07b.stat.R

kjyi

Thu May 31 11:38:06 2018

```
if (F) {
    system("mkdir -p stat")
    rmarkdown::render("07b.stat.R",
                      output_format = "html_document",
                      clean = TRUE)
    rmarkdown::render("07b.stat.R",
                      output_format =
                        rmarkdown::pdf_document(
                            toc = TRUE,
                            latex_engine = 'xelatex',
                            pandoc_args =
                                c("--variable",
                                  "mainfont='NanumGothic'")),
                      clean = TRUE)
}

suppressMessages(library(tidyverse))
library(stringr)
library(gridExtra)

merged <- read_tsv("mutation_summary/merged.tsv")
```

```
## Parsed with column specification:
## cols(
##   id = col_character(),
##   pd1 = col_character(),
##   chr = col_character(),
##   pos = col_integer(),
##   REF = col_character(),
##   ALT = col_character(),
##   gene_name = col_character(),
##   type = col_character(),
##   ML.C = col_double(),
##   ML.M.SEGMENT = col_integer(),
##   ML.M = col_integer(),
##   M.SEGMENT.FLAGGED = col_logical(),
##   CN.SUBCLONAL = col_logical(),
##   CELLFRACTION = col_double(),
##   FLAGGED = col_logical()
## )
```

```
purity <- read_tsv("mutation_summary/purity_ploidy.tsv")
```

```
## Parsed with column specification:
## cols(
##   id = col_character(),
##   Purity = col_double(),
##   Ploidy = col_double(),
```

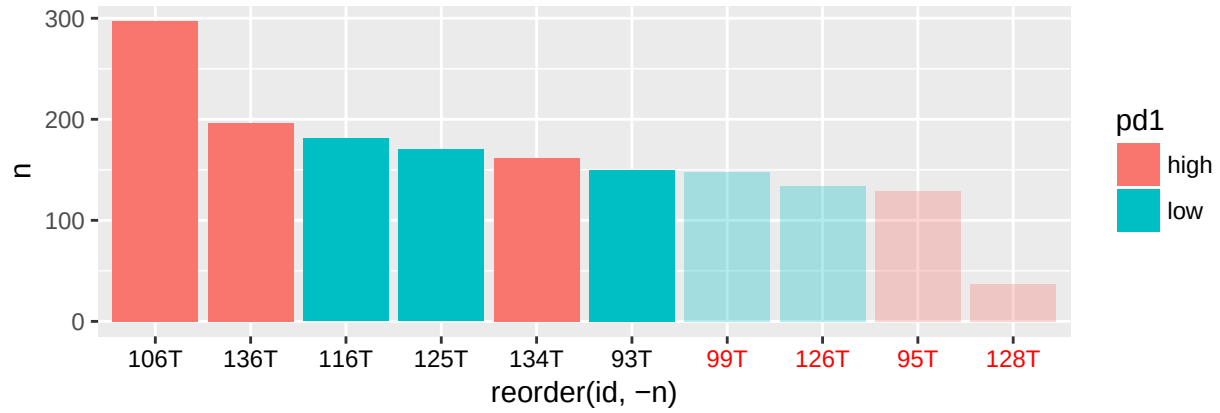
```
##   Sex = col_character(),
##   Contamination = col_integer(),
##   Flagged = col_logical(),
##   Failed = col_logical(),
##   Curated = col_logical(),
##   Comment = col_character()
## )
```

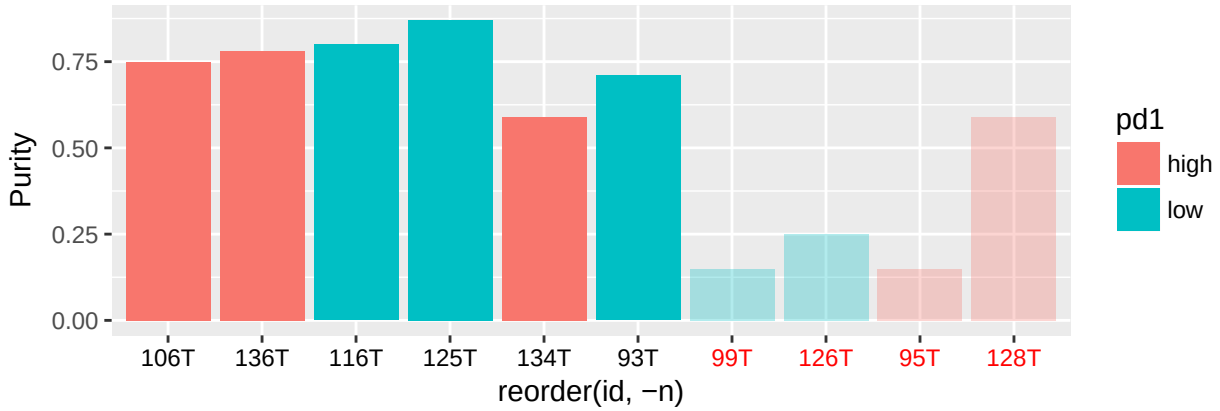
```
# total mutation count, purity, flag ------------------------------------------
count1 <- merged %>% group_by(pd1, id) %>% summarise(n = n()) %>%
    left_join(purity, by = "id") %>% arrange(pd1, -n)
color_vector <- c("red", "black")[c(2,2,2,2,2,2,1,1,1,1)]
count1
```

```
## # A tibble: 10 x 11
## # Groups:   pd1 [2]
##    pd1   id        n Purity Ploidy Sex   Contamination Flagged Failed
##    <chr> <chr> <int>  <dbl>  <dbl> <chr>         <int> <lgl>   <lgl>
##  1 high  106T    297   0.75   2.88 M                 0 FALSE   FALSE
##  2 high  136T    196   0.78   2.12 M                 0 FALSE   FALSE
##  3 high  134T    162   0.59   3.87 F                 0 FALSE   FALSE
##  4 high  95T     129   0.15   2.08 M                 0 TRUE    FALSE
##  5 high  128T     37   0.59   2.01 F                 0 TRUE    FALSE
##  6 low   116T    181   0.8    1.87 M                 0 FALSE   FALSE
##  7 low   125T    170   0.87   2.11 M                 0 FALSE   FALSE
##  8 low   93T     150   0.71   3.76 M                 0 FALSE   FALSE
##  9 low   99T     148   0.15   2.20 M                 0 TRUE    FALSE
## 10 low   126T    134   0.25   2.06 M                 0 TRUE    FALSE
## # ... with 2 more variables: Curated <lgl>, Comment <chr>
```

```
p1 <- count1 %>%
    bind_cols(alpha_vector = ifelse(is.na(count1$Comment), 1, 0.3)) %>%
    ggplot(aes(x = reorder(id, -n), y = n, fill = pd1, alpha = I(alpha_vector))) +
    geom_col() +
    theme(axis.text.x = element_text(colour = color_vector))

p2 <- count1 %>%
    bind_cols(alpha_vector = ifelse(is.na(count1$Comment), 1, 0.3)) %>%
    ggplot(aes(x = reorder(id, -n), y = Purity, fill = pd1, alpha = I(alpha_vector))) +
    geom_col() +
    theme(axis.text.x = element_text(colour = color_vector))
grid.arrange(p1, p2, nrow = 2)
```





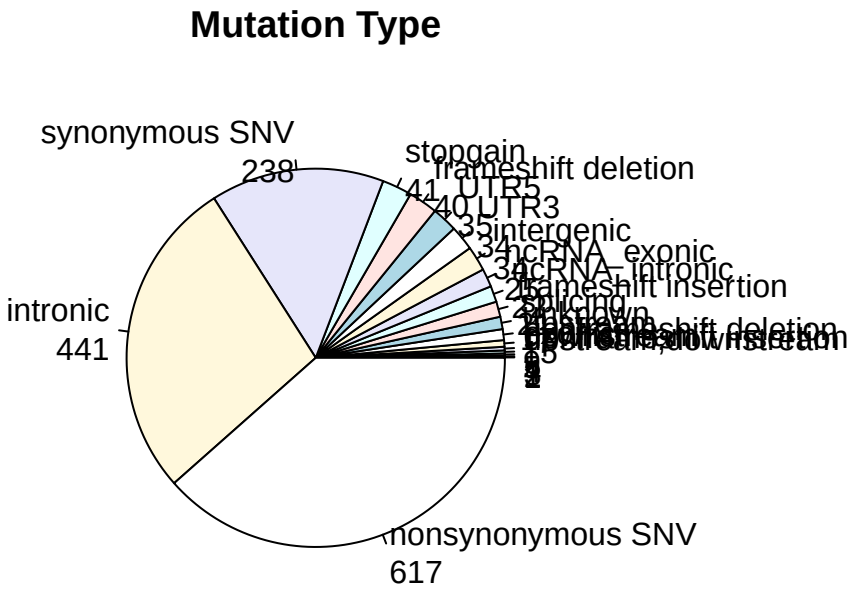

```
count1 %>% filter(is.na(Comment)) %>%
    with(t.test(n ~ factor(pd1)))
```

```
##
##  Welch Two Sample t-test
##
## data:  n by factor(pd1)
## t = 1.2357, df = 2.1999, p-value = 0.3322
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -112.7132  215.3799
## sample estimates:
## mean in group high  mean in group low
##           218.3333           167.0000
```

```
count1 %>% filter(is.na(Comment)) %>%
    with(wilcox.test(n ~ factor(pd1)))
```

```
##
##  Wilcoxon rank sum test
##
## data:  n by factor(pd1)
## W = 7, p-value = 0.4
## alternative hypothesis: true location shift is not equal to 0
```

```
merged$type %>% table %>% sort %>%
    pie(labels = paste(names(.), "\n", ., sep = ""), main = "Mutation Type")
```




```
merged$type %>% table %>% sort(decreasing = T)
```

```
## .
##       nonsynonymous SNV                intronic          synonymous SNV
##                     617                    441                     238
##                stopgain     frameshift deletion                    UTR5
##                      41                     40                      35
##              intergenic                   UTR3             ncRNA_exonic
##                      34                     34                      25
##          ncRNA_intronic    frameshift insertion                 splicing
##                      22                     21                      17
##                 unknown                upstream  nonframeshift deletion
##                      15                      9                       5
##                  exonic             downstream nonframeshift insertion
##                       4                      3                       2
##     upstream;downstream
##                       1
```

```
merged$type2 <- factor(merged$type)
levels(merged$type2)
```

```
##  [1] "downstream"              "exonic"
##  [3] "frameshift deletion"     "frameshift insertion"
##  [5] "intergenic"              "intronic"
##  [7] "ncRNA_exonic"            "ncRNA_intronic"
##  [9] "nonframeshift deletion"  "nonframeshift insertion"
## [11] "nonsynonymous SNV"       "splicing"
## [13] "stopgain"                "synonymous SNV"
## [15] "unknown"                 "upstream"
## [17] "upstream;downstream"     "UTR3"
## [19] "UTR5"
```

```
levels(merged$type2) <- c("Silent",
                          "Undetermined",
                          "frameshift_indel",
                          "Inframe",
                          "unknown",
```

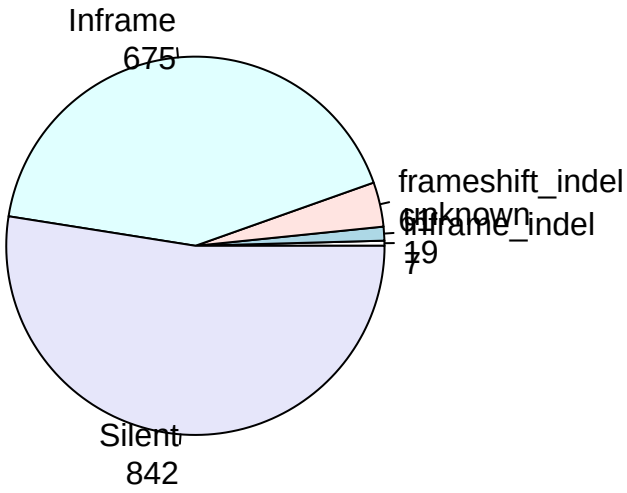
```
                          "inframe_indel"
)[c(1, 5, 3,
    3, 1, 1,
    1, 1, 6,
    6, 4, 4,
    4, 1, 5,
    1, 1, 1,
    1)]

merged %>% group_by(type, type2) %>% summarize(n = n()) %>% arrange(desc(type2))
```

```
## # A tibble: 19 x 3
## # Groups:   type [19]
##    type                    type2               n
##    <chr>                   <fct>           <int>
##  1 nonsynonymous SNV       Inframe           617
##  2 splicing                Inframe            17
##  3 stopgain                Inframe            41
##  4 nonframeshift deletion  inframe_indel       5
##  5 nonframeshift insertion inframe_indel       2
##  6 frameshift deletion     frameshift_indel   40
##  7 frameshift insertion    frameshift_indel   21
##  8 exonic                  unknown             4
##  9 unknown                 unknown            15
## 10 downstream              Silent              3
## 11 intergenic              Silent             34
## 12 intronic                Silent            441
## 13 ncRNA_exonic            Silent             25
## 14 ncRNA_intronic          Silent             22
## 15 synonymous SNV          Silent            238
## 16 upstream                Silent              9
## 17 upstream;downstream     Silent              1
## 18 UTR3                    Silent             34
## 19 UTR5                    Silent             35
```

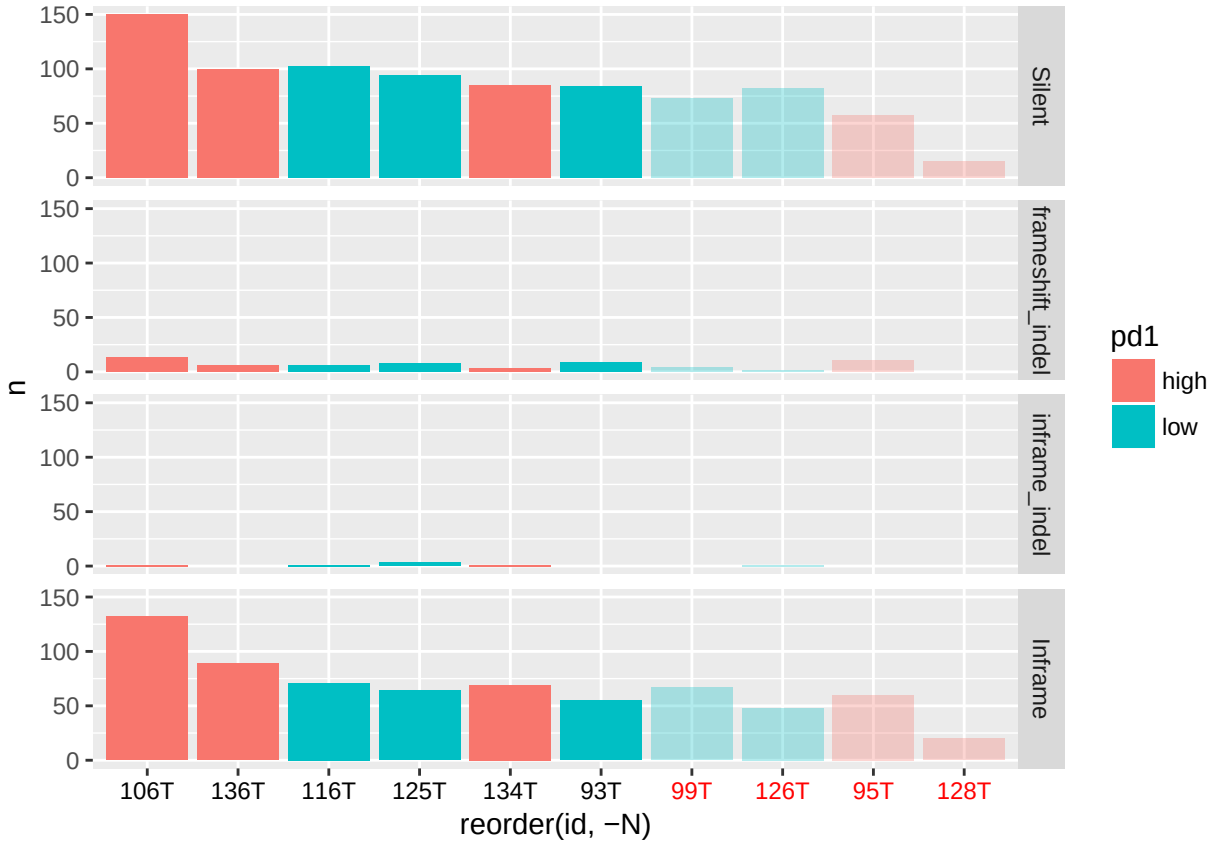
```
merged$type2 %>% table %>% sort %>%
    pie(labels = paste(names(.), "\n", ., sep = ""), main = "Mutation Type")
```

Mutation Type

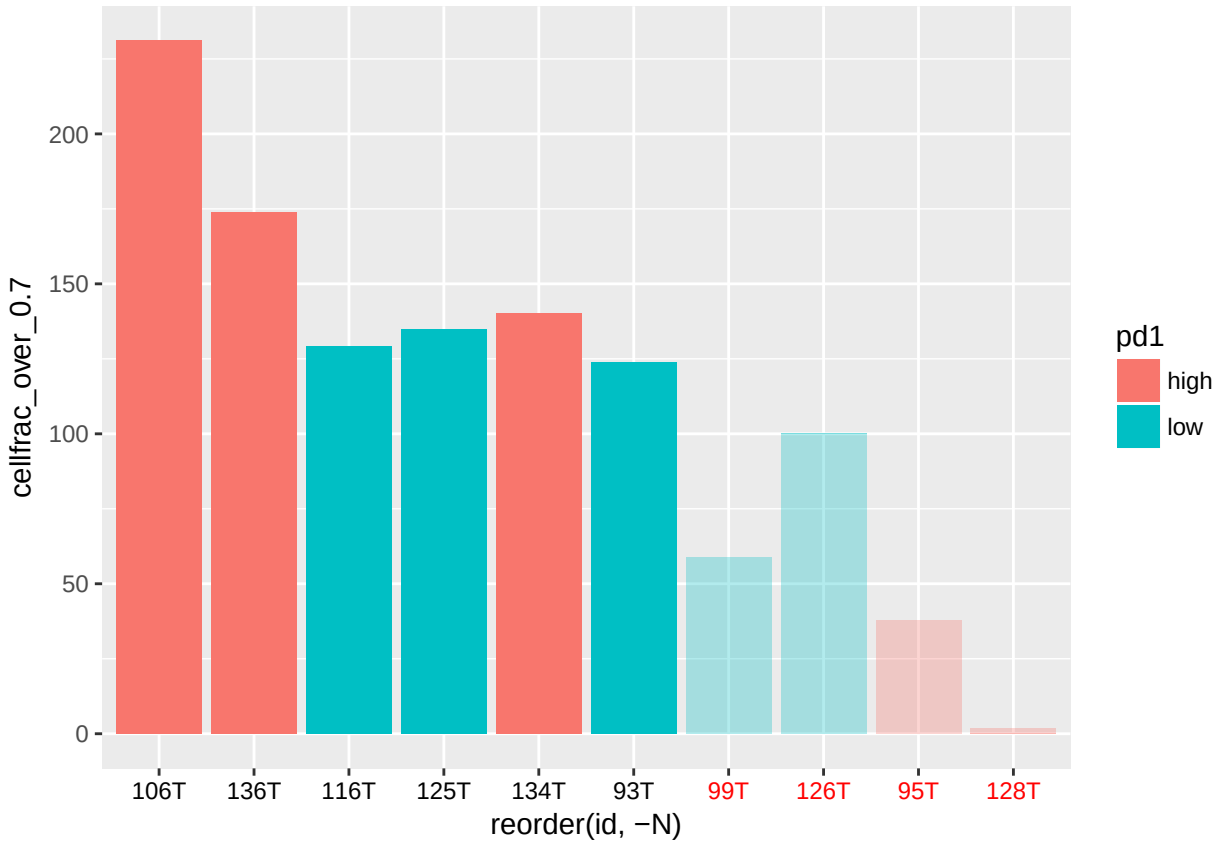



```
count2 <- merged %>%
    group_by(pd1, id) %>%
    mutate(N = n()) %>%
    group_by(pd1, id, type2) %>%

    summarise(n = n(), N = max(N),
              cellfrac_over_0.5 = sum(CELLFRACTION > .5, na.rm = T),
              cellfrac_over_0.7 = sum(CELLFRACTION > .7, na.rm = T)) %>%
    filter(type2 != "unknown") %>%
    left_join(purity, by = "id") %>% arrange(pd1, -n)
count2 %>%
    bind_cols(alpha_vector = ifelse(is.na(count2$Comment), 1, 0.3)) %>%
    ggplot(aes(x = reorder(id, -N), y = n, fill = pd1, alpha = I(alpha_vector))) +
    geom_col() +
    theme(axis.text.x = element_text(colour = color_vector)) +
    facet_grid(type2 ~ .)
```



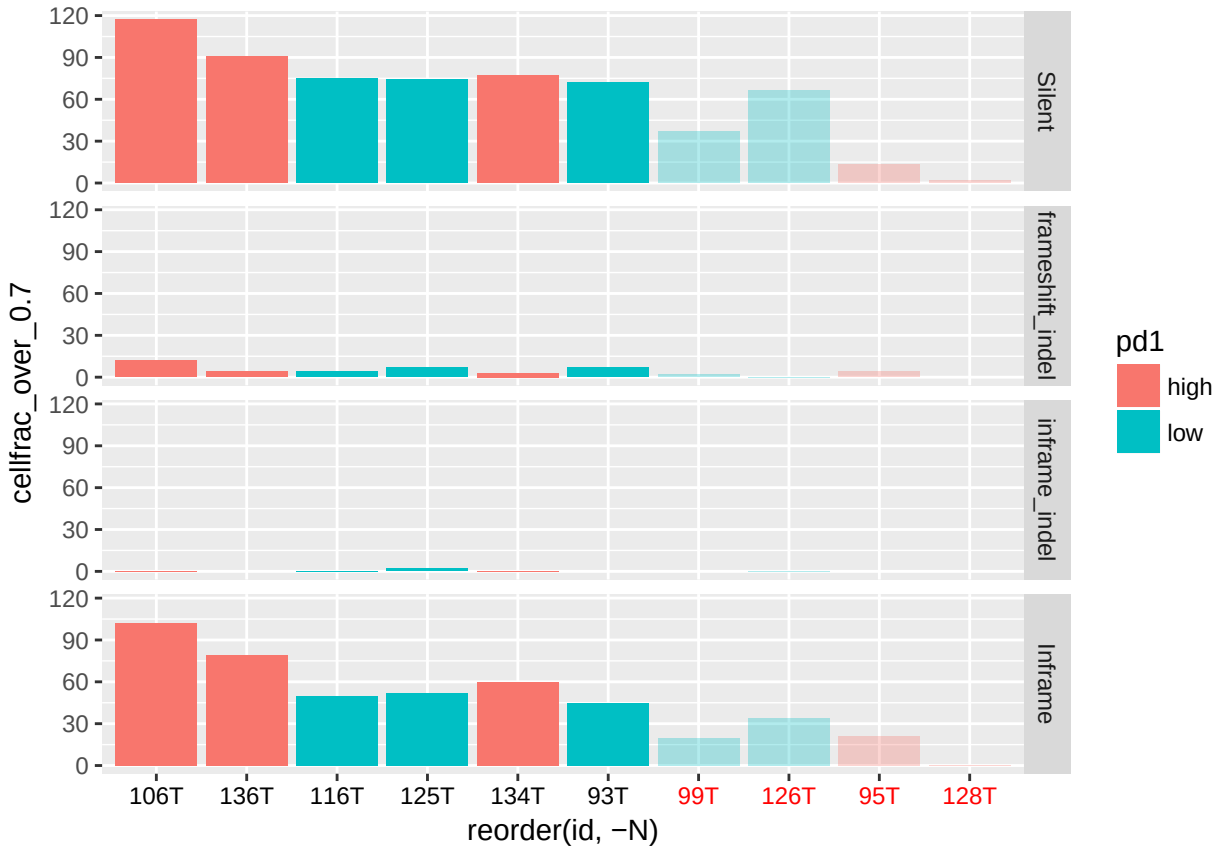

```
count2 %>%
    bind_cols(alpha_vector = ifelse(is.na(count2$Comment), 1, 0.3)) %>%
    ggplot(aes(x = reorder(id, -N), y = cellfrac_over_0.7, fill = pd1, alpha = I(alpha_vector))) +
    geom_col() +
    theme(axis.text.x = element_text(colour = color_vector))
```




```
count2 %>% filter(is.na(Comment)) %>%
    with(t.test(cellfrac_over_0.7 ~ factor(pd1)))
```

```
##
##  Welch Two Sample t-test
##
## data:  cellfrac_over_0.7 by factor(pd1)
## t = 0.8467, df = 17.665, p-value = 0.4085
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21.19045  49.73590
## sample estimates:
## mean in group high  mean in group low
##           49.54545           35.27273
```

```
count2 %>%
    bind_cols(alpha_vector = ifelse(is.na(count2$Comment), 1, 0.3)) %>%
    ggplot(aes(x = reorder(id, -N), y = cellfrac_over_0.7, fill = pd1, alpha = I(alpha_vector))) +
    geom_col() +
    theme(axis.text.x = element_text(colour = color_vector)) +
    facet_grid(type2 ~ .)
```




```
count2 %>% filter(is.na(Comment)) %>%
    with(wilcox.test(cellfrac_over_0.7 ~ factor(pd1)))
```

```
## Warning in wilcox.test.default(x = c(117L, 102L, 91L, 79L, 77L, 60L, 12L, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  cellfrac_over_0.7 by factor(pd1)
## W = 73.5, p-value = 0.411
## alternative hypothesis: true location shift is not equal to 0
```