02_TNBC_Phenotype_Distribution

2022-11-18

Contents

Overview 1
Libraries 1
Read CSV 1
Phenotype-Sample distribution 2
Phenotype-cluster distribution 6
Reference 11

Overview

This document aims to analyze the phenotype distribution with `sample_id` and `cluster_id` in the `TNBC` dataset. The analysis will include following steps:

1. Count the number of cells for each cell type in each sample, then draw histogram to determine the cell count interval. Construct a heatmap of phenotype distribution with sample number.
2. Inverstigate the cell type information for each cluster number by FlowSOM and their proportion, and construct a heatmap of phenotype distribution with cluster number

Libraries

```
library(readr)
library(dplyr)
library(ggplot2)
library(RColorBrewer)
```

Read CSV

Read in the MATLAB revised TNBC output CSV files.

```
TNBC <- readr::read_csv("/Users/henzhwang/Desktop/TNBC_training/MIBI-TNBC_scdata_counts_mm_matlab_revise
```

```
## Rows: 179194 Columns: 64
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): SITE_02, RECURRENCE_LABEL, mm
```

```
## dbl (61): sample_id, patient_id, AGE_AT_DX, STAGE, LATERAL, GRADE, Survival_...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Phenotype-Sample distribution

We want to construct a heatmap of the phenotype distribution of sample number and cell types.

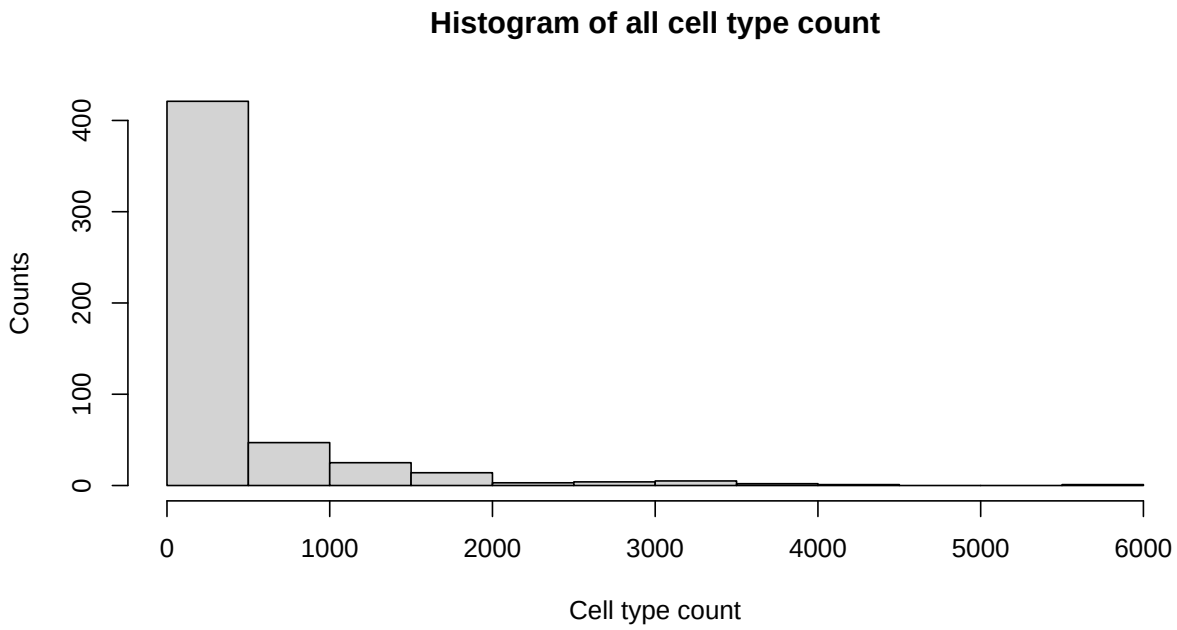
Cell type count histogram

To perform this, we first need to obtain the cell type count in each sample. Then, construct histogram to observe the distribution of cell type counts for all sample in order to determine the cut-off values for further analysis.

```
# cell type count in each sample
cell_count <- dplyr::count(TNBC, sample_id, mm) %>%
  dplyr::rename(count = n)
head(cell_count)
```

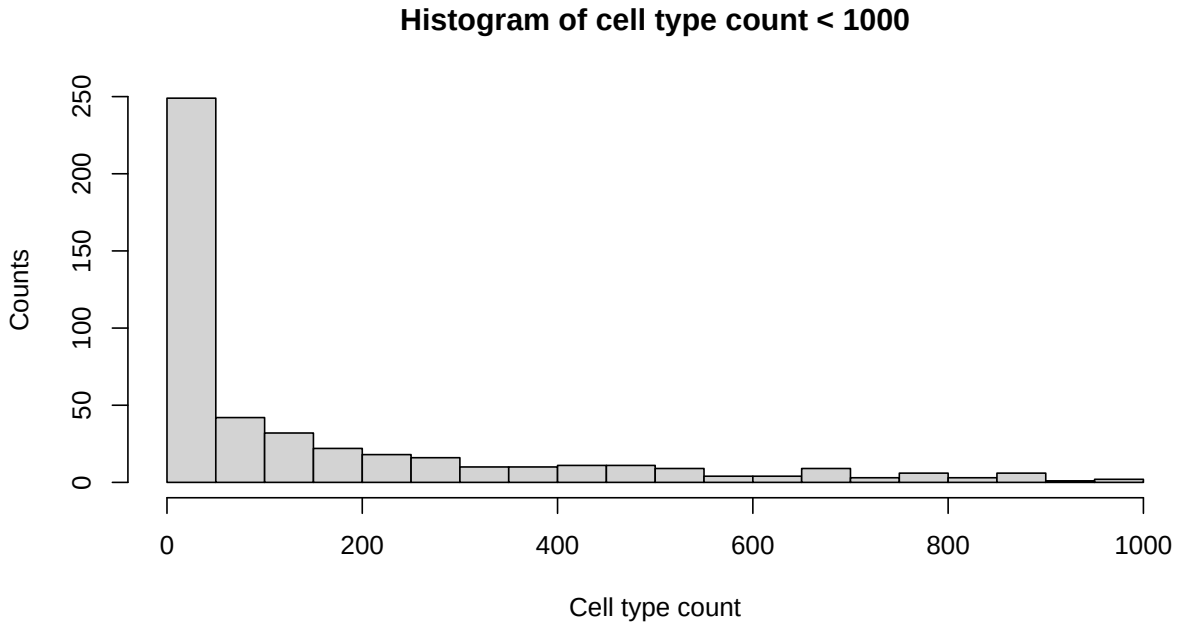
```
## # A tibble: 6 x 3
##   sample_id mm         count
##       <dbl> <chr>      <int>
## 1         1 B            734
## 2         1 CD4 T        152
## 3         1 CD8 T        147
## 4         1 DC             1
## 5         1 Epithelial    20
## 6         1 Mac           10
```

```
# histogram 1(all count)
hist(cell_count$count,
     main = "Histogram of all cell type count",
     xlab = "Cell type count", ylab = "Counts")
```



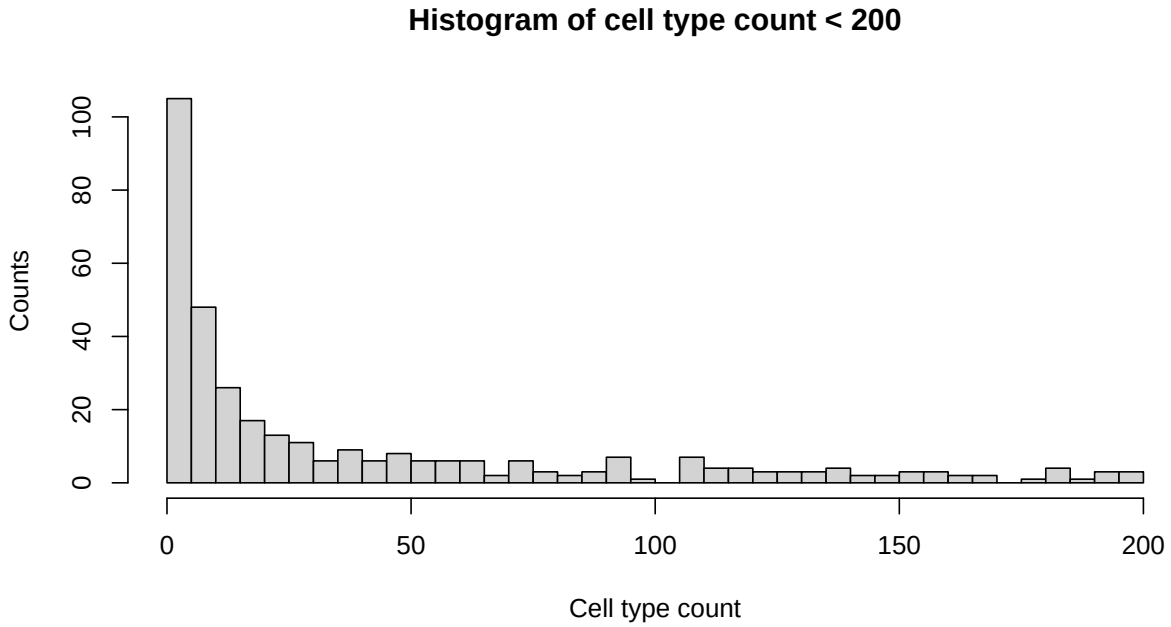
We notice that majority of the distribution is less then 1000 count, we want to draw a new histogram to observe closer. We now draw the histrogram for count up to 1000.

```
# histogram 2(count up to 1000)
hist(cell_count$count[cell_count$count < 1000],
     main = "Histogram of cell type count < 1000",
     xlab = "Cell type count", ylab = "Counts",
     breaks = 20)
```



We notice again that majority of this new distribution is less than 200, we now draw another histogram for cell type count up to 200.

```
# histogram 2(count up to 200)
hist(cell_count$count[cell_count$count < 200],
     main = "Histogram of cell type count < 200",
     xlab = "Cell type count", ylab = "Counts",
     breaks = 40)
```



Phenotype-sample distribution heatmap

By the above histograms, we can clearly see most of the cell type count are less than 50, and thus we have conentrate our cut-off value on $0-50$. We will have 11 value interval to be plotted in the distribution heatmap, and the value interval are 0-1, 1-5, 5-10, 10-25, 25-50, 50-100, 100-250, 250-500, 500-1000, 1000-2000, and greater than 2000.

```
# adding additional 0 to sample_id 1-9
TNBC$sample_id[TNBC$sample_id < 10] <- paste0("0", TNBC$sample_id[TNBC$sample_id < 10])
```

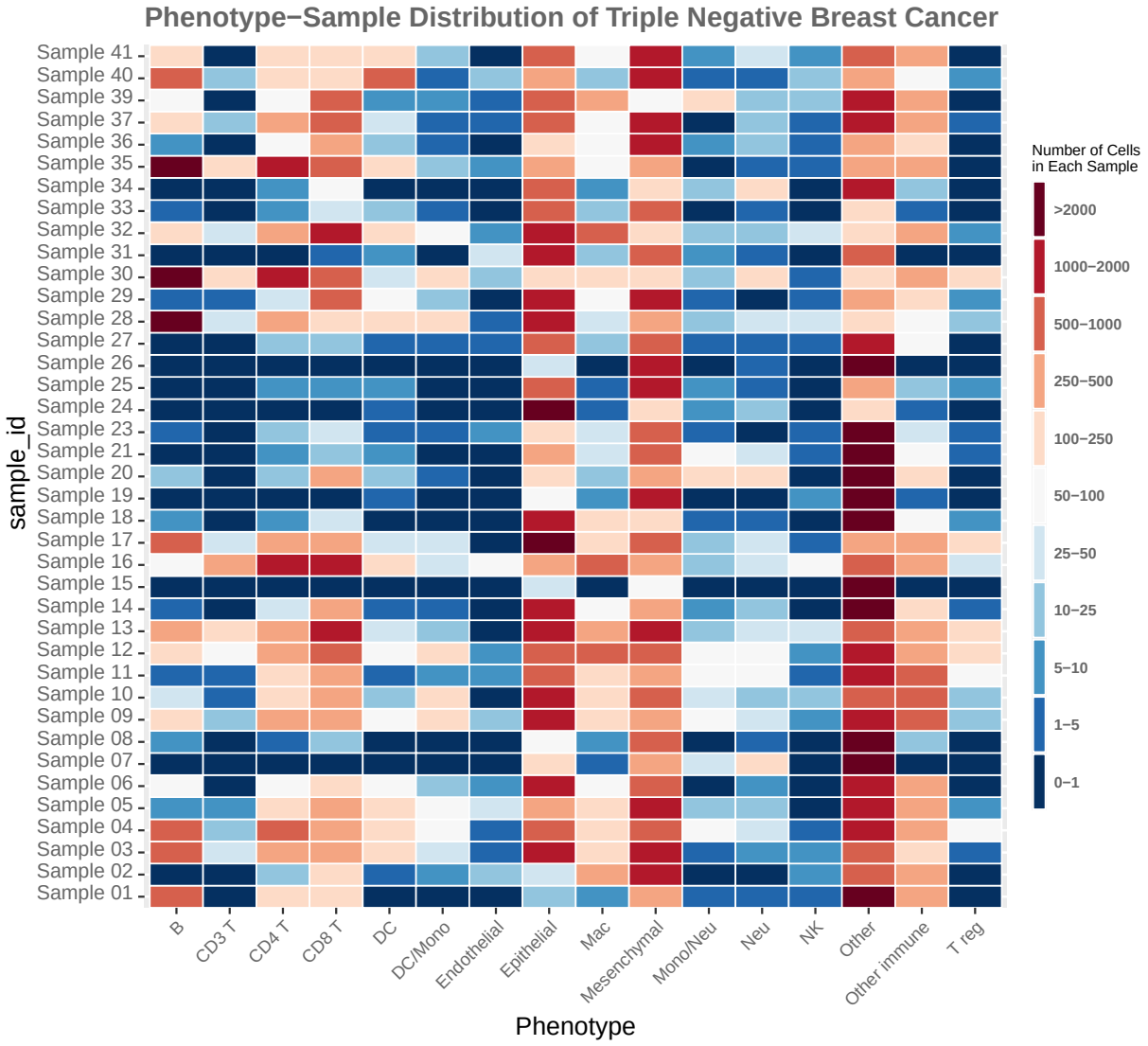
```
# phenotype-sample distribution heatmap
TNBC %>%
  with(table(sample_id, mm)) %>%
  as.data.frame() %>%
  mutate(sample_id = paste("Sample", sample_id)) %>%
  mutate(count = cut(as.numeric(Freq), breaks = c(-1, 1.1, 5.1, 10.1,
                                    25.1, 50.1, 100.1, 250.1, 500.1, 1000.1, 2000.1, max(Freq, na.rm = T
                     labels = c("0-1", "1-5", "5-10", "10-25", "25-50",
                                "50-100", "100-250", "250-500", "500-1000", "1000-2000", ">2000"))) %>%
  mutate(count = factor(as.character(count), levels = rev(levels(count)))) %>%
```

```r
ggplot(mapping = aes(x = mm, y = sample_id, fill = count)) +
  geom_tile(colour = "white", linewidth = 0.3) +

  # labels
  guides(fill=guide_legend(title = "Number of Cells \nin Each Sample"))+
  labs(x = "Phenotype", y = " sample_id", title = "Phenotype-Sample Distribution of Triple Negative Br
  scale_y_discrete(expand = c(0, 0)) +
  scale_x_discrete(expand = c(0, 0)) +

  # color palette
  #scale_fill_manual(values = rev(brewer.pal(11, "RdGy")), na.value = "azure4") + #RdGy RdBu
  scale_fill_manual(values = brewer.pal(11, "RdBu"), na.value = "azure4") +

  # theme
  theme_grey(base_size = 10)+
  theme(legend.position = "right", legend.direction = "vertical",
      legend.title = element_text(colour = "black", size = 6),
      legend.margin = margin(grid::unit(0, "cm")),
      legend.text = element_text(colour = "grey40", size = 6, face = "bold"),
      legend.key.height = grid::unit(0.8, "cm"),
      legend.key.width = grid::unit(0.2, "cm"),
      axis.text.x = element_text(size = 7, colour = "grey40"),
      axis.text.y = element_text(vjust = 0.2, colour = "grey40"),
      axis.ticks = element_line(linewidth = 0.4),
      plot.background = element_blank(),
      panel.border = element_blank(),
      plot.margin = margin(0.7, 0.4, 0.1, 0.2, "cm"),
      plot.title = element_text(colour = "grey40", hjust = 0, size = 11, face = "bold")
   )+
  #scale_x_discrete(guide = guide_axis(n.dodge = 2)) # separate x label into two levels
  scale_x_discrete(guide = guide_axis(angle = 45))  # rotating x label angle
```



Phenotype-cluster distribution

We want to construct a heatmap of the phenotype distribution of cluster number by FlowSOM and cell types.

Cluster_id Diagnostics

To perform this, we first want to investigate the count of cell type in each cluster and their correponding proportion. There are 113 clusters present in the dataset where there are at most different cell types in each cluster. The following table shows information about the clusters such as cluster number, cell types in the cluster, and the proportion of cell types. The proportion is able to indicate if the clusters are dominated by one cell type and possibly implies the accuracy of the FlowSOM clustering technique.

Variables used in the tables are:

- cluster_id: Cluster number

- cell_n: Number of cells in the cluster
- n_ct: Number of unique cell types in the cluster
- ct1, ct2, ct3: Cell type
- ct1_n, ct2_n, ct3_n: Number of corresponding cell types in the cluster
- ct1_f, ct2_f, ct3_f: Proportion of corresponding cell type in the cluster

```
# number of cell_type for each cluster_id
TNBC %>%
  group_by(cluster_id) %>%
  summarise(cell_n = n(),
            n_ct = n_distinct(mm), # mm and cell_type are the same
            ct1 = as.character(as.list(unique(mm))[1]),
            ct1_n = sum(mm == as.character(as.list(unique(mm))[1])),
            ct2 = as.character(as.list(unique(mm))[2]),
            ct2_n = sum(mm == as.character(as.list(unique(mm))[2])),
            ct3 = as.character(as.list(unique(mm))[3]),
            ct3_n = sum(mm==as.character(as.list(unique(mm))[3]))) %>%
  mutate(ct1_f = formattable::percent(ct1_n / cell_n),
         ct2_f = formattable::percent(ct2_n / cell_n),
         ct3_f = formattable::percent(ct3_n / cell_n)) %>%
  dplyr::relocate(ct1_f, .after = ct1_n) %>%
  dplyr::relocate(ct2_f, .after = ct2_n) %>%
  dplyr::relocate(ct3_f, .after = ct3_n) %>%
  knitr::kable()
```

cluster_id	cell_n	n_ct	ct1	ct1_n	ct1_f	ct2	ct2_n	ct2_f	ct3	ct3_n	ct3_f
1	1034	2	B	309	29.88%	Epithelial	725	70.12%	NULL	0	0.00%
2	1450	2	B	725	50.00%	Epithelial	725	50.00%	NULL	0	0.00%
3	1991	2	B	913	45.86%	Other	1078	54.14%	NULL	0	0.00%
4	1403	2	B	676	48.18%	Other	727	51.82%	NULL	0	0.00%
5	2334	2	DC/Mono	341	14.61%	Other	1993	85.39%	NULL	0	0.00%
6	3148	2	B	438	13.91%	Other	2710	86.09%	NULL	0	0.00%
7	3057	2	CD4 T	339	11.09%	Other	2718	88.91%	NULL	0	0.00%
8	906	2	CD3 T	391	43.16%	Other	515	56.84%	NULL	0	0.00%
9	1973	2	CD4 T	227	11.51%	Other	1746	88.49%	NULL	0	0.00%
10	783	2	CD4 T	244	31.16%	Mesenchymal	539	68.84%	NULL	0	0.00%
11	2140	2	B	843	39.39%	Epithelial	1297	60.61%	NULL	0	0.00%
12	2301	2	B	1372	59.63%	Epithelial	929	40.37%	NULL	0	0.00%
13	1681	2	B	499	29.68%	Epithelial	1182	70.32%	NULL	0	0.00%
14	2437	2	B	749	30.73%	Other	1688	69.27%	NULL	0	0.00%
15	1388	2	B	439	31.63%	Other	949	68.37%	NULL	0	0.00%
16	2238	2	B	646	28.87%	Other	1592	71.13%	NULL	0	0.00%
17	1879	2	B	645	34.33%	Other	1234	65.67%	NULL	0	0.00%
18	1287	2	CD4 T	306	23.78%	Other	981	76.22%	NULL	0	0.00%
19	1642	2	CD4 T	599	36.48%	Mesenchymal	1043	63.52%	NULL	0	0.00%
20	1924	2	CD4 T	514	26.72%	Mesenchymal	1410	73.28%	NULL	0	0.00%
21	2023	2	B	1119	55.31%	Epithelial	904	44.69%	NULL	0	0.00%
22	2483	2	B	938	37.78%	Epithelial	1545	62.22%	NULL	0	0.00%
23	2494	2	B	966	38.73%	Other	1528	61.27%	NULL	0	0.00%
24	1985	2	CD8 T	533	26.85%	Other	1452	73.15%	NULL	0	0.00%
25	1212	2	B	552	45.54%	Other	660	54.46%	NULL	0	0.00%
26	1920	2	B	704	36.67%	Other	1216	63.33%	NULL	0	0.00%

cluster_id	cell_n	n_ct	ct1	ct1_n	ct1_f	ct2	ct2_n	ct2_f	ct3	ct3_n	ct3_f
27	1585	2	CD4 T	609	38.42%	Epithelial	976	61.58%	NULL	0	0.00%
28	1579	2	CD4 T	519	32.87%	Other	1060	67.13%	NULL	0	0.00%
29	1885	2	CD4 T	787	41.75%	Mesenchymal	1098	58.25%	NULL	0	0.00%
30	1859	2	CD3 T	394	21.19%	Mesenchymal	1465	78.81%	NULL	0	0.00%
31	1087	2	B	719	66.15%	Epithelial	368	33.85%	NULL	0	0.00%
32	1975	2	B	1060	53.67%	Epithelial	915	46.33%	NULL	0	0.00%
33	1856	2	B	534	28.77%	Other	1322	71.23%	NULL	0	0.00%
34	1889	2	B	907	48.01%	Other	982	51.99%	NULL	0	0.00%
35	1564	2	CD8 T	593	37.92%	Other	971	62.08%	NULL	0	0.00%
36	1799	2	CD4 T	497	27.63%	Epithelial	1302	72.37%	NULL	0	0.00%
37	1208	2	CD4 T	609	50.41%	Other	599	49.59%	NULL	0	0.00%
38	1096	2	Mesenchymal	672	61.31%	T reg	424	38.69%	NULL	0	0.00%
39	2082	2	CD4 T	1024	49.18%	Mesenchymal	1058	50.82%	NULL	0	0.00%
40	1769	2	CD4 T	584	33.01%	Mesenchymal	1185	66.99%	NULL	0	0.00%
41	1387	2	Epithelial	1130	81.47%	Neu	257	18.53%	NULL	0	0.00%
42	1147	2	B	792	69.05%	Epithelial	355	30.95%	NULL	0	0.00%
43	1933	2	DC	439	22.71%	Epithelial	1494	77.29%	NULL	0	0.00%
44	1437	2	B	539	37.51%	Other	898	62.49%	NULL	0	0.00%
45	1257	2	Other	876	69.69%	Other immune	381	30.31%	NULL	0	0.00%
46	927	2	Epithelial	457	49.30%	Other immune	470	50.70%	NULL	0	0.00%
47	1120	2	Epithelial	681	60.80%	T reg	439	39.20%	NULL	0	0.00%
48	1857	2	CD4 T	855	46.04%	Epithelial	1002	53.96%	NULL	0	0.00%
49	1860	2	CD4 T	863	46.40%	Mesenchymal	997	53.60%	NULL	0	0.00%
50	1066	2	Mesenchymal	811	76.08%	Other immune	255	23.92%	NULL	0	0.00%
51	966	2	Epithelial	605	62.63%	Neu	361	37.37%	NULL	0	0.00%
52	1094	2	Epithelial	554	50.64%	Neu	540	49.36%	NULL	0	0.00%
53	1006	2	DC	288	28.63%	Epithelial	718	71.37%	NULL	0	0.00%
54	1537	2	CD4 T	514	33.44%	Other	1023	66.56%	NULL	0	0.00%
55	1148	2	Other	446	38.85%	Other immune	702	61.15%	NULL	0	0.00%
56	1803	2	Mesenchymal	515	28.56%	Other immune	1288	71.44%	NULL	0	0.00%
57	1848	2	Other	617	33.39%	Other immune	1231	66.61%	NULL	0	0.00%
58	1404	2	CD4 T	828	58.97%	Epithelial	576	41.03%	NULL	0	0.00%
59	981	2	CD8 T	393	40.06%	Mesenchymal	588	59.94%	NULL	0	0.00%
60	907	2	CD3 T	350	38.59%	Mesenchymal	557	61.41%	NULL	0	0.00%
61	1268	2	Epithelial	804	63.41%	Mono/Neu	464	36.59%	NULL	0	0.00%
62	858	2	Epithelial	487	56.76%	Mono/Neu	371	43.24%	NULL	0	0.00%
63	813	2	Epithelial	311	38.25%	Mac	502	61.75%	NULL	0	0.00%
64	1532	2	Epithelial	836	54.57%	Mac	696	45.43%	NULL	0	0.00%
65	1046	2	DC	676	64.63%	Mesenchymal	370	35.37%	NULL	0	0.00%
66	1096	2	Mesenchymal	511	46.62%	Other immune	585	53.38%	NULL	0	0.00%
67	1240	2	Epithelial	449	36.21%	Other immune	791	63.79%	NULL	0	0.00%
68	1706	2	CD8 T	686	40.21%	Mesenchymal	1020	59.79%	NULL	0	0.00%
69	1055	2	CD8 T	573	54.31%	Mesenchymal	482	45.69%	NULL	0	0.00%
70	713	2	CD8 T	434	60.87%	Endothelial	279	39.13%	NULL	0	0.00%

cluster_id	cell_n	n_ct	ct1	ct1_n	ct1_f	ct2	ct2_n	ct2_f	ct3	ct3_n	ct3_f
71	1087	2	Epithelial	622	57.22%	Mac	465	42.78%	NULL	0	0.00%
72	1201	2	Epithelial	994	82.76%	Neu	207	17.24%	NULL	0	0.00%
73	960	2	Epithelial	393	40.94%	Mac	567	59.06%	NULL	0	0.00%
74	1144	2	DC/Mono	425	37.15%	Epithelial	719	62.85%	NULL	0	0.00%
75	1644	2	DC	978	59.49%	Mesenchymal	666	40.51%	NULL	0	0.00%
76	899	2	Mesenchymal	614	68.30%	NK	285	31.70%	NULL	0	0.00%
77	2061	2	CD8 T	950	46.09%	Mesenchymal	1111	53.91%	NULL	0	0.00%
78	2523	2	CD8 T	955	37.85%	Mesenchymal	1568	62.15%	NULL	0	0.00%
79	1208	2	CD8 T	459	38.00%	Mesenchymal	749	62.00%	NULL	0	0.00%
80	918	2	CD8 T	577	62.85%	Mesenchymal	341	37.15%	NULL	0	0.00%
81	1273	2	DC/Mono	314	24.67%	Epithelial	959	75.33%	NULL	0	0.00%
82	845	2	Epithelial	400	47.34%	Mac	445	52.66%	NULL	0	0.00%
83	1714	3	Epithelial	524	30.57%	Mac	488	28.47%	Other	702	40.96%
84	1286	2	Epithelial	658	51.17%	Mac	628	48.83%	NULL	0	0.00%
85	1655	2	Epithelial	891	53.84%	Other immune	764	46.16%	NULL	0	0.00%
86	1797	2	Mesenchymal	908	50.53%	Other immune	889	49.47%	NULL	0	0.00%
87	2633	2	CD8 T	1184	44.97%	Mesenchymal	1449	55.03%	NULL	0	0.00%
88	2465	2	CD8 T	1193	48.40%	Mesenchymal	1272	51.60%	NULL	0	0.00%
89	2210	2	CD8 T	944	42.71%	Mesenchymal	1266	57.29%	NULL	0	0.00%
90	1106	2	CD8 T	565	51.08%	Epithelial	541	48.92%	NULL	0	0.00%
91	1428	2	DC/Mono	200	14.01%	Epithelial	1228	85.99%	NULL	0	0.00%
92	1658	2	Epithelial	1128	68.03%	Mac	530	31.97%	NULL	0	0.00%
93	846	2	Epithelial	571	67.49%	Mac	275	32.51%	NULL	0	0.00%
94	1326	2	Epithelial	370	27.90%	Mac	956	72.10%	NULL	0	0.00%
95	1297	2	Epithelial	546	42.10%	Other immune	751	57.90%	NULL	0	0.00%
96	1180	2	Mesenchymal	618	52.37%	Other immune	562	47.63%	NULL	0	0.00%
97	2019	3	CD8 T	976	48.34%	Mesenchymal	590	29.22%	Other	453	22.44%
98	2774	3	CD8 T	1115	40.19%	Mesenchymal	802	28.91%	Other	857	30.89%
99	2448	3	CD8 T	684	27.94%	Mesenchymal	683	27.90%	Other	1081	44.16%
100	1302	2	CD8 T	562	43.16%	Mesenchymal	740	56.84%	NULL	0	0.00%
101	817	1	Other	817	100.00%	NULL	0	0.00%	NULL	0	0.00%
113	910	1	Other	910	100.00%	NULL	0	0.00%	NULL	0	0.00%
114	1846	1	Other	1846	100.00%	NULL	0	0.00%	NULL	0	0.00%
128	1428	1	Other	1428	100.00%	NULL	0	0.00%	NULL	0	0.00%
129	2023	1	Other	2023	100.00%	NULL	0	0.00%	NULL	0	0.00%
143	2503	1	Other	2503	100.00%	NULL	0	0.00%	NULL	0	0.00%
157	2002	1	Other	2002	100.00%	NULL	0	0.00%	NULL	0	0.00%
158	1090	1	Other	1090	100.00%	NULL	0	0.00%	NULL	0	0.00%
159	1750	1	Other	1750	100.00%	NULL	0	0.00%	NULL	0	0.00%
172	971	1	Other	971	100.00%	NULL	0	0.00%	NULL	0	0.00%
175	978	1	Other	978	100.00%	NULL	0	0.00%	NULL	0	0.00%
187	1229	1	Other	1229	100.00%	NULL	0	0.00%	NULL	0	0.00%
211	4382	1	Other	4382	100.00%	NULL	0	0.00%	NULL	0	0.00%

```
# formattable::formattable(list(ct1_f = color_bar(color = "lightblue"),
#                               ct2_f = color_bar(color = "lightpink"),
#                               ct3_f = color_bar(color = "lightgreen"))) # for html
```

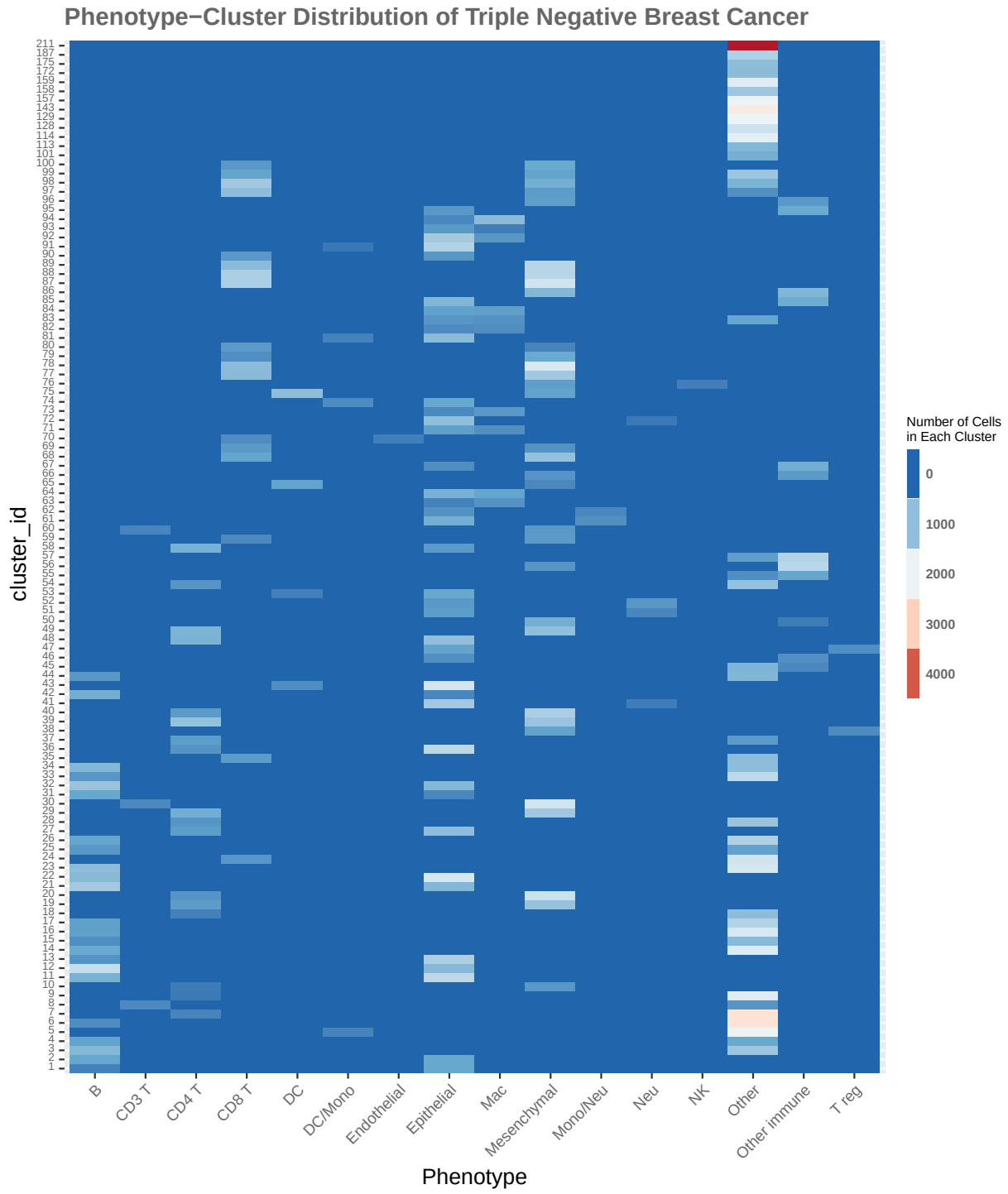
Phenotype-cluster distribution

Now we want to draw a heatmap of prototypes and cluster number.

```
as.data.frame(with(TNBC, table(cluster_id, mm))) %>%
  ggplot() +
    #geom_tile(aes(x = cluster_id, y = mm, fill = Freq)) +
    geom_tile(aes(x = mm, y = cluster_id, fill = Freq)) +

    # colour palettes
    scale_fill_distiller(palette = "RdBu", na.value = "azure4")+
    # labels
    guides(fill=guide_legend(title = "Number of Cells \nin Each Cluster"))+
    labs(x = "Phenotype", y = " cluster_id",
         title = "Phenotype-Cluster Distribution of Triple Negative Breast Cancer")+
    scale_y_discrete(expand = c(0, 0)) +
    scale_x_discrete(expand = c(0, 0)) +

    # theme
    theme_grey(base_size = 10)+
    theme(legend.position = "right", legend.direction = "vertical",
        legend.title = element_text(colour = "black", size = 6),
        legend.margin = margin(grid::unit(0, "cm")),
        legend.text = element_text(colour = "grey40", size = 6, face = "bold"),
        legend.key.height = grid::unit(0.8, "cm"),
        legend.key.width = grid::unit(0.2, "cm"),
        axis.text.x = element_text(size = 7, colour = "grey40"),
        axis.text.y = element_text(size = 5, vjust = 0.2, colour = "grey40"),
        axis.ticks = element_line(linewidth = 0.4),
        plot.background = element_blank(),
        panel.border = element_blank(),
        plot.margin = margin(0.7, 0.4, 0.1, 0.2, "cm"),
        plot.title = element_text(colour = "grey40", hjust = 0, size = 11, face = "bold")
     ) +
    scale_x_discrete(guide = guide_axis(angle = 45))
```



Reference

A guide to elegant tiled heatmaps in R

Formattable Vignettes