Histology distribution of putative onocgene annotated fusions

Code

- Show All Code
- Hide All Code
-
- Download Rmd

Histology distribution of putative onocgene annotated fusions

K S Gaonkar

Putative Driver (only oncogene annotated) : Filtering for general cancer specific genes Fusions with genes in either onco Fusions distribution for filtering criteria removed

This notebook assumes you are in OpenPBTA-analysis project folder structure.

```
#rootdir
root_dir <- rprojroot::find_root(rprojroot::has_dir(".git"))

####load required packages
suppressPackageStartupMessages(library("readr"))
suppressPackageStartupMessages(library("tidyverse"))
```

```
Warning: package 'tidyverse' was built under R version 3.5.2

Warning: package 'ggplot2' was built under R version 3.5.2

Warning: package 'tibble' was built under R version 3.5.2

Warning: package 'tidyr' was built under R version 3.5.2

Warning: package 'purrr' was built under R version 3.5.2
```

```
Warning: package 'dplyr' was built under R version 3.5.2

Warning: package 'stringr' was built under R version 3.5.2

Warning: package 'forcats' was built under R version 3.5.2
```

```
suppressPackageStartupMessages(library("reshape2"))
suppressPackageStartupMessages(library("qdapRegex"))

####read filtFusion files
strandedQCGeneFiltered_filtFusion<-readRDS(file.path(root_dir, params$dataStranded))
polyaQCGeneFiltered_filtFusion<-readRDS(file.path(root_dir, params$dataPolya))

####read files from results folder
outputfolder<-params$outputfolder
QCGeneFiltered_filtFusion<-rbind(strandedQCGeneFiltered_filtFusion,polyaQCGeneFiltered_filtF

fusion_calls<-unique(QCGeneFiltered_filtFusion)
#### remove distance from intergenic fusions
fusion_calls$FusionName<-unlist(lapply(fusion_calls$FusionName,function(x) rm_between(x, "('

#### get histology file
clinical<-read.delim(file.path(root_dir, params$histology), stringsAsFactors = FALSE)
clinical<-clinical[,c("Kids_First_Biospecimen_ID","Kids_First_Participant_ID","broad_histolo

#### get count cutoff for histology
countHistology<-params$countHistology
```

Format and filter

```
#aggregate caller
fusion_caller.summary <- fusion_calls %>%
  dplyr::select(Sample,FusionName,Caller,Fusion_Type) %>%
  group_by(FusionName, Sample ,Fusion_Type) %>%
  unique() %>%
  dplyr::mutate(CalledBy = toString(Caller), caller.count = n()) %>%
  dplyr::select(-Caller)

#to add aggregated caller from fusion_caller.summary
fusion_calls<-fusion_calls %>%
  left_join(fusion_caller.summary,by=(c("Sample","FusionName","Fusion_Type"))) %>%
  dplyr::select(-JunctionReadCount,-SpanningFragCount,-Confidence,-LeftBreakpoint,-RightBrea

#merge with histology file
fusion_calls<-merge(fusion_calls,clinical,by.x="Sample",by.y="Kids_First_Biospecimen_ID")
```

```r
#filter for putative driver genes
putative_driver_annotated_fusions <- fusion_calls %>%
  dplyr::select(-Caller,-annots) %>%
  unique() %>%
  dplyr::filter(!is.na(Gene1A_anno) | !is.na(Gene1B_anno) | !is.na(Gene2A_anno) | !is.na(Ger
  unique()

#filter other fusion genes
putative_driver_annotated_other_fusions <- fusion_calls %>%
  dplyr::select(-Caller,-annots) %>%
  unique() %>%
  dplyr::filter(!is.na(Gene1A_anno) | !is.na(Gene1B_anno) | !is.na(Gene2A_anno) | !is.na(Ger
  dplyr::filter(Fusion_Type=="other") %>%
  unique()

#local rearrangements
putative_driver_annotated_fusions_local<-fusion_calls %>%
  # local rearrangement/adjacent genes
  dplyr::filter(grepl("LOCAL_REARRANGEMENT|LOCAL_INVERSION",annots)) %>%
  dplyr::select(-Caller,-annots) %>%
  unique() %>%
  dplyr::filter(!is.na(Gene1A_anno) | !is.na(Gene1B_anno) | !is.na(Gene2A_anno) | !is.na(Ger
  unique()

#function to plot fusion found in N histology
#standardFusionCalls: standardized fusion calls
#filterN: filter to plot fusions found in more than filterN histologies
plotNhist<-function(standardFusionCalls=standardFusionCalls,filterN=filterN){
plotNhist<-standardFusionCalls %>% dplyr::select(FusionName,broad_histology) %>% unique() %>
plotNhist_total_count<-standardFusionCalls %>% dplyr::select(FusionName) %>% group_by(Fusion

plotNhist<-plotNhist %>% left_join(plotNhist_total_count,by=c("FusionName"))
plotNhist$FusionNameTotal<-paste0(plotNhist$FusionName,"(",plotNhist$totalcount,")")
plotNhist<-plotNhist[order(plotNhist$count,decreasing = TRUE),]

if (!is_empty(filterN)){
  plotNhist<-plotNhist[plotNhist$count>filterN,]
}

plotNhist$FusionNameTotal<-factor(plotNhist$FusionNameTotal,levels=unique(plotNhist$FusionNa
ggplot(plotNhist)+geom_col(aes(x=plotNhist$FusionNameTotal,y=plotNhist$count))+theme(axis.te
}
```

Plots

plot "other" reading-frame fusion found in more than N (countHistology) histologies which might indicate false calls

x axis is number of histologies y axis is the fusion name (total number of calls in putative oncogene list)

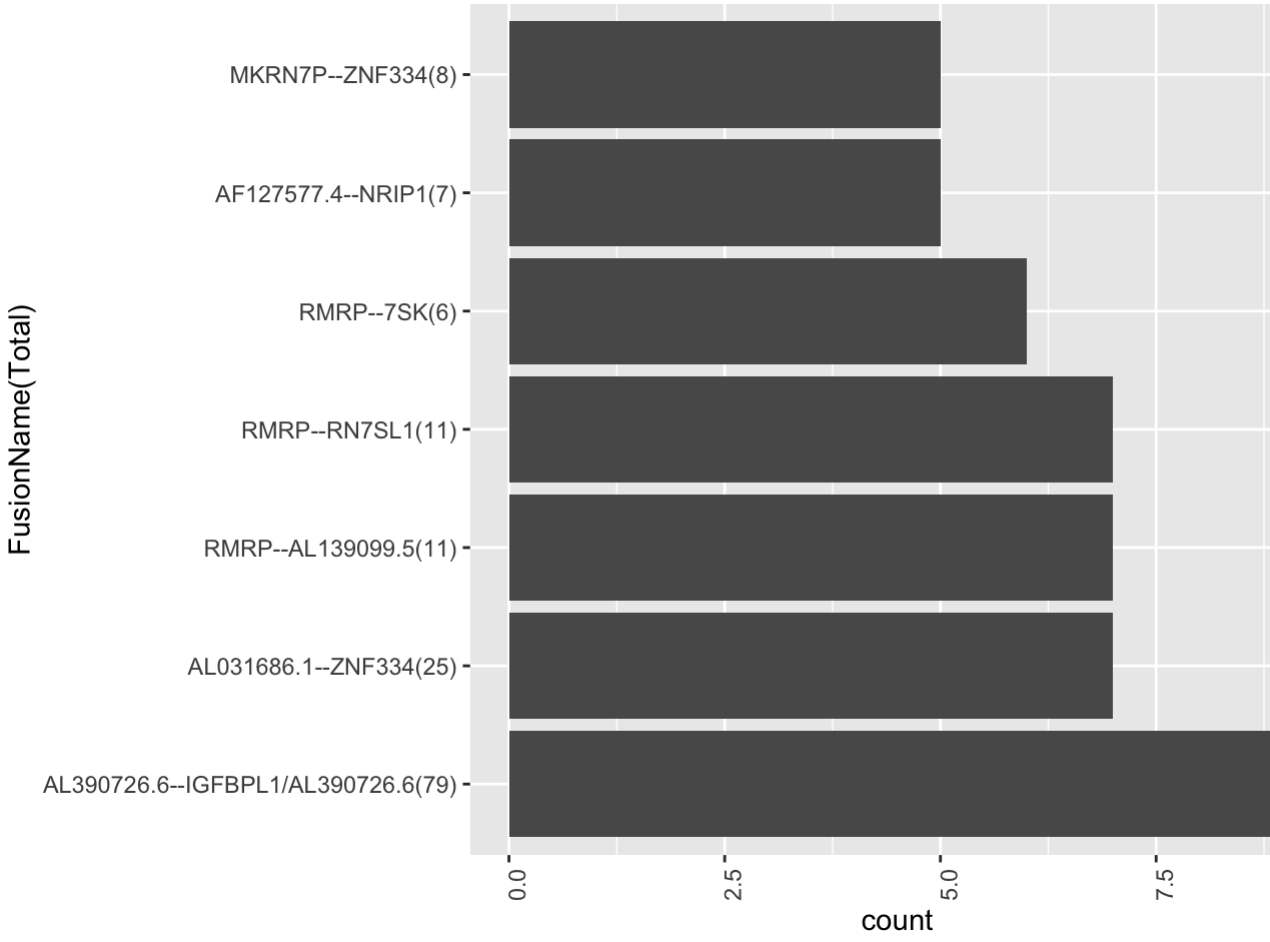
```
print( "total putative oncogene other fusion")
```

```
[1] "total putative oncogene other fusion"
```

```
nrow(putative_driver_annotated_other_fusions)
```

```
[1] 1056
```

```
plotNhist(putative_driver_annotated_other_fusions,filterN = countHistology)
```



plot fusion annotated as local rearrangements found in more than N (countHistology) histologies which might indicate false calls

x axis is number of histologies y axis is the fusion name (total number of calls in putative oncogene list)

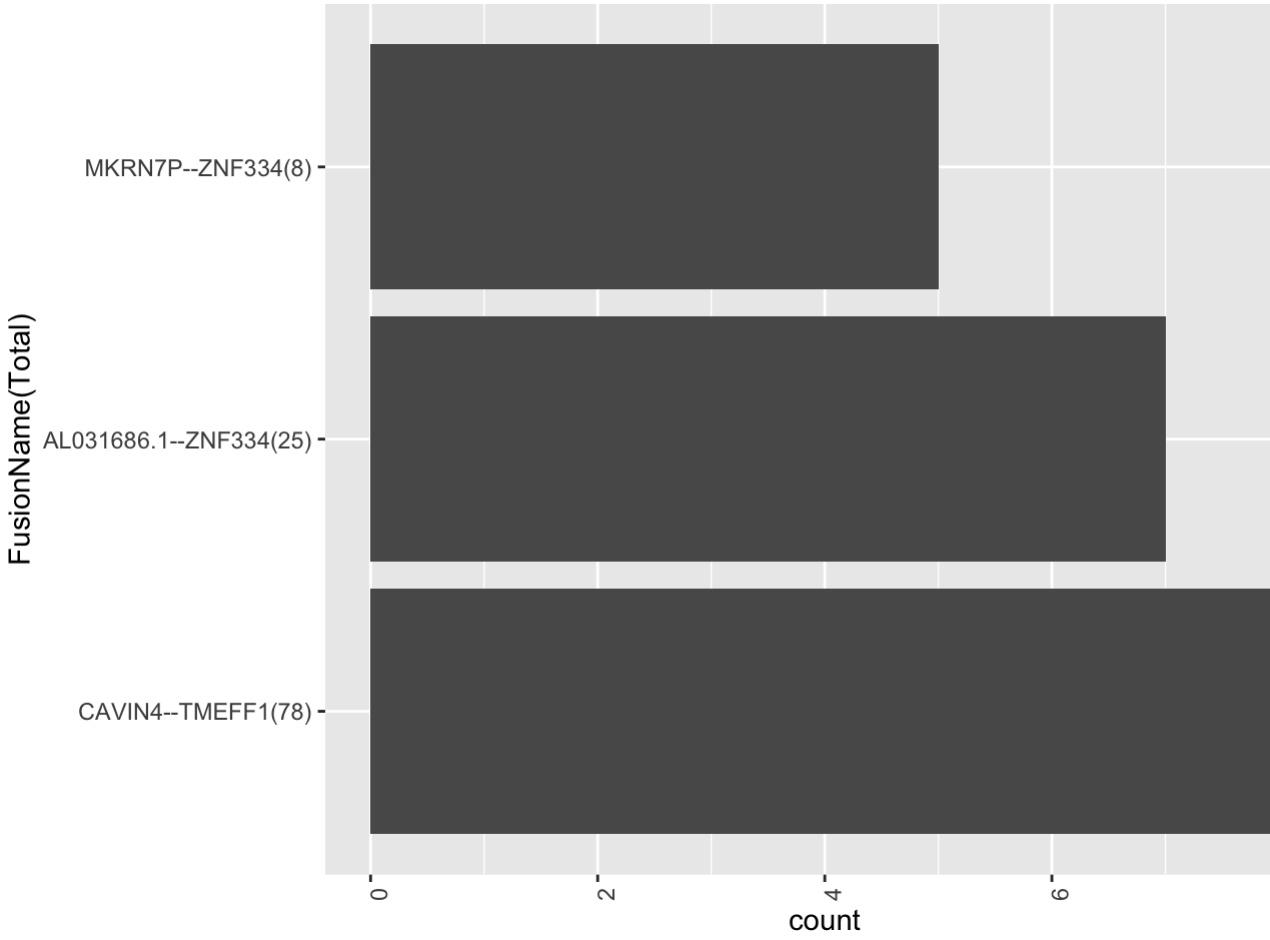
```
print( "total putative oncogene local rearrangement fusion")
```

```
[1] "total putative oncogene local rearrangement fusion"
```

```
nrow(putative_driver_annotated_fusions_local)
```

```
[1] 253
```

```
plotNhist(putative_driver_annotated_fusions_local,filterN = countHistology)
```



plot all putative oncogene fusion found in more than N (countHistology) histologies which might indicate false calls

x axis is number of histologies y axis is the fusion name (total number of calls in putative oncogene list)

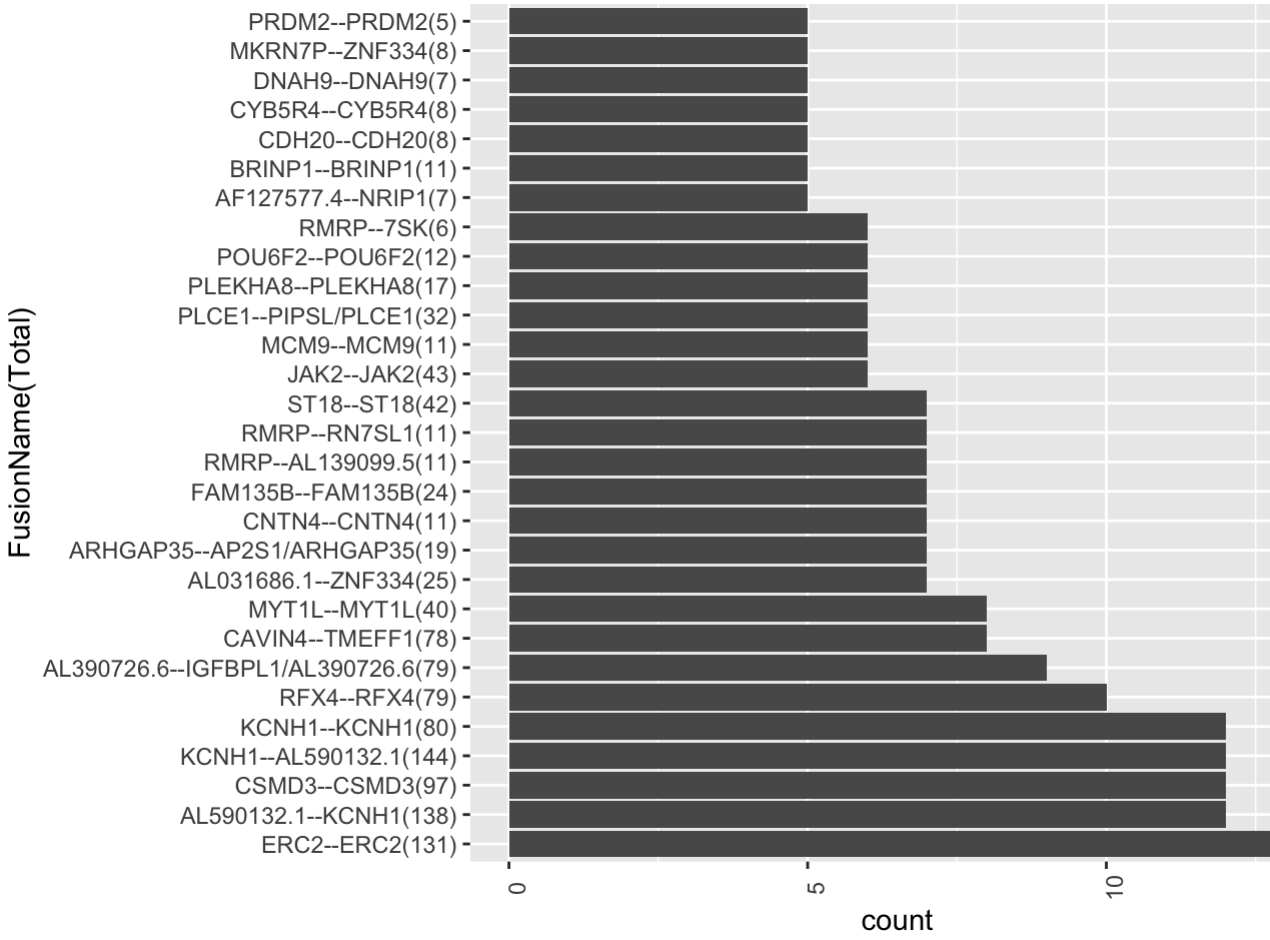
```
print( "total putative oncogene fusion")
```

```
[1] "total putative oncogene fusion"
```

```
nrow(putative_driver_annotated_fusions)
```

```
[1] 4366
```

```
plotNhist(putative_driver_annotated_fusions, filterN = countHistology)
```




```
# count number of fusions in putative oncogene annotated fused gene are in more than N (cou
FusionInNhist<-putative_driver_annotated_fusions %>% dplyr::select(FusionName,broad_histolog
```

```
FusionInNhist<-FusionInNhist[FusionInNhist$count>countHistology,]
FusionInNhist
```

```
putative_driver_annotated_fusions %>% filter(FusionName %in% FusionInNhist$FusionName) %>% r
```

```
[1] 1184
```

LS0tCnRpdGxlOiAiSGlzdG9sb2d5IGRpc3RyaWJ1dGlvbiBvZiBwdXRhdGl2ZSBvbm9jZ2VuZSBhbm5vdGF0