FIG_4

Compiled: May 15, 2022

```
library(grid)
library(RColorBrewer)
```

```
## Warning: package 'RColorBrewer' was built under R version 4.0.5
```

```
library(pheatmap)
library(ggplot2)
library(reshape2)
library(ggpubr)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggVennDiagram)
library(VennDiagram)
```

```
## Loading required package: futile.logger
```

```
##
## Attaching package: 'VennDiagram'
```

```
## The following object is masked from 'package:ggVennDiagram':
##
##     ellipse
```

```
## The following object is masked from 'package:ggpubr':
##
##     rotate
```

```
library(ComplexUpset)
```

Fig. 4a - Variant prevalence (condition/databse annotation)

```
df3c <- read.csv("v2.variants.ALL_INDV.annot_hypergeom.var_prevalence.txt", sep = " ", header = F)
names(df3c) <- c("IND", "var", "region", "gene", "annot", "var_type", "label", "facets")

df3c$facets <- paste0(df3c$IND, "_", df3c$label, "_", df3c$var_type)

head(df3c)
```

IND	var	region	gene	annot	var_type	label	facets	NA	NA	NA	NA

IND	var	region	gene	annot	var_type	label	facets	NA	NA	NA	NA
TH179	chr1:14522:G>A	UTR3	WASH7P	Novel	SNP	germline	TH179_germline_SNP	1	4	2	2
TH179	chr1:14542:A>G	UTR3	WASH7P	Novel	SNP	germline	TH179_germline_SNP	1	4	2	2
TH179	chr1:14574:A>G	UTR3	WASH7P	Novel	SNP	germline	TH179_germline_SNP	1	3	3	3
TH179	chr1:14653:C>T	UTR3	WASH7P	Novel	SNP	germline	TH179_germline_SNP	3	2	2	3
TH179	chr1:14673:G>C	UTR3	WASH7P	Novel	SNP	germline	TH179_germline_SNP	0	3	5	4
TH179	chr1:14677:G>A	UTR3	WASH7P	Novel	SNP	germline	TH179_germline_SNP	2	1	3	5

```
for (indv in c("TH179", "TH238")) {
    for (var_type in c("SNP", "INDEL")) {
        df <- table(df3c[df3c$IND == indv & df3c$var_type == var_type, c("label", "annot")])
        c <- chisq.test(df)
        print(paste(indv, var_type, c$p.value, sep = " "))
    }
}
```

```
## [1] "TH179 SNP 0"
## [1] "TH179 INDEL 3.093973188804e-08"
## [1] "TH238 SNP 0"
## [1] "TH238 INDEL 2.74102112220463e-25"
```

```
sd3c_table <- table(df3c[c("IND", "annot", "var_type", "label", "facets")])

sd3c_table <- as.data.frame(sd3c_table)

g1 <- ggplot(sd3c_table, aes(x = "", y = Freq, fill = annot)) + geom_bar(stat = "identity", position = position_f
ill()) +
    geom_text(aes(label = Freq), position = position_fill(vjust = 0.5)) + coord_polar(theta = "y") +
    scale_fill_brewer(palette = "Pastel1") + facet_wrap(~facets, ncol = 2) + theme(axis.title.x = element_blank
(),
    axis.title.y = element_blank()) + theme(legend.position = "bottom") + guides(fill = guide_legend(nrow = 2,
    byrow = TRUE)) + theme_void()

pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/4a.pdf")
g1
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
g1
```

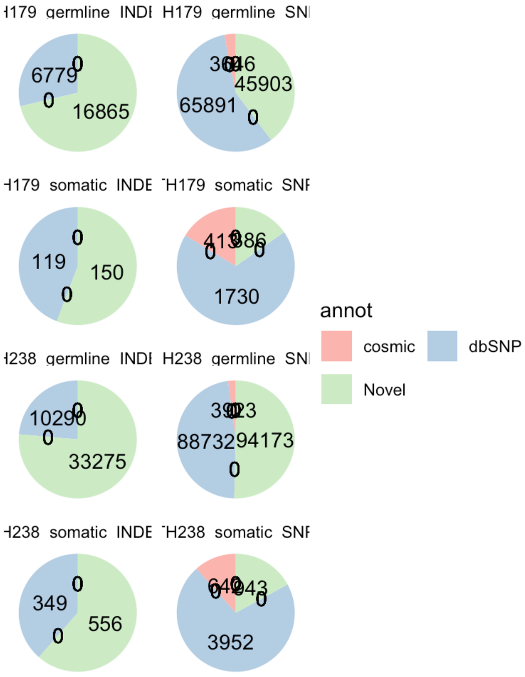


Fig. 4b - Variant prevalence (region annotation)

```
df3c$region <- factor(df3c$region, levels = c("exonic", "UTR3", "UTR5", "intronic", "intergenic",
    "downstream", "upstream"))

g1 <- ggplot(df3c, aes(x = label)) + geom_bar(color = "white", position = "fill", aes(fill = region)) +
    geom_text(aes(label = ..count..), stat = "count", position = position_fill()) + scale_fill_brewer(palette = "
Set1") +
    facet_grid(IND ~ var_type) + theme(legend.position = "bottom") + #guides(fill=guide_legend(nrow=2, byrow=TRU
E))+ facet_grid(IND
    facet_grid(IND ~ var_type) + theme(legend.position = "bottom") + #guides(fill=guide_legend(nrow=2, byrow=TRU
E))+ ~
    facet_grid(IND ~ var_type) + theme(legend.position = "bottom") + #guides(fill=guide_legend(nrow=2, byrow=TRU
E))+ var_type)
    facet_grid(IND ~ var_type) + theme(legend.position = "bottom") + #guides(fill=guide_legend(nrow=2, byrow=TRU
E))+ +
    facet_grid(IND ~ var_type) + theme(legend.position = "bottom") + #guides(fill=guide_legend(nrow=2, byrow=TRU
E))+ theme(legend.position
    facet_grid(IND ~ var_type) + theme(legend.position = "bottom") + #guides(fill=guide_legend(nrow=2, byrow=TRU
E))+ =
    facet_grid(IND ~ var_type) + theme(legend.position = "bottom") + #guides(fill=guide_legend(nrow=2, byrow=TRU
E))+ "bottom")
    facet_grid(IND ~ var_type) + theme(legend.position = "bottom") + #guides(fill=guide_legend(nrow=2, byrow=TRU
E))+ +
    facet_grid(IND ~ var_type) + theme(legend.position = "bottom") + #guides(fill=guide_legend(nrow=2, byrow=TRU
E))+ #guides(fill=guide_legend(nrow=2,
    facet_grid(IND ~ var_type) + theme(legend.position = "bottom") + #guides(fill=guide_legend(nrow=2, byrow=TRU
E))+ byrow=TRUE))+
theme_classic() + theme(axis.text.x = element_text(size = 9, angle = 45))

pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/4b.pdf")
g1
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
g1
```

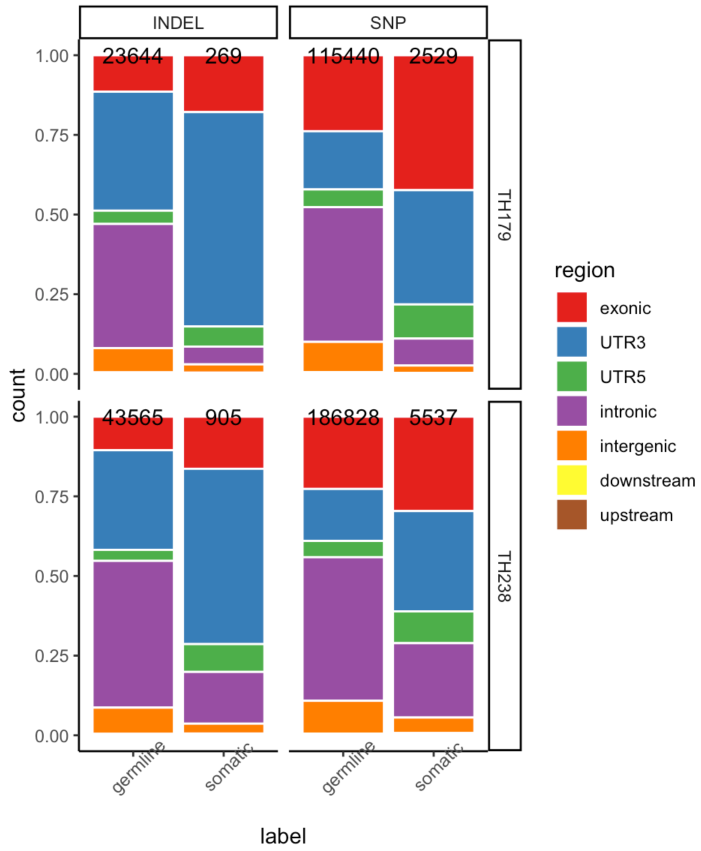
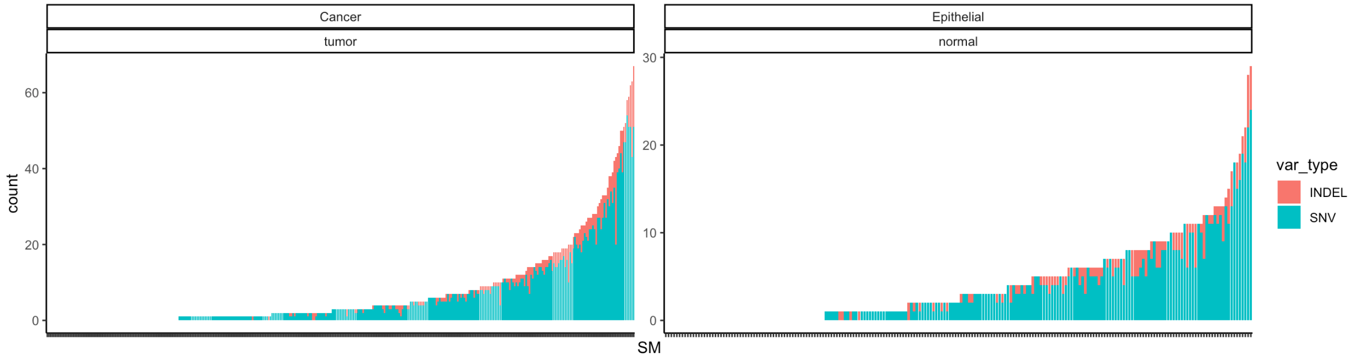
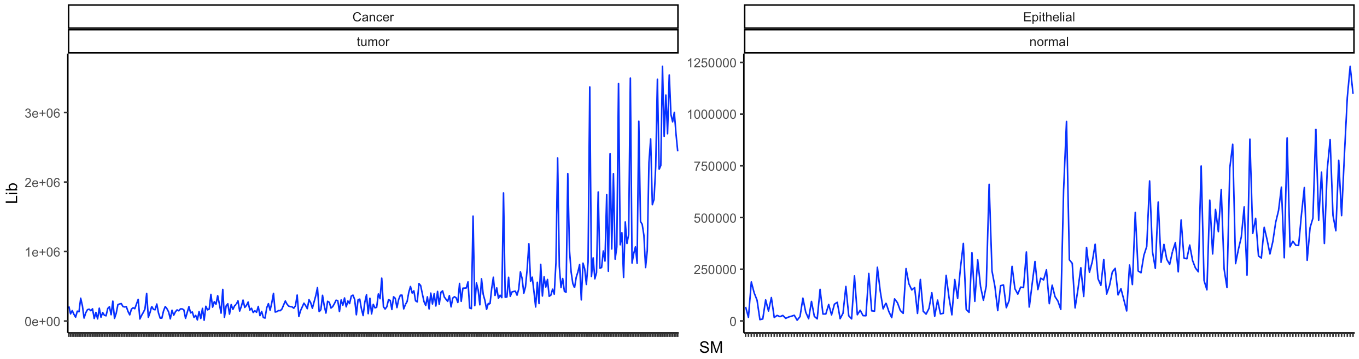


Fig. 4d - Event Count (Fig. 4d)

```
df3c <- read.csv("v2.linkage_events.ALL_INDV.counts.txt", sep = "\t", header = F)
names(df3c) <- c("SM", "var_type", "count", "total", "IND", "Tissue", "Condition", "Lib")

df3cu <- unique(df3c[c("SM", "total")])
df3cu <- arrange(df3cu, total)

df3c$SM <- factor(df3c$SM, levels = df3cu$SM)


g1 <- ggplot(df3c, aes(y = count, x = SM)) + geom_bar(stat = "identity", aes(fill = var_type)) +
    facet_wrap(Tissue ~ Condition, scales = "free") + theme_classic() + theme(axis.text.x = element_blank())

df3c$comv <- paste0(df3c$Condition, "-", df3c$Tissue)

g2 <- ggplot(df3c, aes(y = Lib, x = SM, group = comv)) + geom_line(color = "blue") + facet_wrap(Tissue ~
    Condition, scales = "free") + theme_classic() + theme(axis.text.x = element_blank())

g1
```



```
g2
```



```
pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/4d.pdf")
g1
g2
dev.off()
```

```
## quartz_off_screen
##                 2
```

Fig. 4e - Event Count downsampling (Fig. 4e)

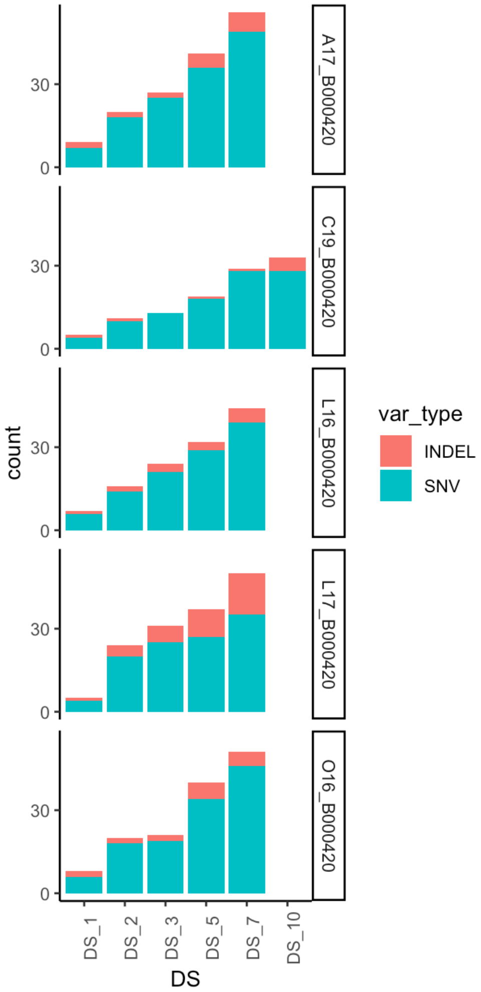
```
df3c <- read.csv("v2.linkage_events.DownSampled.counts.txt", sep = "\t", header = F)
names(df3c) <- c("SM", "var_type", "count", "total", "DS")
head(df3c)
```

SM	var_type	count	total	DS
A17_B000420	SNV	7	9	DS_1
A17_B000420	INDEL	2	9	DS_1
L17_B000420	SNV	4	5	DS_1
L17_B000420	INDEL	1	5	DS_1
L16_B000420	SNV	6	7	DS_1
L16_B000420	INDEL	1	7	DS_1

```
df3c$group = paste0(df3c$SM, df3c$var_type)
df3c$DS <- factor(df3c$DS, levels = c("DS_1", "DS_2", "DS_3", "DS_5", "DS_7", "DS_10", "DS_15",
    "DS_20", "whole"))

g1 <- ggplot(df3c, aes(y = count, x = DS, group = group)) + geom_bar(stat = "identity", aes(fill = var_type)) +
    facet_grid(SM ~ .) + scale_y_continuous(breaks = c(0, 30, 60)) + theme_classic() + theme(axis.text.x = elemen
t_text(angle = 90))

pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/4e.pdf")
g1
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
g1
```

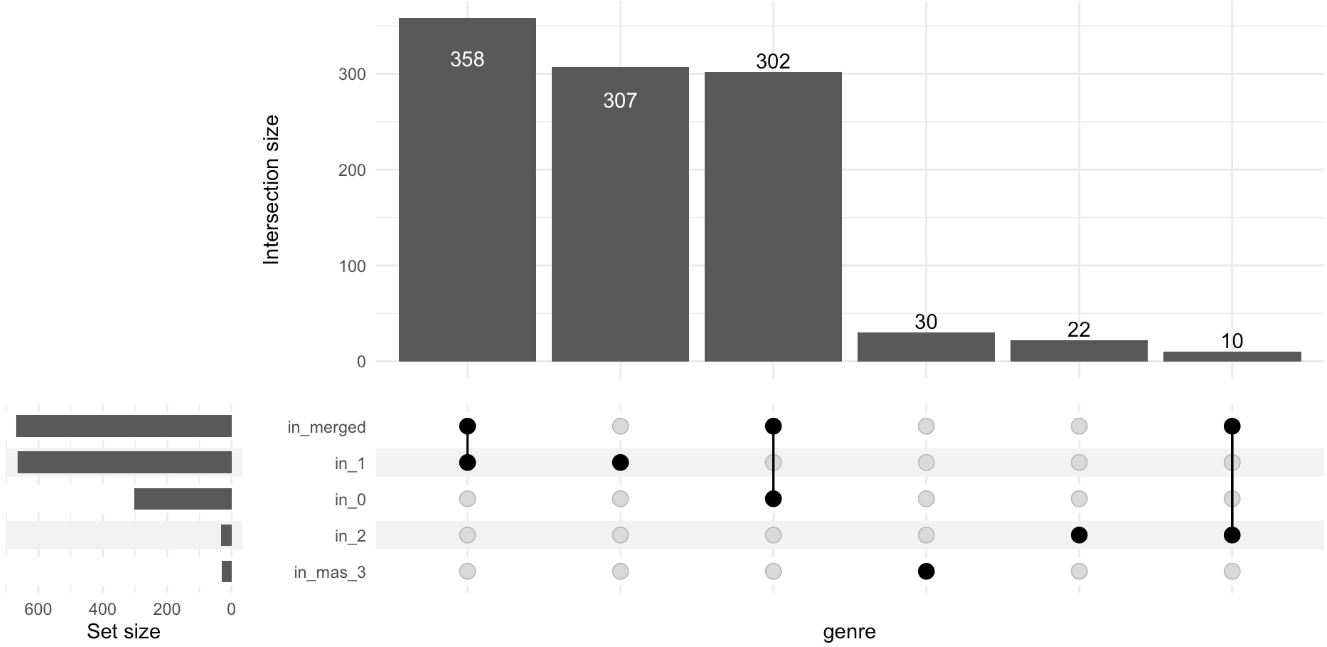


### Fig. 4e - Overlap with joint-called linkage events

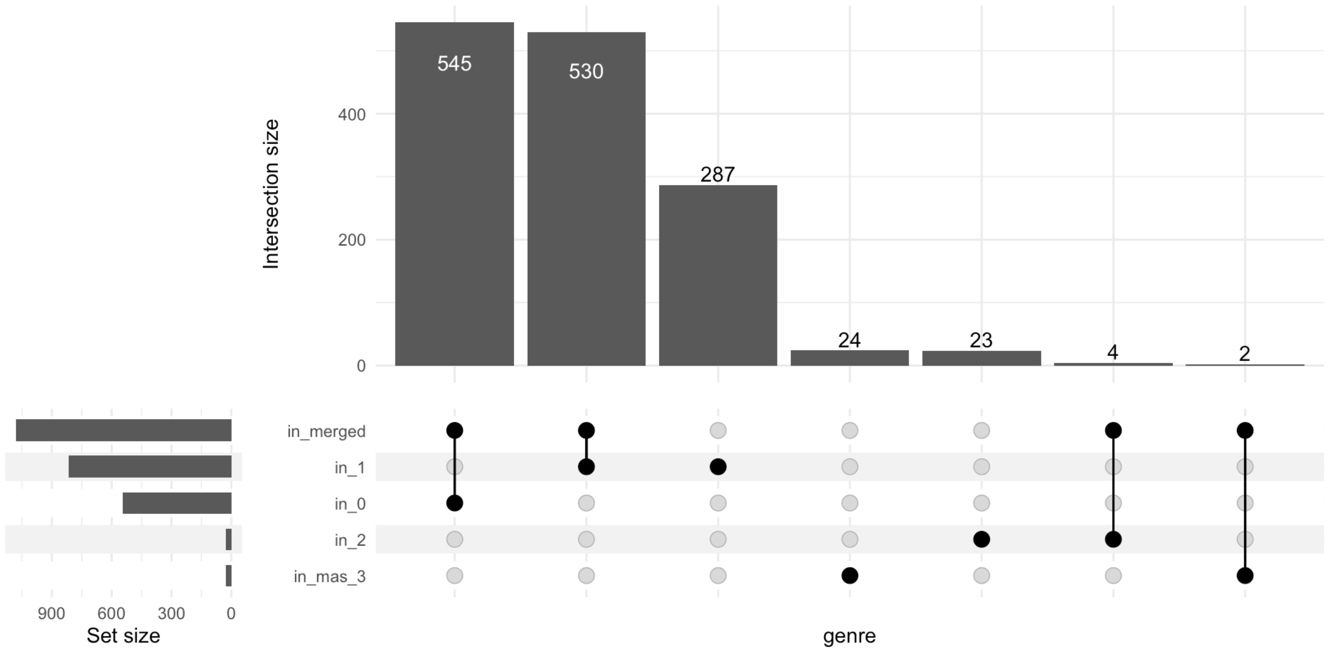
```
df3a <- read.csv("v2.linkage_matrix.ALL_INDV.merged_vs_single.txt", sep = "\t", header = F)

names(df3a) <- c("indv", "event", "in_0", "in_1", "in_2", "in_mas_3", "in_merged")

df3a <- df3a[df3a$in_0 == 0 | df3a$in_merged == 1, ]
head(df3a)
```

indv	event	in_0	in_1	in_2	in_mas_3	in_merged
TH179.Cancer-tumor	('1375184-1375185', <GenomicInterval object 'chr1', [1371201,1372305), strand '.'>)	1	0	0	0	1
TH179.Cancer-tumor	('2280239-2280240', <GenomicInterval object 'chr1', [2280136,2280159), strand '.'>)	0	1	0	0	1
TH179.Cancer-tumor	('8021777-8021778', <GenomicInterval object 'chr1', [8021854,8022822), strand '.'>)	0	1	0	0	0
TH179.Cancer-tumor	('11810392-11810393', <GenomicInterval object 'chr1', [11808667,11810132), strand '.'>)	0	1	0	0	1
TH179.Cancer-tumor	('12638789-12638790', <GenomicInterval object 'chr1', [12638985,12639320), strand '.'>)	0	1	0	0	1
TH179.Cancer-tumor	('12638789-12638790', <GenomicInterval object 'chr1', [12639441,12640548), strand '.'>)	0	1	0	0	1

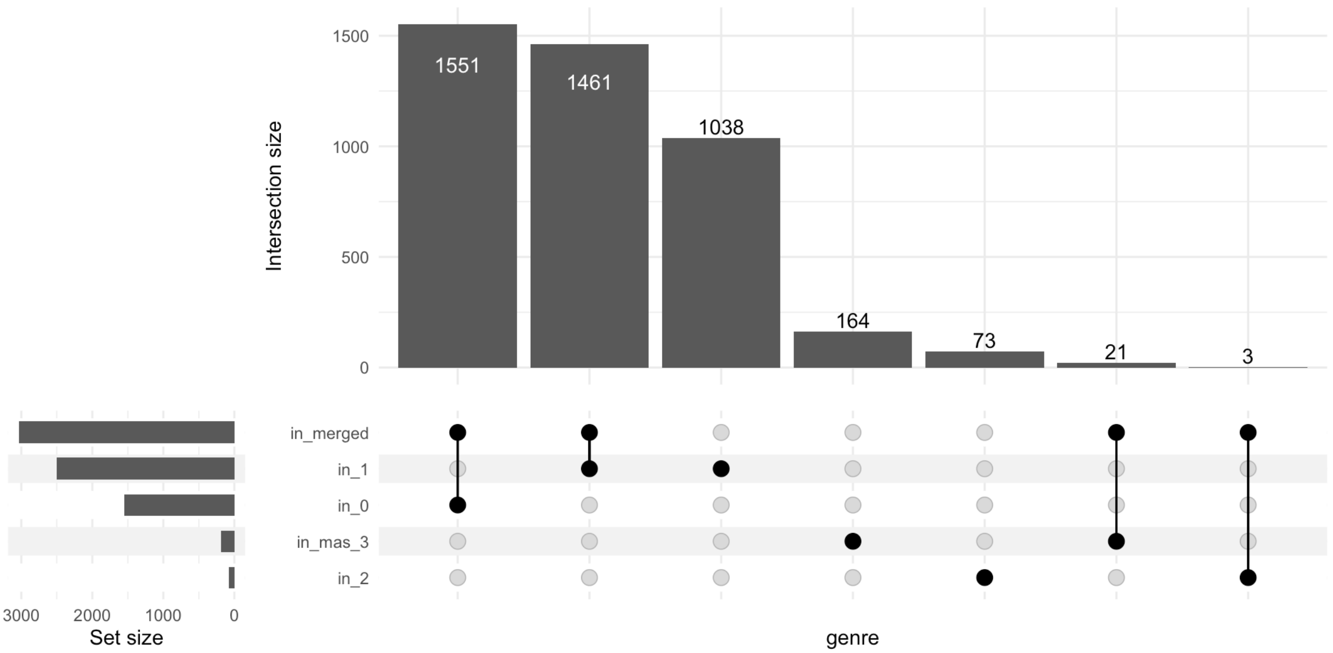
```
vars <- c("in_0", "in_1", "in_2", "in_mas_3", "in_merged")

intersections = list(c("in_0"), c("in_1"), c("in_2"), c("in_mas_3"), c("in_merged"), c("in_0", "in_merged"),
    c("in_1", "in_merged"), c("in_2", "in_merged"), c("in_mas_3", "in_merged"))

p1 = upset(df3a[df3a$indv == "TH179.Cancer-tumor", ], vars, name = "genre", keep_empty_groups = TRUE,
    width_ratio = 0.2)

p2 = upset(df3a[df3a$indv == "TH179.Epithelial-normal", ], vars, name = "genre", keep_empty_groups = TRUE,
    width_ratio = 0.2)

p3 = upset(df3a[df3a$indv == "TH238.Cancer-tumor", ], vars, name = "genre", keep_empty_groups = TRUE,
    width_ratio = 0.2)

p4 = upset(df3a[df3a$indv == "TH238.Epithelial-normal", ], vars, name = "genre", keep_empty_groups = TRUE,
    width_ratio = 0.2)

print(p1)
```



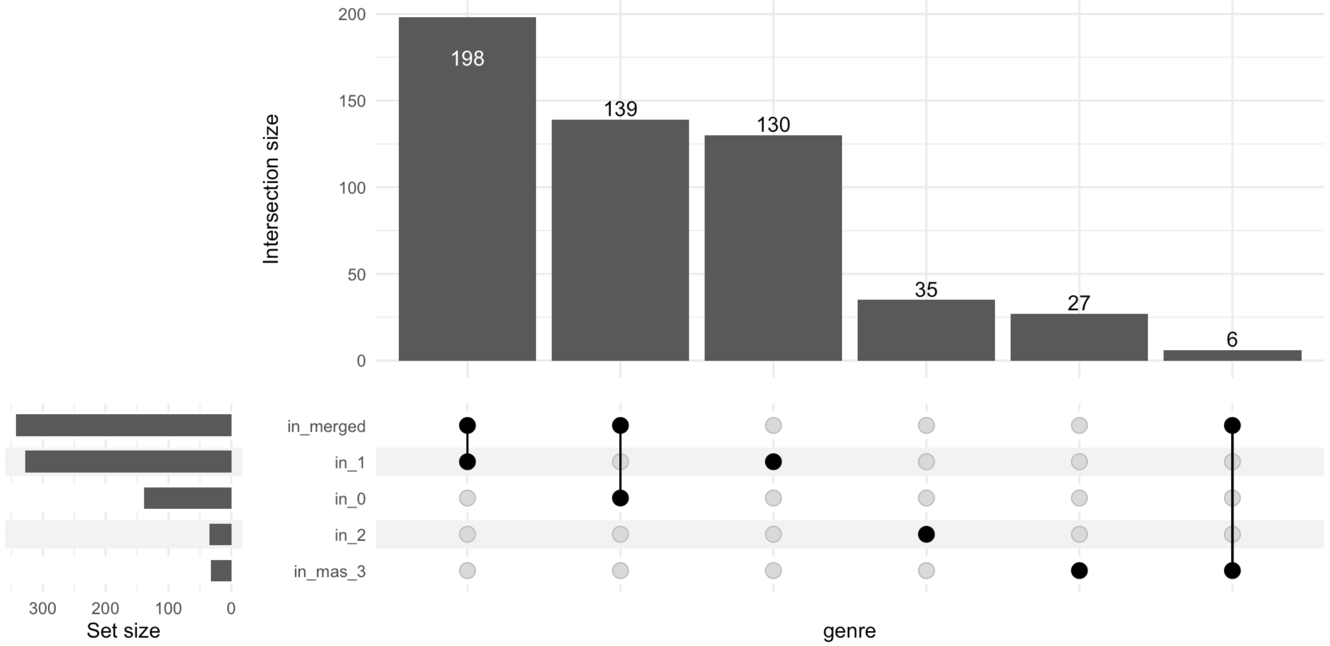
```
print(p2)
```



```
print(p3)
```



```
print(p4)
```



```
pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/4f.pdf")
print(p1)
print(p2)
print(p3)
print(p4)
dev.off()
```

```
## quartz_off_screen
##                 2
```

Eventprevalnce

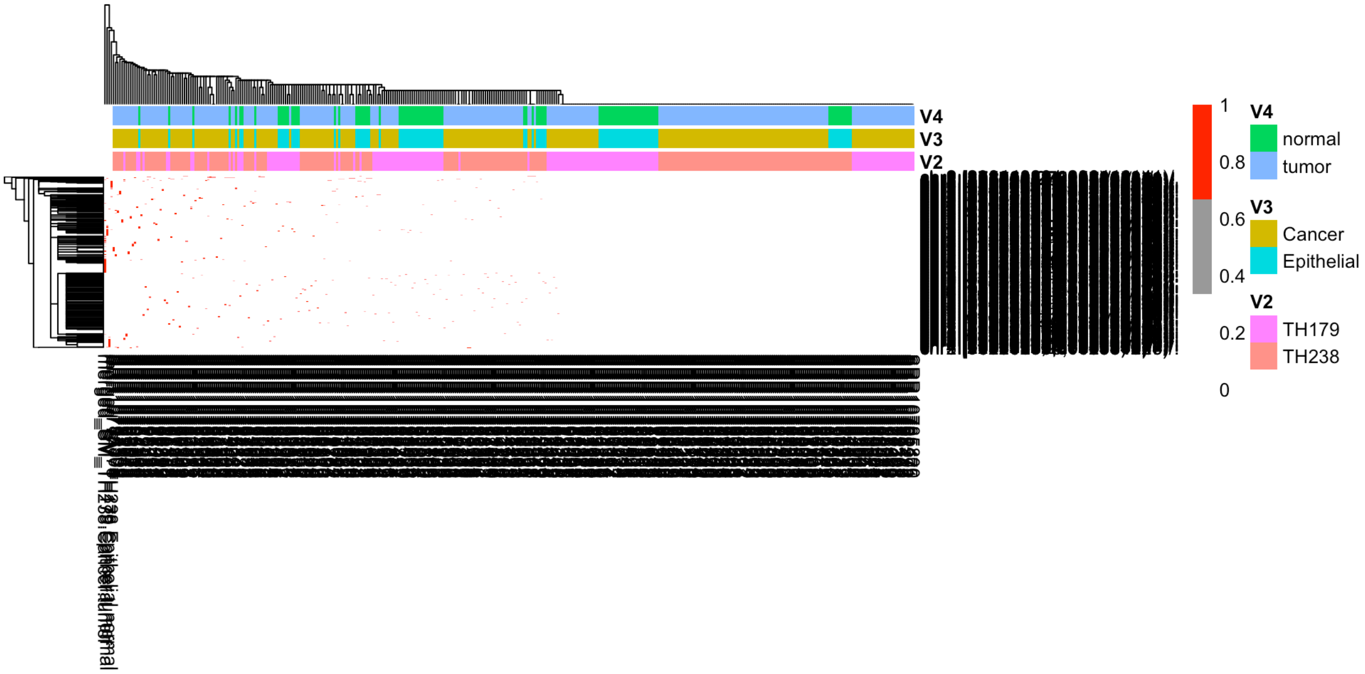
```
annot = read.csv("v2.linkage_matrix.ALL_INDV.annot.txt", sep = " ", header = F)
head(annot)
```

V1	V2	V3	V4
SRR10778546	TH179	Cancer	tumor
SRR10795475	TH179	Cancer	tumor
SRR10796628	TH179	Cancer	tumor
SRR10793483	TH179	Cancer	tumor
SRR10785800	TH179	Cancer	tumor
SRR10781571	TH179	Cancer	tumor

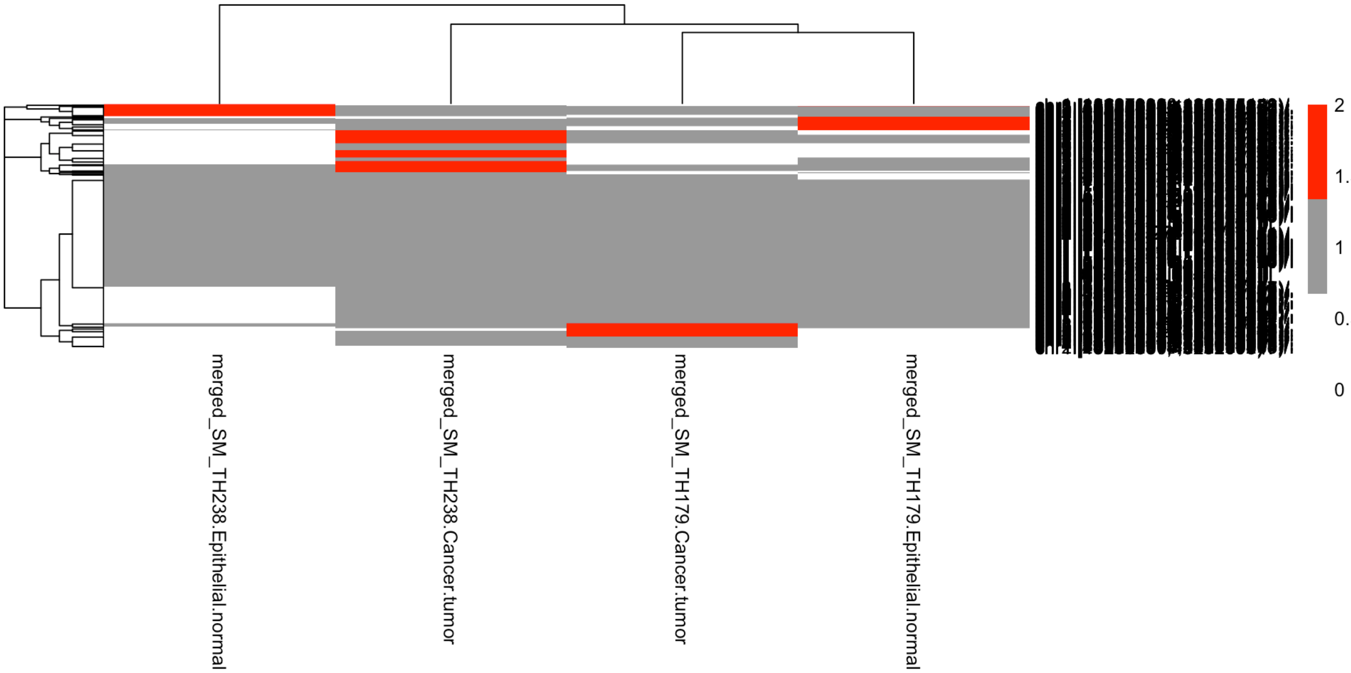
```
mat <- read.csv("v2.linkage_matrix.ALL_INDV.txt", sep = " ", header = T)

rownames(mat) <- mat$event
mat <- as.matrix(mat[, 2:ncol(mat)])
mat[1:10, 1:10]
```

```
##                        SRR10778546 SRR10795475 SRR10796628 SRR10793483
## chr1:[880527,880896)/.           0           0           0           0
## chr1:[881034,881327)/.           0           0           0           0
## chr1:[881667,881780)/.           0           0           0           0
## chr1:[881926,883510)/.           0           0           0           0
## chr1:[883613,883868)/.           0           0           0           0
## chr1:[883984,886506)/.           0           0           0           0
## chr1:[886619,887378)/.           0           0           0           0
## chr1:[887521,887790)/.           0           0           0           0
## chr1:[888669,889161)/.           0           0           0           0
## chr1:[889273,889383)/.           0           0           0           0
##                        SRR10785800 SRR10781571 SRR10792298 SRR10795413
## chr1:[880527,880896)/.           0           0           0           0
## chr1:[881034,881327)/.           0           0           0           0
## chr1:[881667,881780)/.           0           0           0           0
## chr1:[881926,883510)/.           0           0           0           0
## chr1:[883613,883868)/.           0           0           0           0
## chr1:[883984,886506)/.           0           0           0           0
## chr1:[886619,887378)/.           0           0           0           0
## chr1:[887521,887790)/.           0           0           0           0
## chr1:[888669,889161)/.           0           0           0           0
## chr1:[889273,889383)/.           0           0           0           0
##                        SRR10797178 SRR10783077
## chr1:[880527,880896)/.           0           0
## chr1:[881034,881327)/.           0           0
## chr1:[881667,881780)/.           0           0
## chr1:[881926,883510)/.           0           0
## chr1:[883613,883868)/.           0           0
## chr1:[883984,886506)/.           0           0
## chr1:[886619,887378)/.           0           0
## chr1:[887521,887790)/.           0           0
## chr1:[888669,889161)/.           0           0
## chr1:[889273,889383)/.           0           0
```

```
rownames(annot) <- annot$V1
annot <- annot[, 2:ncol(annot)]

mat <- mat[rowSums(mat == 2) >= 1, ]
mat <- mat[, colSums(mat >= 1) > 10]

mat2 <- mat[, colnames(mat) %in% c("merged_SM_TH179.Cancer.tumor", "merged_SM_TH238.Cancer.tumor",
    "merged_SM_TH179.Epithelial.normal", "merged_SM_TH238.Epithelial.normal")]

pheatmap((mat == 2) * 1, color = c("white", "gray60", "red"), annotation = annot)
```



```
pheatmap(mat2, color = c("white", "gray60", "red"))
```


merged_SM_TH238.Epithelial.normal
merged_SM_TH238.Cancer.tumor
merged_SM_TH179.Cancer.tumor
merged_SM_TH179.Epithelial.normal