Structural Variant Analysis on the Exceptional Responders Cohort

David Fuh, Summer Intern, Garvan Institute of Medical Research, Kinghorn Centre for Clinical Genomi

2022-01-27

Contents

Packages **1**

Data import **1**

I. Single-Variable Analysis **1**
1. SV call location 1
2. Types of Structural Variants 4
3. Size and length of structural variants detected 5
4. Number of variants detected by all methods 7
5. ACMG class of variants (essentially unpresentable as many caveats) 7
6. Detected variants affecting CDS 8

II. Visualising high-dimensional data **17**

Sandbox **17**
Create a smaller subset of the master file 17

Footnotes **17**

Session Info **17**

Packages

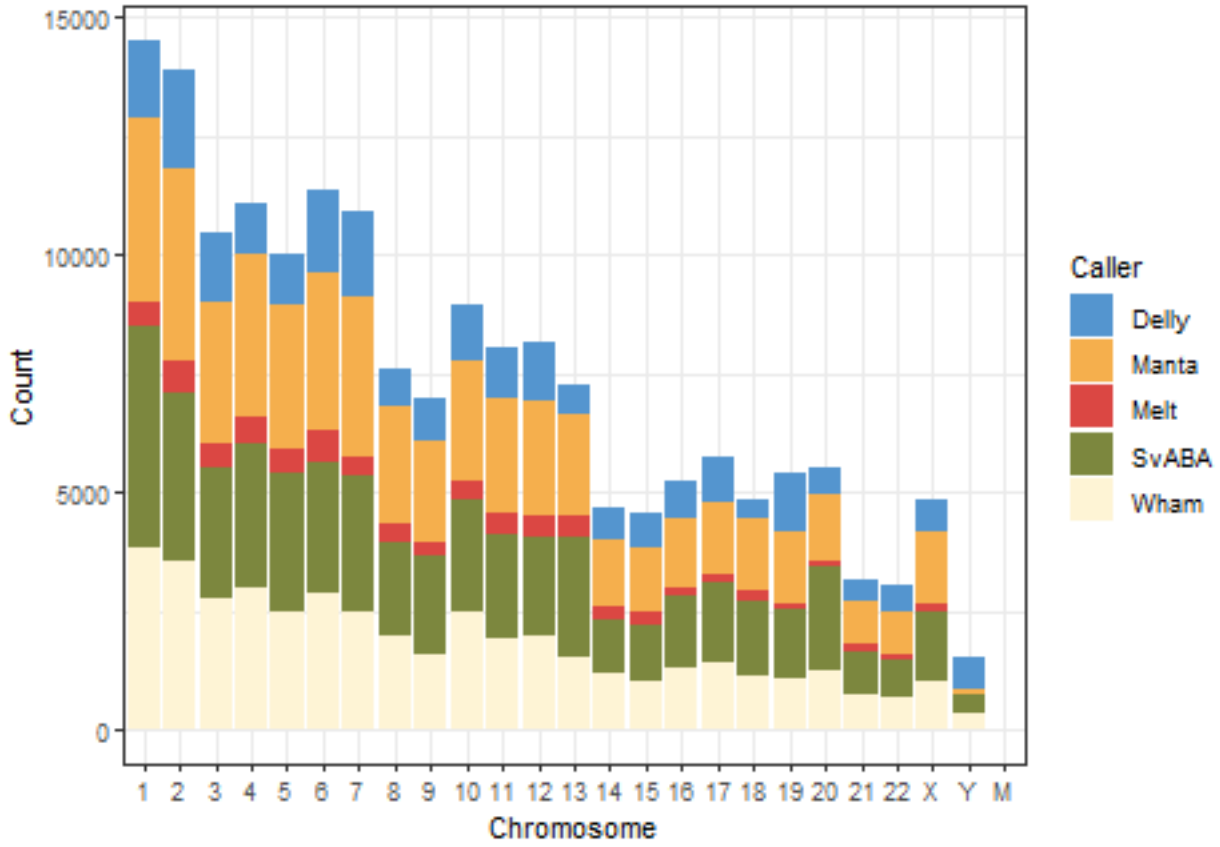
Data import

I. Single-Variable Analysis

1. SV call location

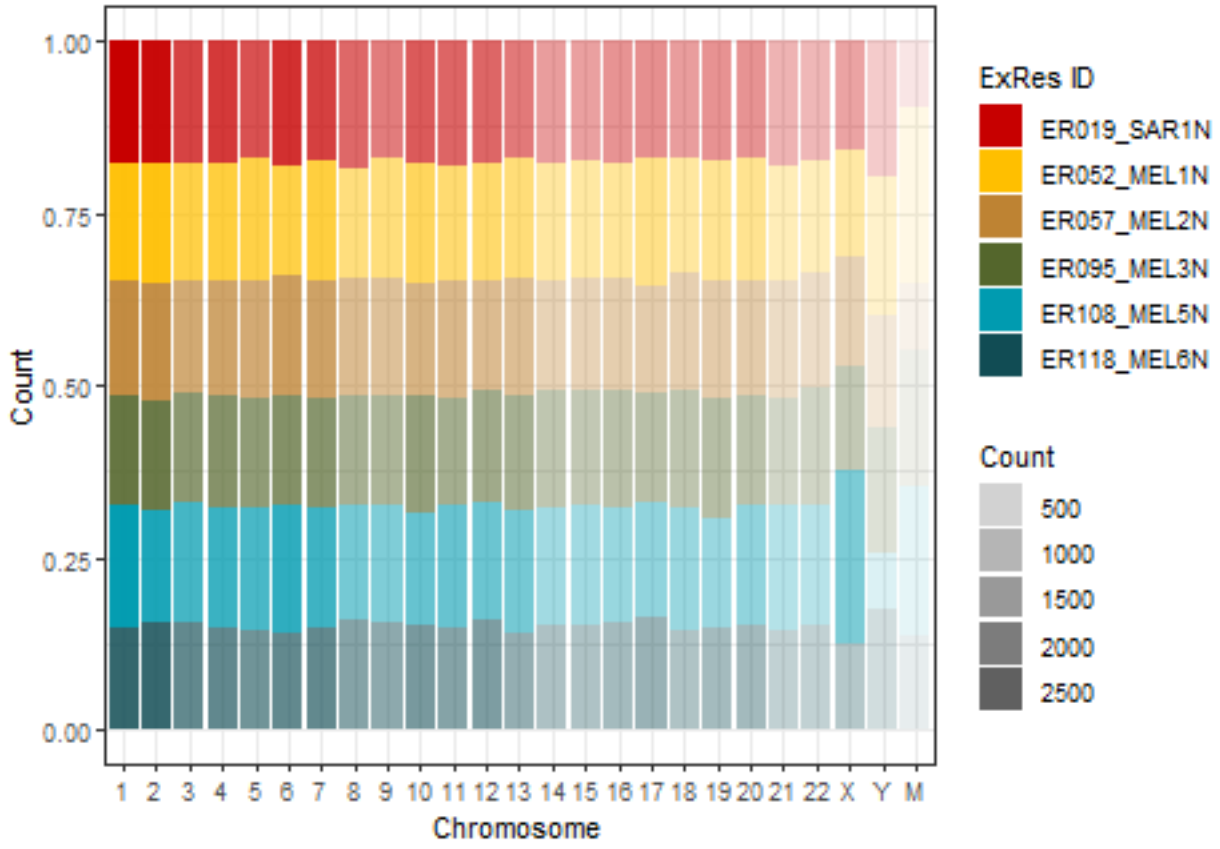
```
##
##      1      2      3      4      5      6      7      8      9     10     11
## 204069 610820 104155  46524  40185  55793  87177  19506  33269  28504 123780
##     12     13     14     15     16     17     18     19     20     21     22
##  65126  18002  21334  29935  57179  41346   8640  28737   9483  10537  11739
##      X      Y      M
##  22984   7948     51
```

1-1. SV location by caller



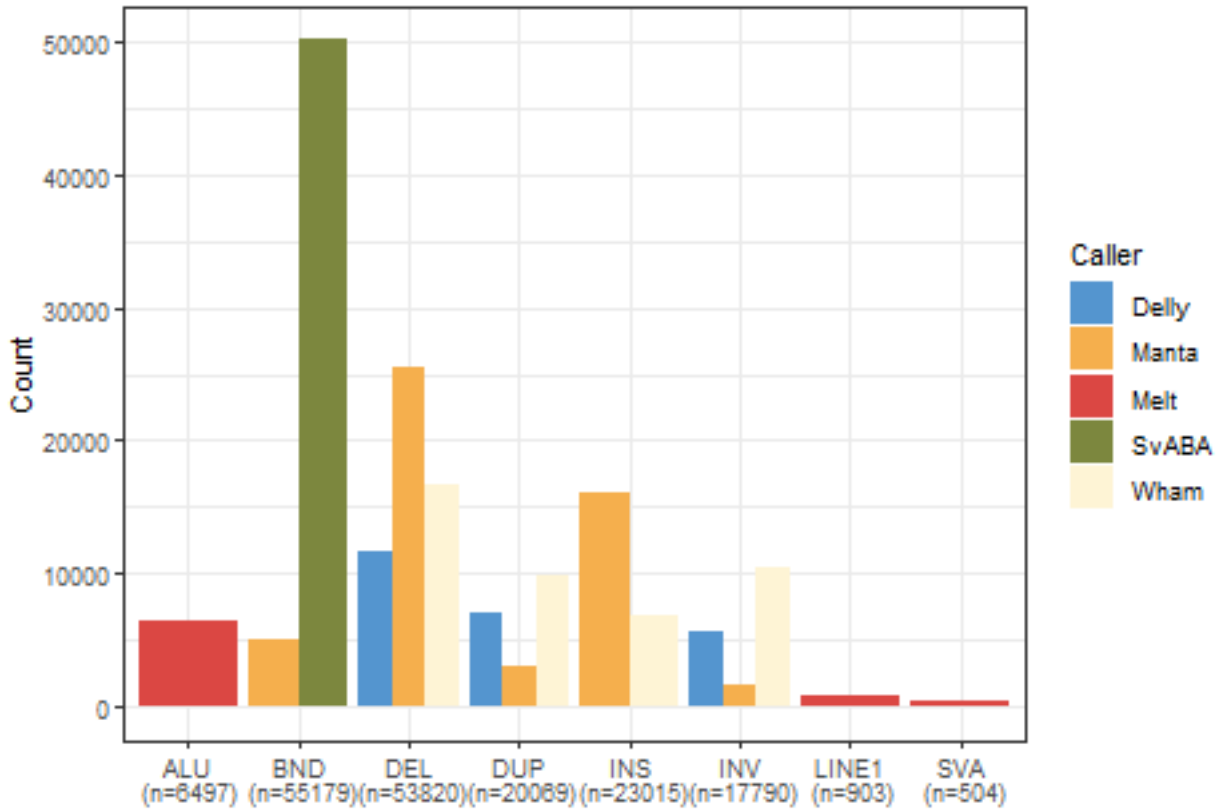
(Superimpose line graph of chromosomal length on diaxis plot)

1-2. SV location by ExRes ID



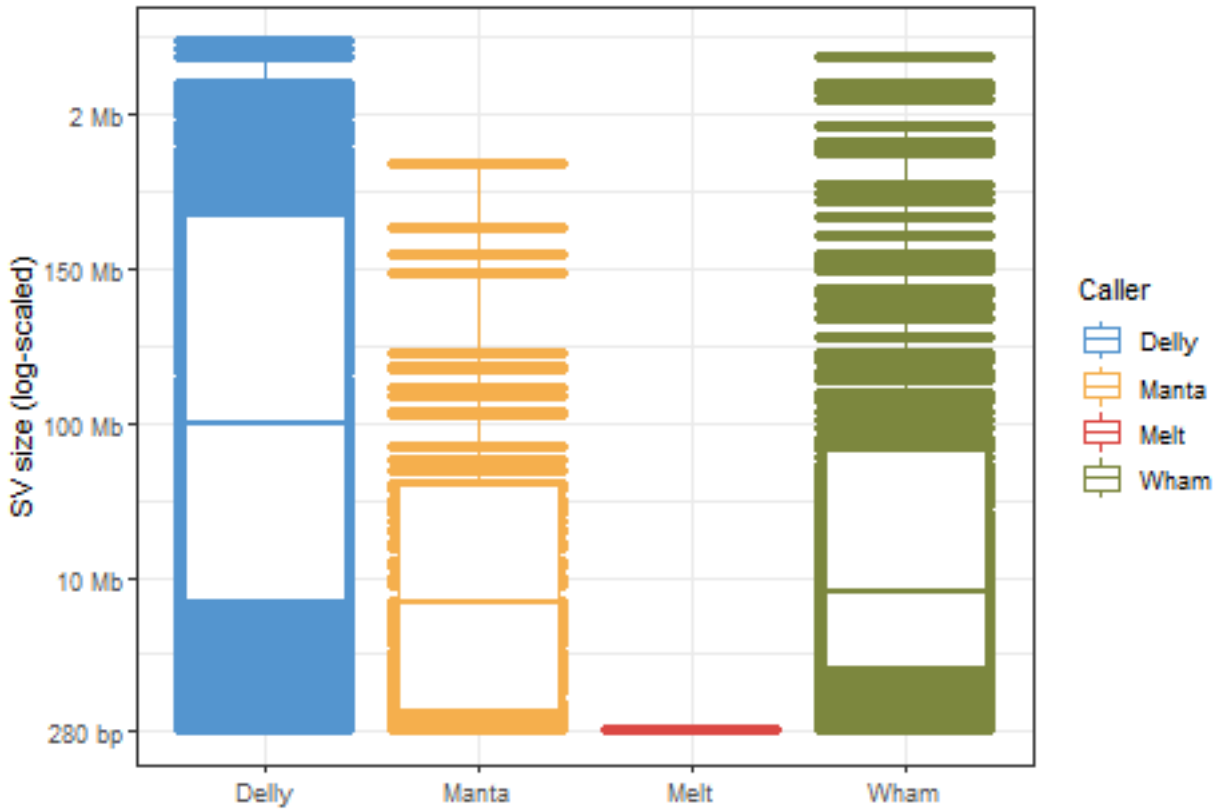
2. Types of Structural Variants

2-1. SV type between callers

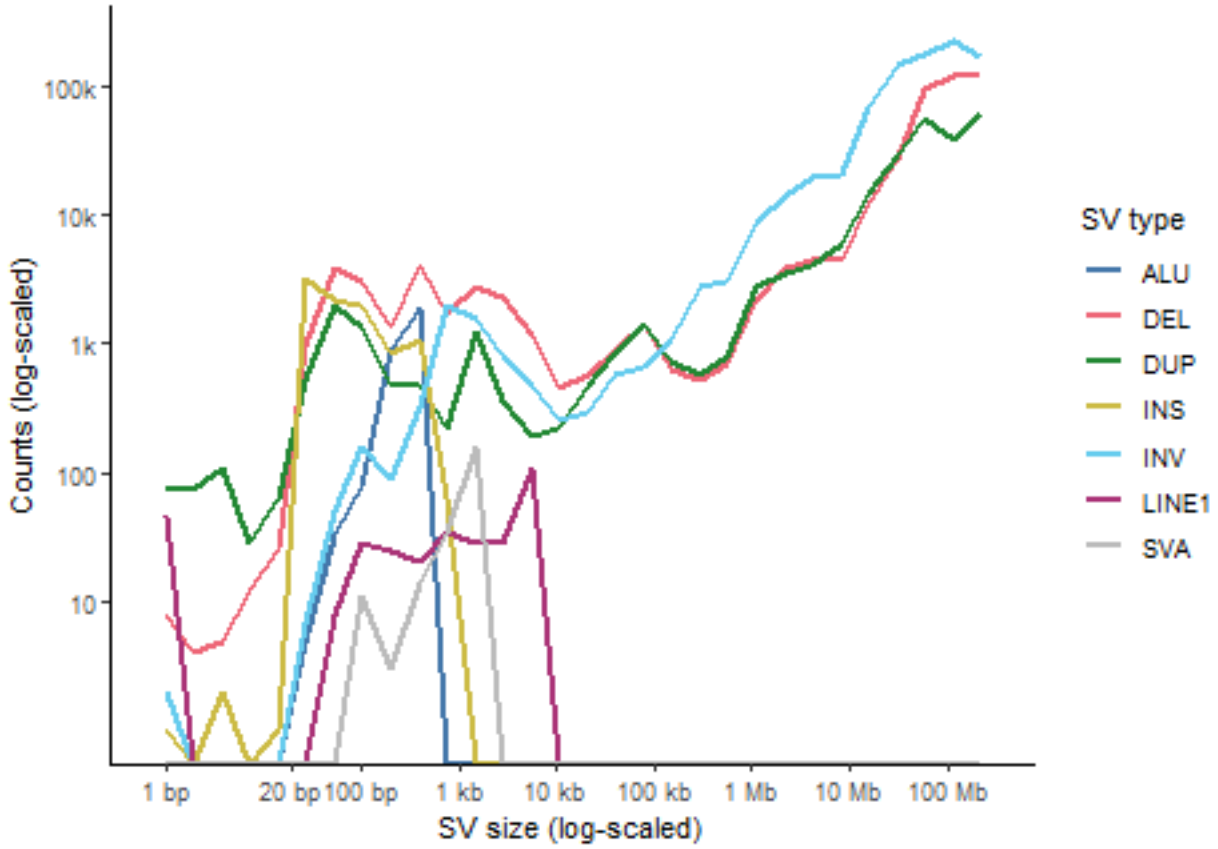


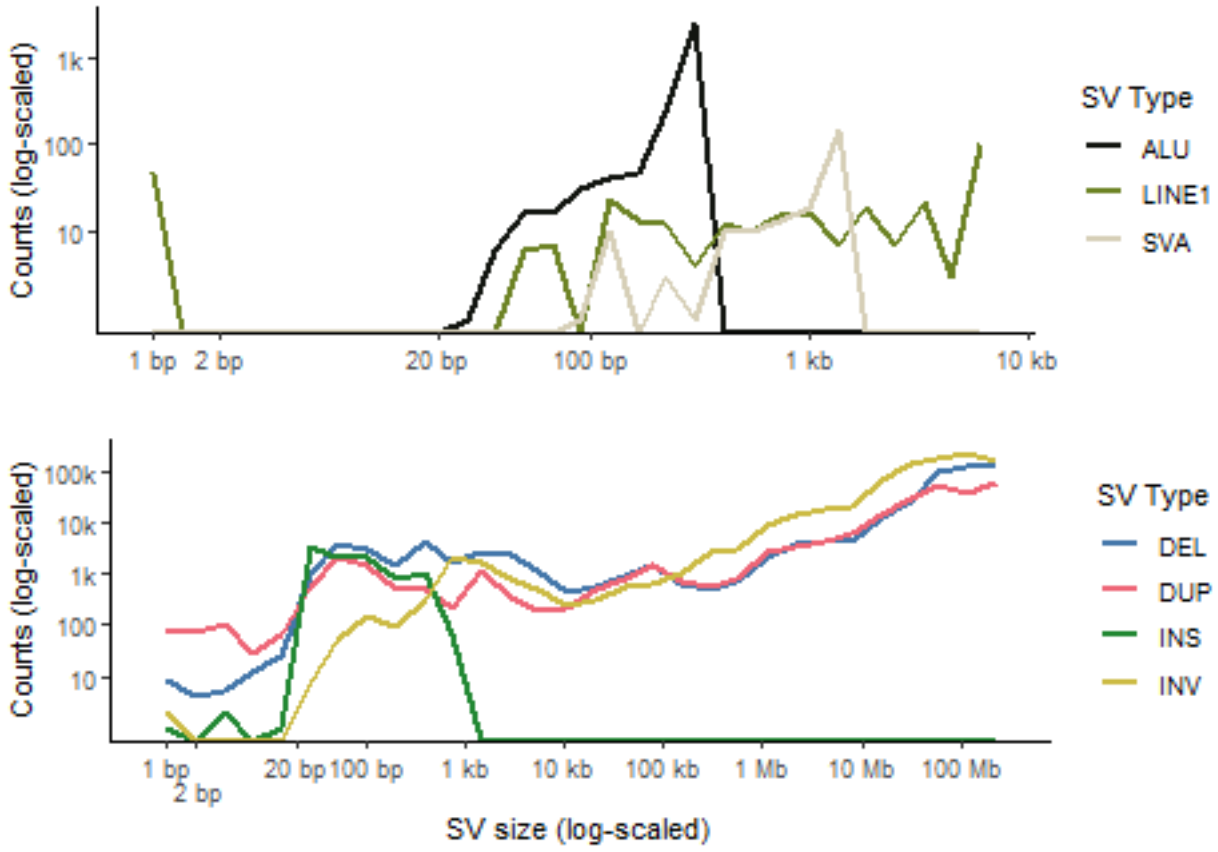
3. Size and length of structural variants detected

3-1. SV length by caller



3-2. SV length vs SV type





Problem: 1) BND (svaba) calls ignored 2) INS = ALU + LINE1 + SVA + ... (the bottom graph didn't synthesise them)

4. Number of variants detected by all methods

See Common_caller_test.R.

5. ACMG class of variants (essentially unpresentable as many caveats)

```
##
##       1       3       4       5  full=1  full=3  full=4  full=5 full=NA
##   24689   47241     200    1759   11932   30229    2795  587547  876543
```

```
sample_size = ExRes %>% filter(Annotation_mode == "full") %>% group_by(ACMG_class) %>% summarize(num=n()

ExRes %>%
  filter(Annotation_mode == "full") %>%
  left_join(sample_size) %>%
  mutate(ACMG.class = paste0(ACMG_class, "\n (n=", num, ")")) %>%

ggplot(aes(x = ACMG.class, fill=SV_type)) +
    geom_bar() +
    theme_bw() +
    labs(x="ACMG class", y="Count") +
    scale_fill_discrete(name = "SV type")

rm("sample_size")
```

Try removing ACMG = NA entries:

```
filter(ExRes, !is.na(ACMG_class)) %>% filter(ACMG_class != "full=NA") %>%

ggplot(aes(x = ACMG_class, fill=SV_type)) +
    geom_bar() +
    theme_bw() +
    labs(x="ACMG class", y="count") +
    scale_fill_discrete(name = "SV type")
```

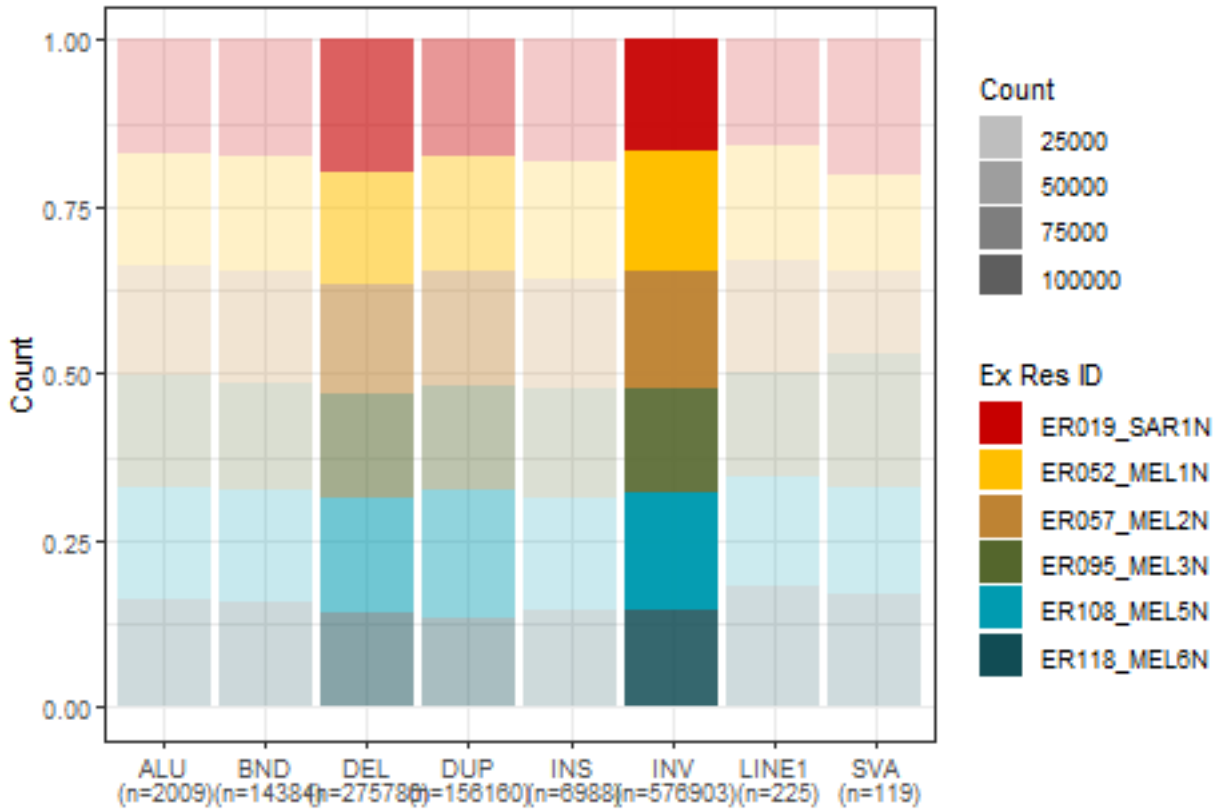
6. Detected variants affecting CDS

6-1. CDS-affecting SVs

```
##
##       3'UTR       5'UTR 5'UTR-3'UTR   5'UTR-CDS         CDS   CDS-3'UTR
##        1590        9953      980128        2598       47118        2730
##         UTR
##      464929
```

Note that only split AM contains topology data (confirmed).

SV curation set: CDS-affecting only



(run multi ANOVA: In each SV type, which patient has the most sig. deviation? Star it.)

6-2. Prioritised variant list

<1st layer: AnnotSV input>

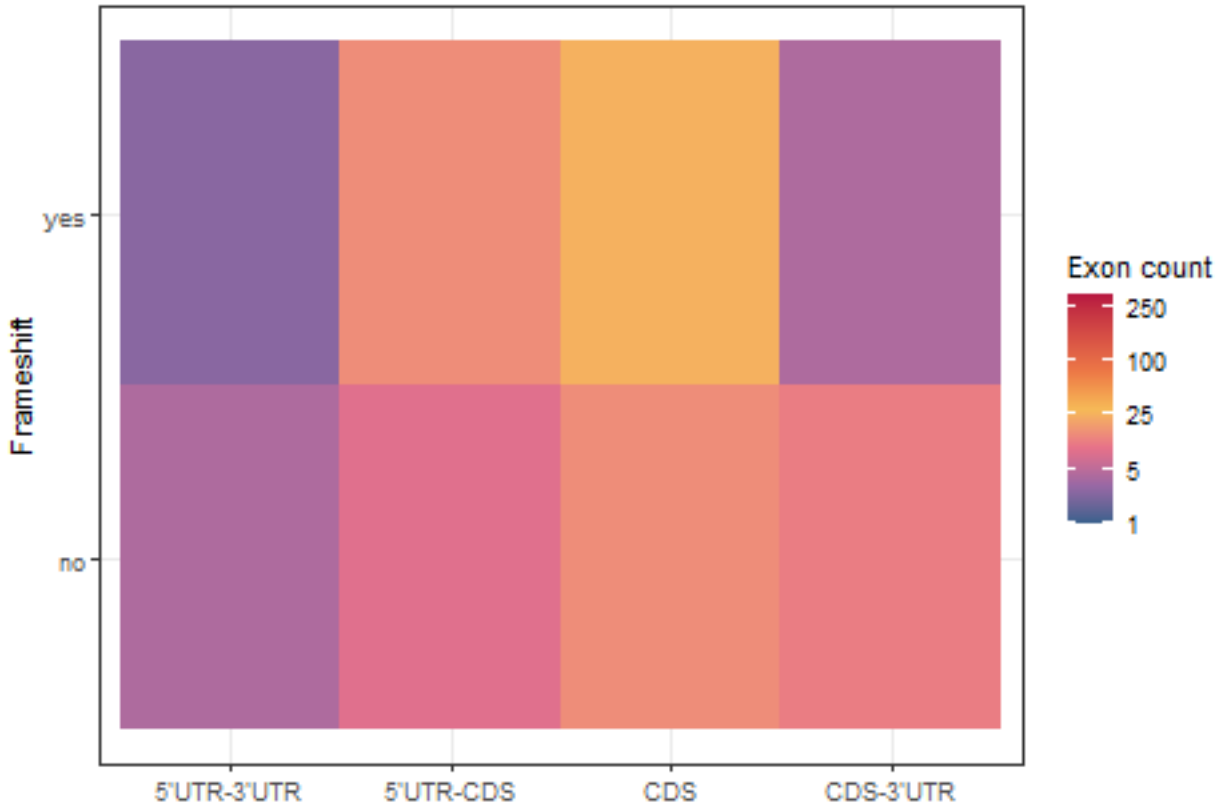
```
Raw call set %>% filter(PASS) %>% filter(ACMG = {4,5}) %>%

[R scripts]       [R scripts]           [Task 4]
```

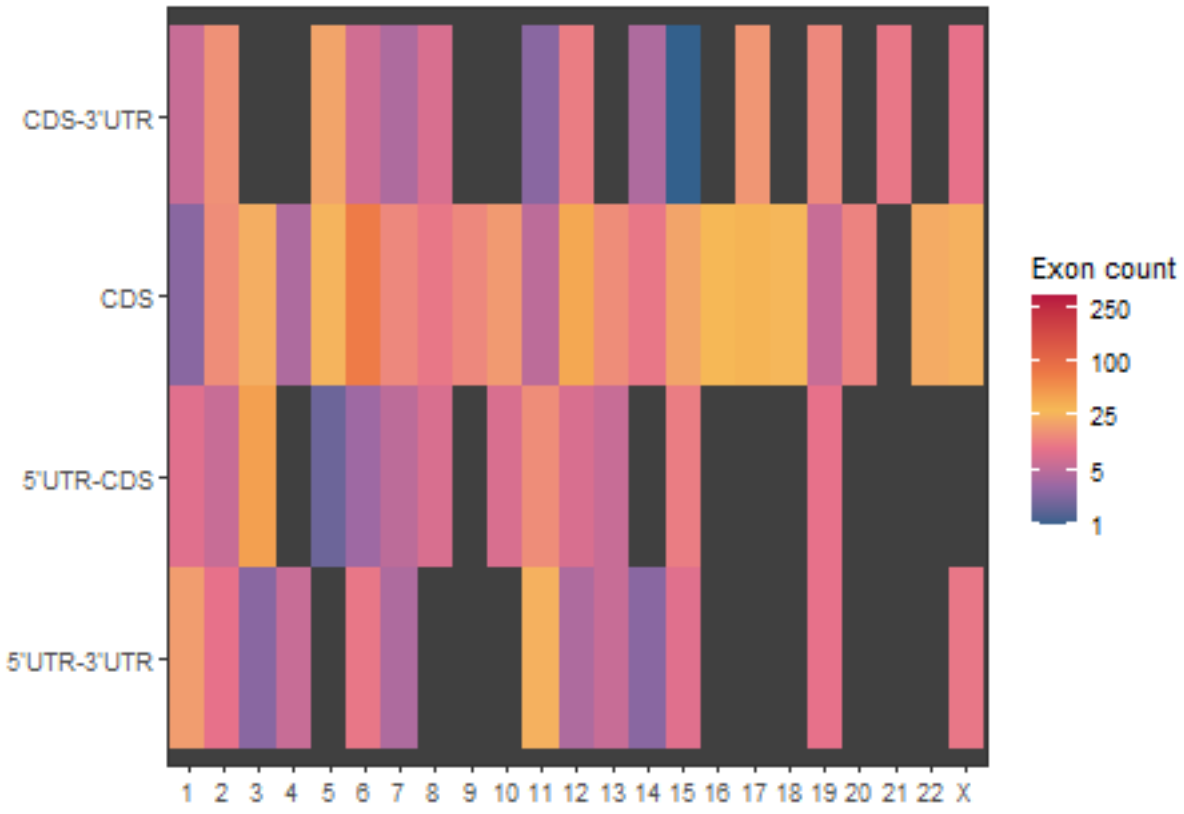
<2nd layer>

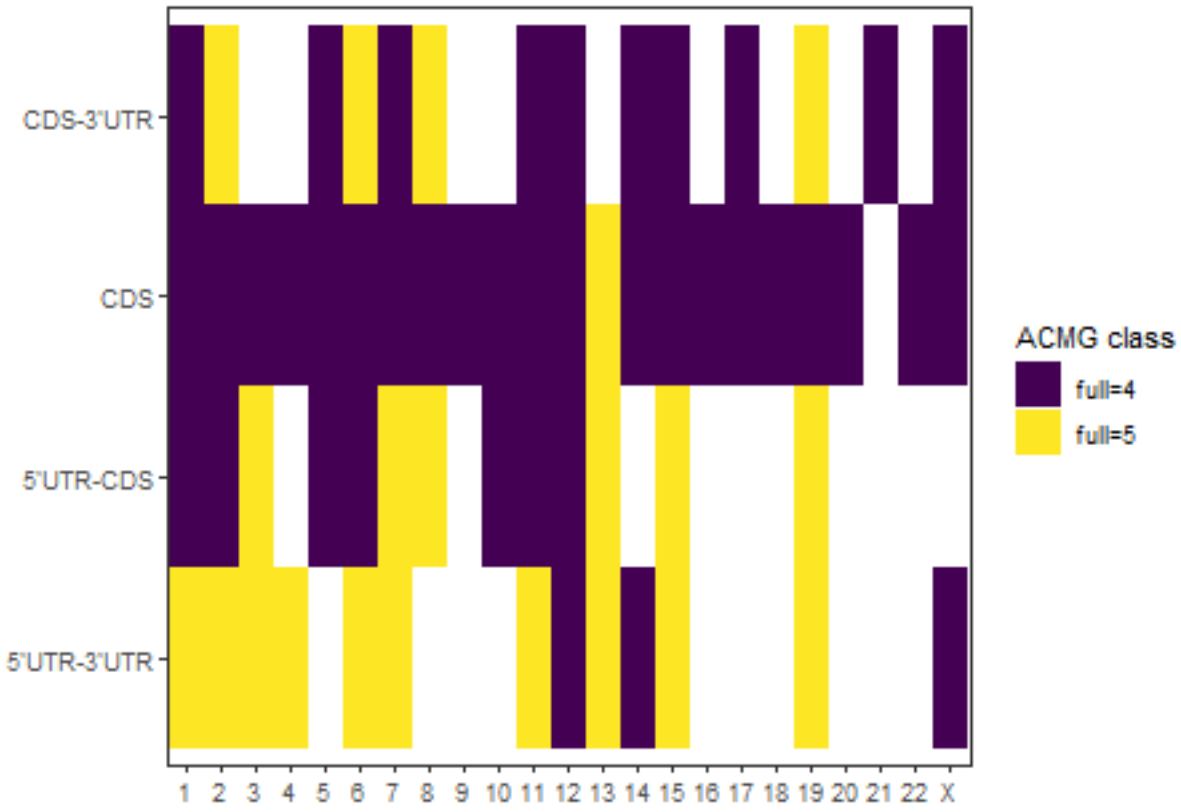
```
      filter(Called by e.g. at least 3 callers out of 5 used) %>%
      [Task 3]


      filter(CDS-affecting) %>%
      [Task 5]


      filter(Known to affect XX tissue)
      [This task]
```

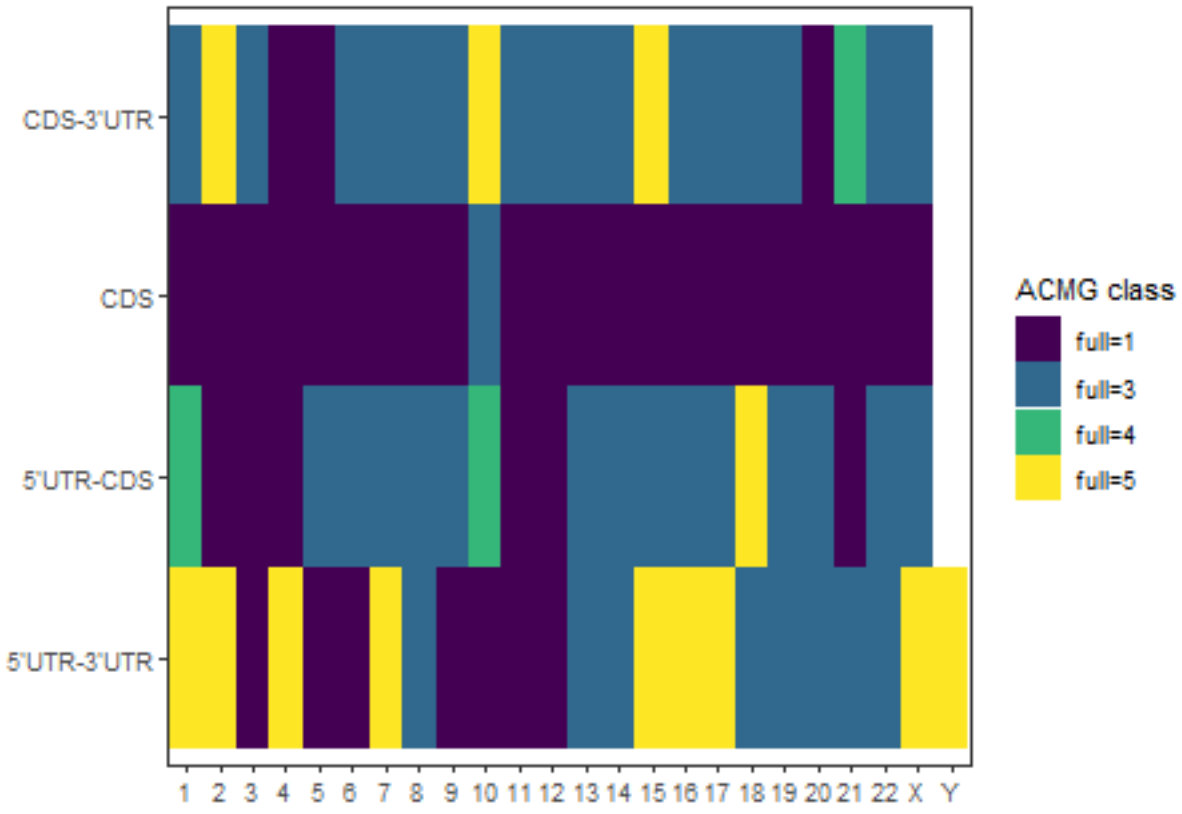
List curation

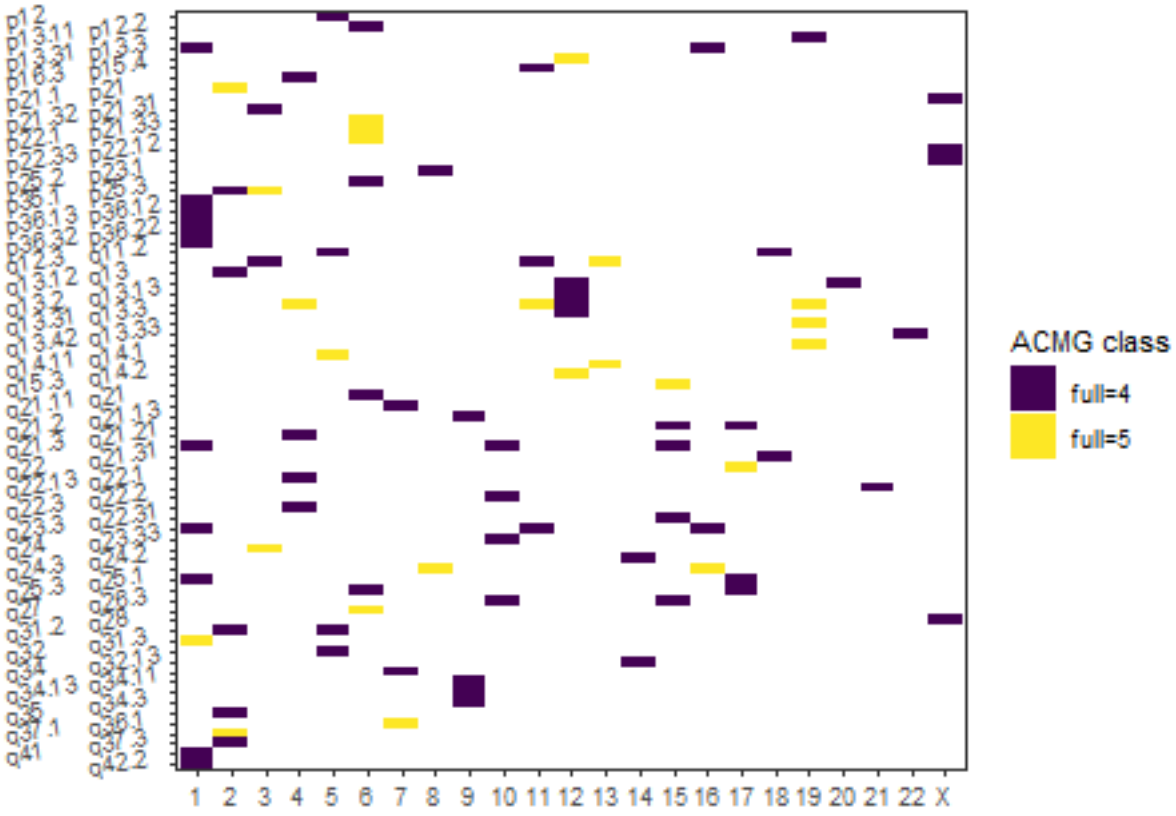
6-3. Analysis on prioritised variants

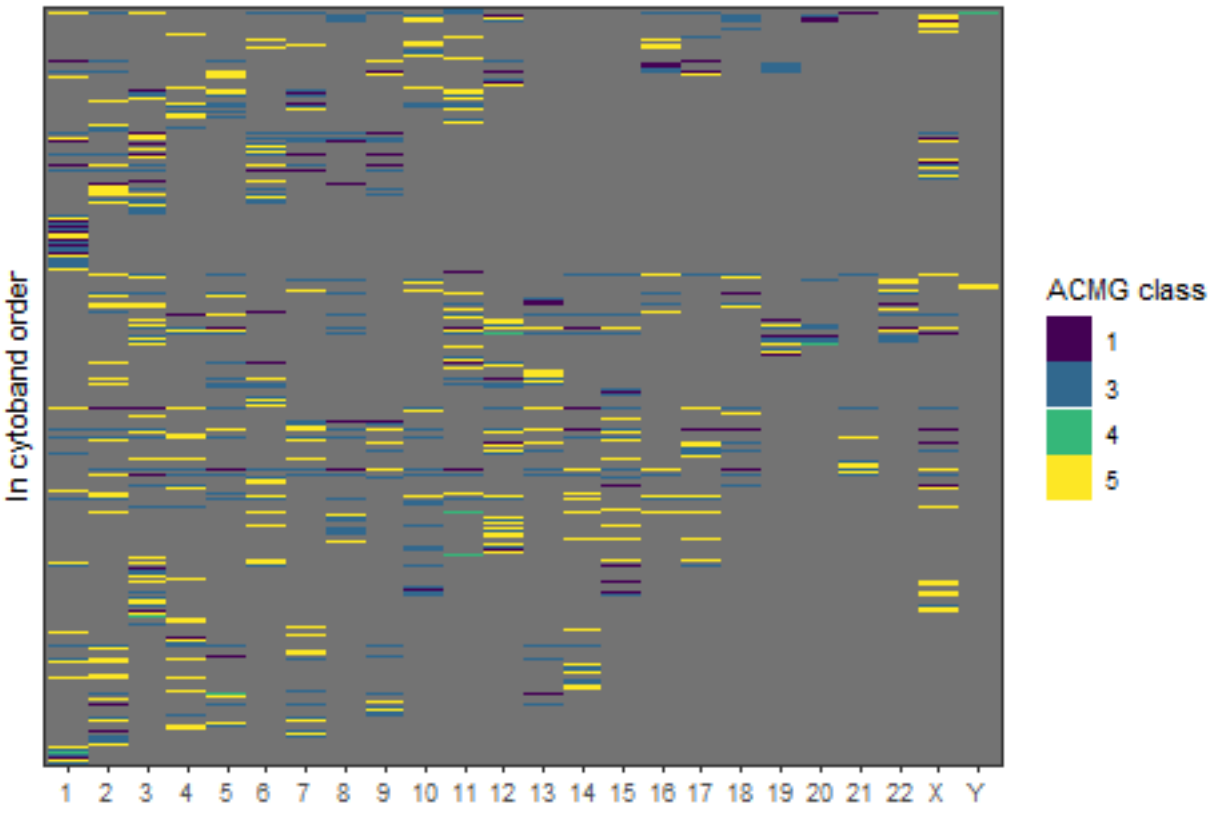



CDS-3'UTR
CDS
5'UTR-CDS
5'UTR-3'UTR
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Exon count
250
100
25
5
1

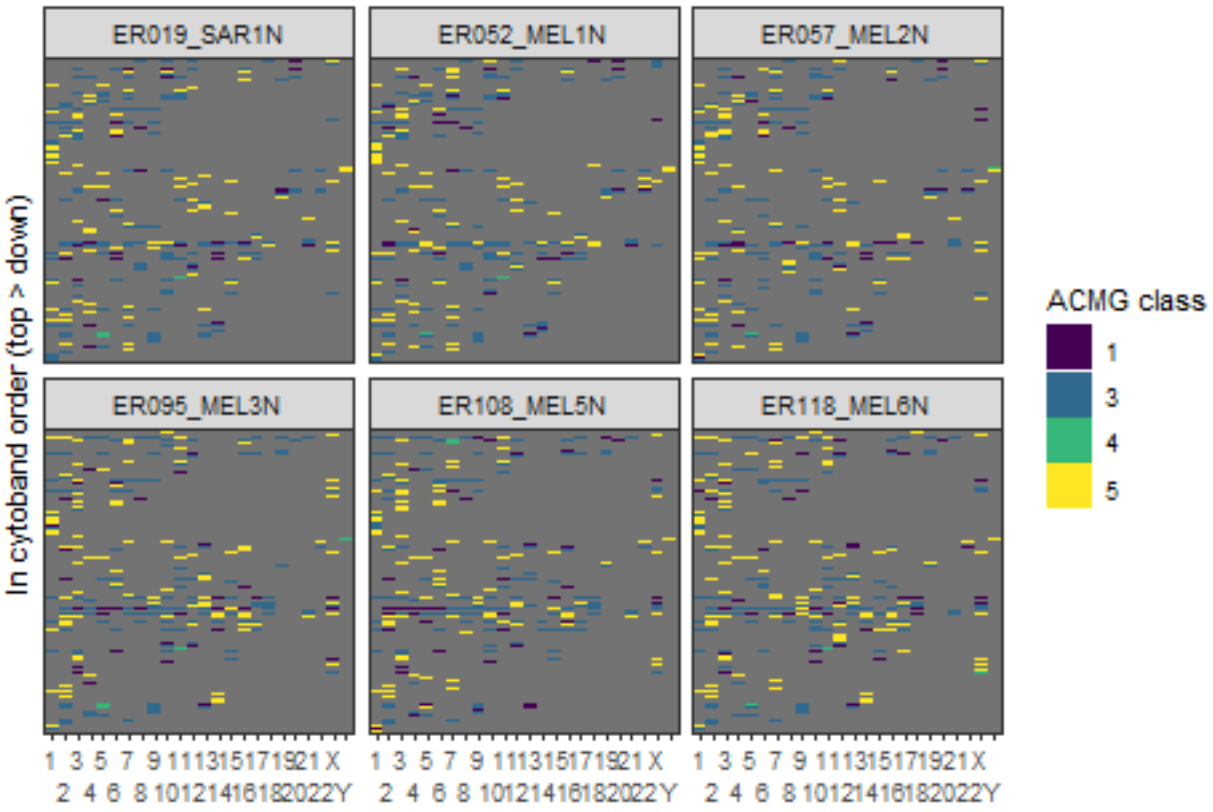

CDS-3'UTR
CDS
5'UTR-CDS
5'UTR-3'UTR
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
ACMG class
full=4
full=5

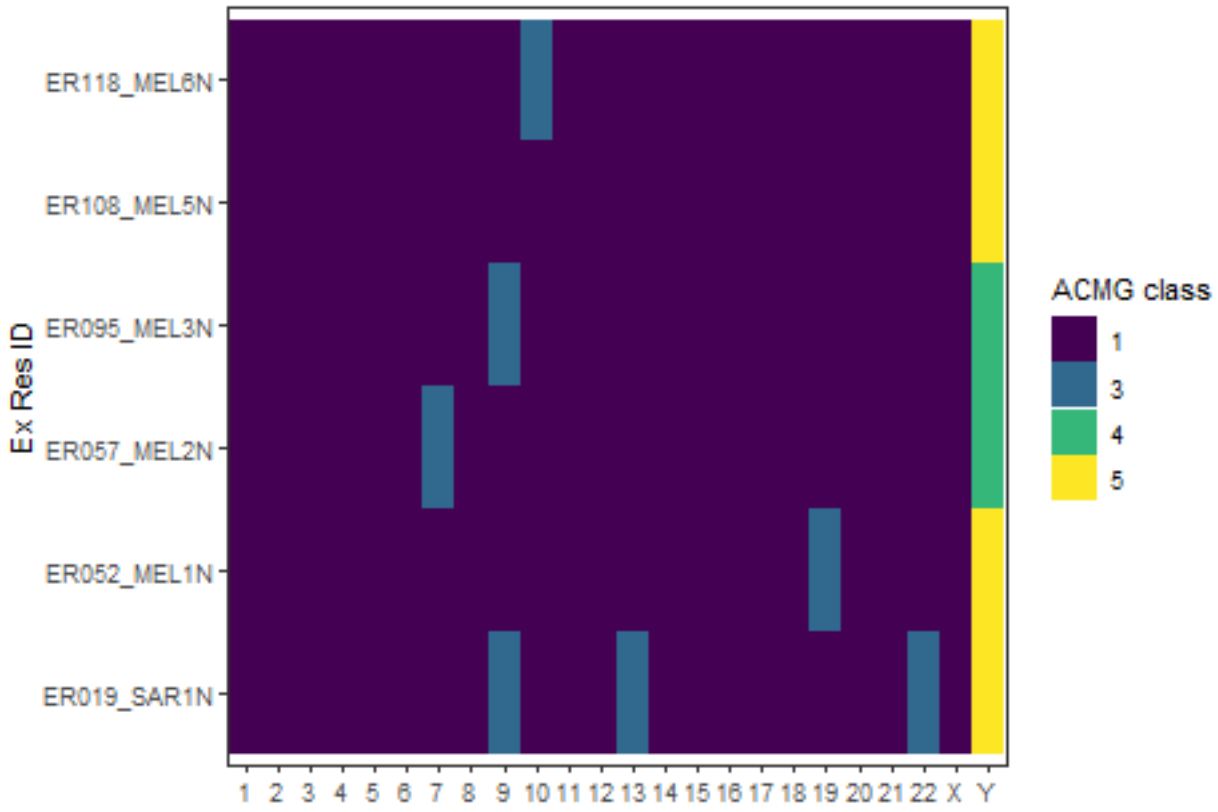

CDS-3'UTR
CDS
5'UTR-CDS
5'UTR-3'UTR
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
ACMG class
full=1
full=3
full=4
full=5


ACMG class
full=4
full=5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X



How to correctly interpret this plot?


ER019_SAR1N
ER052_MEL1N
ER057_MEL2N
ER095_MEL3N
ER108_MEL5N
ER118_MEL6N
In cytoband order (top > down)
1 3 5 7 9 11 13 15 17 19 21 X
2 4 6 8 10 12 14 16 18 20 22 Y
ACMG class
1
3
4
5


Ex Res ID
ER118_MEL6N
ER108_MEL5N
ER095_MEL3N
ER057_MEL2N
ER052_MEL1N
ER019_SAR1N
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
ACMG class
1
3
4
5

II. Visualising high-dimensional data

Sandbox

Create a smaller subset of the master file

Footnotes

Session Info

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19043)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_Australia.1252  LC_CTYPE=English_Australia.1252
## [3] LC_MONETARY=English_Australia.1252 LC_NUMERIC=C
## [5] LC_TIME=English_Australia.1252
## system code page: 950
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
```

```
##
## other attached packages:
##  [1] viridis_0.6.2     viridisLite_0.4.0 cowplot_1.1.1     ggforce_0.3.3
##  [5] paletteer_1.4.0   forcats_0.5.1     stringr_1.4.0     dplyr_1.0.7
##  [9] purrr_0.3.4       readr_2.1.1       tidyr_1.1.4       tibble_3.1.6
## [13] ggplot2_3.3.5     tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] httr_1.4.2       bit64_4.0.5      vroom_1.5.7      jsonlite_1.7.2
##  [5] modelr_0.1.8     assertthat_0.2.1 cellranger_1.1.0 yaml_2.2.1
##  [9] pillar_1.6.4     backports_1.4.1  glue_1.6.0       digest_0.6.29
## [13] polyclip_1.10-0  rvest_1.0.2      colorspace_2.0-2 htmltools_0.5.2
## [17] pkgconfig_2.0.3  broom_0.7.11     haven_2.4.3      scales_1.1.1
## [21] tweenr_1.0.2     tzdb_0.2.0       generics_0.1.1   farver_2.1.0
## [25] ellipsis_0.3.2   withr_2.4.3      cli_3.1.0        magrittr_2.0.1
## [29] crayon_1.4.2     readxl_1.3.1     evaluate_0.14    fs_1.5.2
## [33] fansi_0.5.0      MASS_7.3-54      xml2_1.3.3       ggthemes_4.2.4
## [37] tools_4.1.2      hms_1.1.1        lifecycle_1.0.1  munsell_0.5.0
## [41] reprex_2.0.1     compiler_4.1.2   rlang_0.4.12     grid_4.1.2
## [45] rstudioapi_0.13  labeling_0.4.2   rmarkdown_2.11   gtable_0.3.0
## [49] DBI_1.1.2        rematch2_2.1.2   R6_2.5.1         gridExtra_2.3
## [53] lubridate_1.8.0  knitr_1.37       fastmap_1.1.0    bit_4.0.4
## [57] utf8_1.2.2       prismatic_1.1.0  stringi_1.7.6    parallel_4.1.2
## [61] Rcpp_1.0.8       vctrs_0.3.8      dbplyr_2.1.1     tidyselect_1.1.1
## [65] xfun_0.29
```