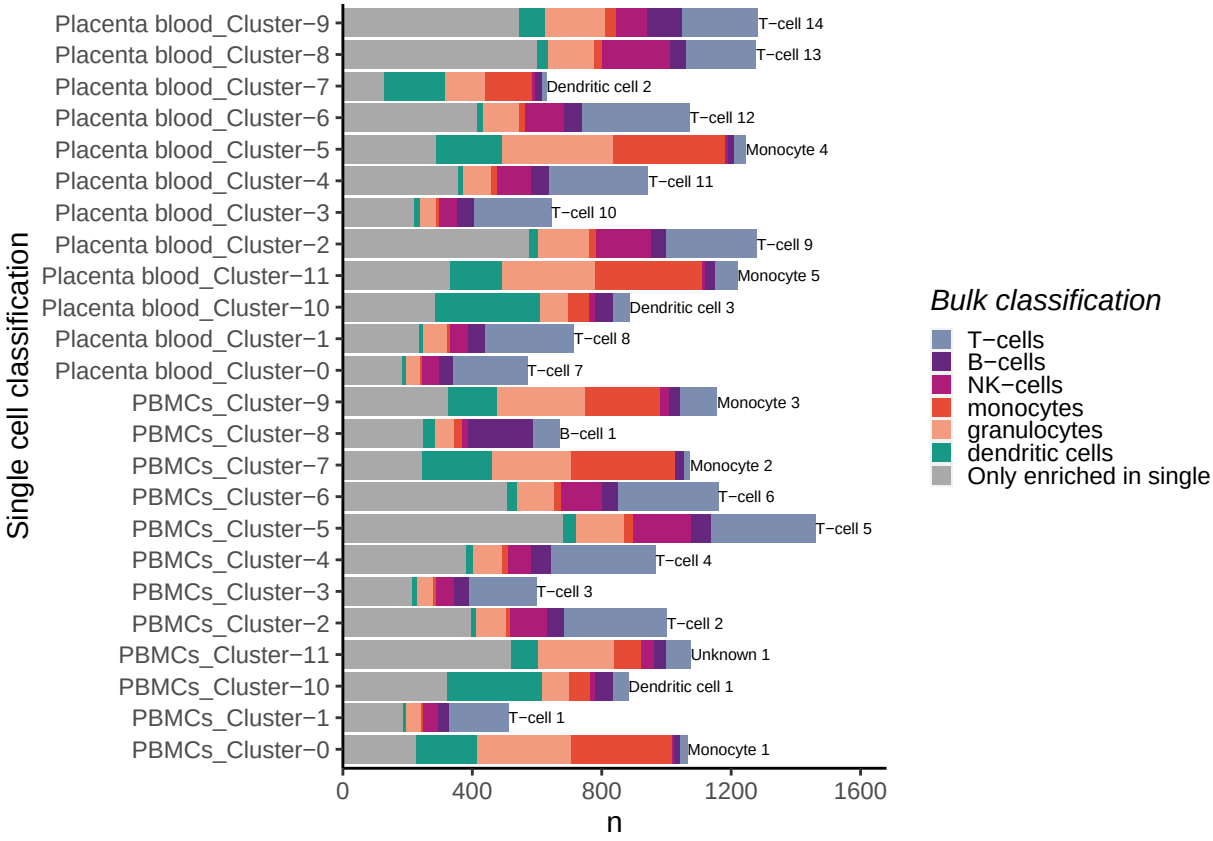
```
classification_sep_norm_count %>%
  filter(spec_category %in% c("tissue enriched")) %>%
  select(gene, spec_category, enriched_tissues, enriched_samples) %>%

  separate_rows(enriched_samples, sep = ";") %>%
  separate_rows(enriched_tissues, sep = ";") %>%
```

```
  group_by(enriched_samples, enriched_tissues) %>%
  summarise(n = n()) %>%
  ungroup() %>%
  select(1:3) %>%
  spread(enriched_tissues, n, fill = 0) %>%
  column_to_rownames("enriched_samples") %>%
  pheatmap(color = heatmap_palette)
```




```
classification_sep_norm_count %>%
  left_join(blood_cell_category,
            by = c("gene" = "ensg_id")) %>%
  filter(spec_category %in% c("tissue enriched", "group enriched")) %>%
  select(gene, spec_category, enriched_samples, enhanced_tissues) %>%

  separate_rows(enriched_samples, sep = ";") %>%
  separate_rows(enhanced_tissues, sep = ",") %>%
  group_by(spec_category, enriched_samples, enhanced_tissues) %>%
  summarise(n = n()) %>%
  ungroup() %>%
  filter(!is.na(enriched_samples)) %>%

  mutate(enhanced_tissues = ifelse(is.na(enhanced_tissues),
                                   "Only enriched in single",
                                   enhanced_tissues)) %>%
  mutate(enhanced_tissues = factor(enhanced_tissues,
                                   levels = c("T-cells",
```

```
                                          "B-cells",
                                          "NK-cells",
                                          "monocytes",
                                          "granulocytes",
                                          "dendritic cells",
                                          "Only enriched in single"))) %>%
  left_join(cluster_annotation,
            by = c("enriched_samples" = "unique_cluster_id")) %>%
  ggplot(aes(cluster, n, fill = enhanced_tissues)) +
  geom_col() +
  geom_text(data = . %>%
              group_by(cluster, enriched_samples) %>%
              summarise(n = sum(n)),
            aes(cluster, n,
                label = enriched_samples),
            inherit.aes = F,
            hjust = 0,
            size = 2) +
  scale_fill_manual(values = c(tissue_colors, "Only enriched in single" = "darkgray"), name = "Bulk cla
  xlab("Single cell classification") +
  stripped_theme +
  coord_flip() +
  scale_y_continuous(expand = expand_scale(c(0,0.15)))
```



```
ggsave(savepath("N enriched genes bulk - single sample comparison.pdf"), width = 7, height = 5)
```