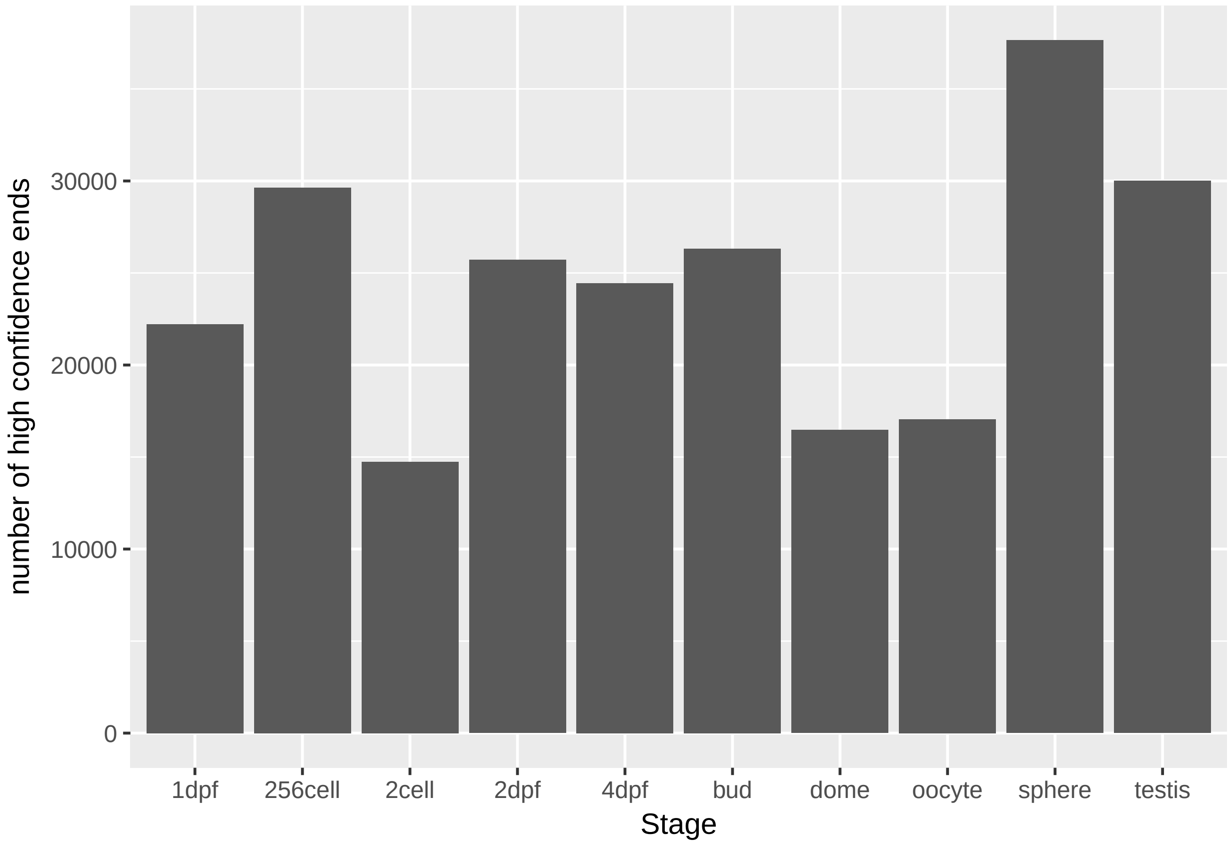
comparing counting windows from different stages

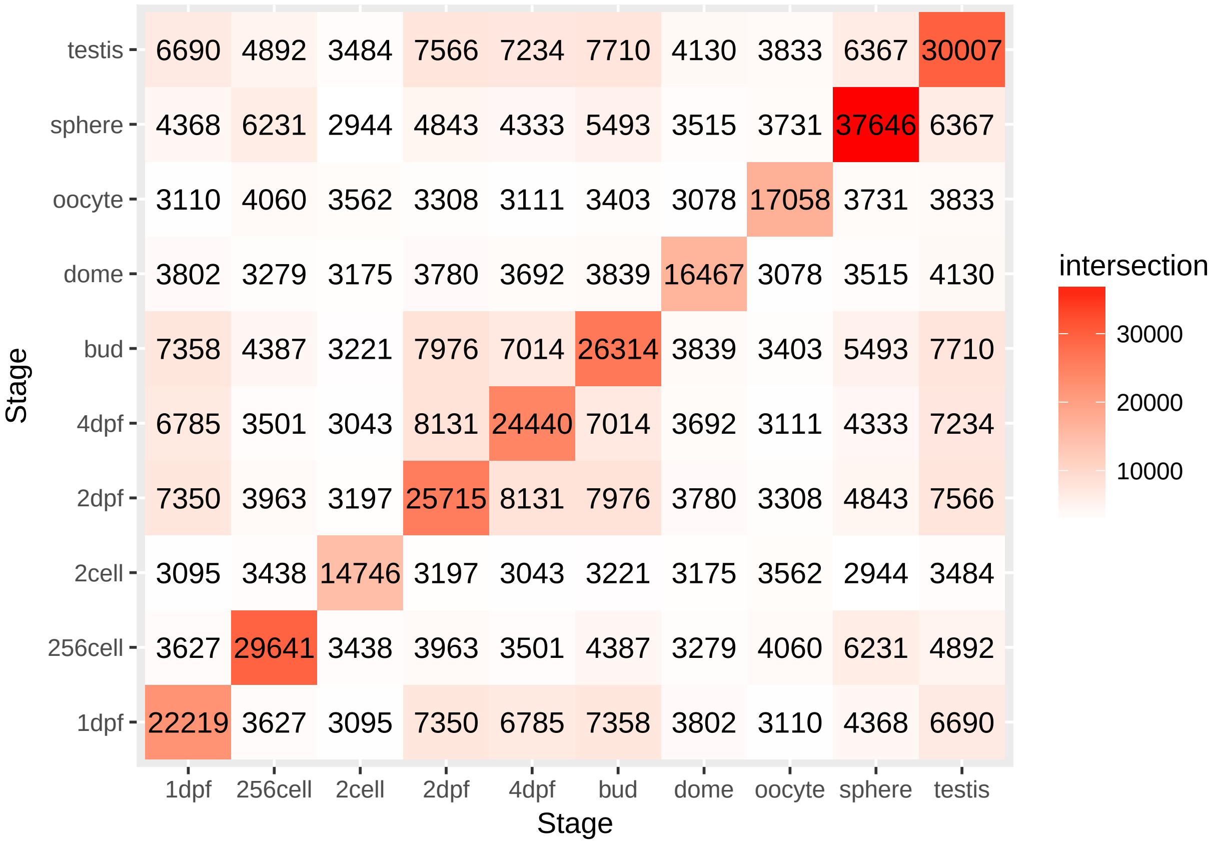
Pooja Bhat

July 19, 2017

R Markdown



```
## pdf
##   2
```



```
## pdf
##   2
```

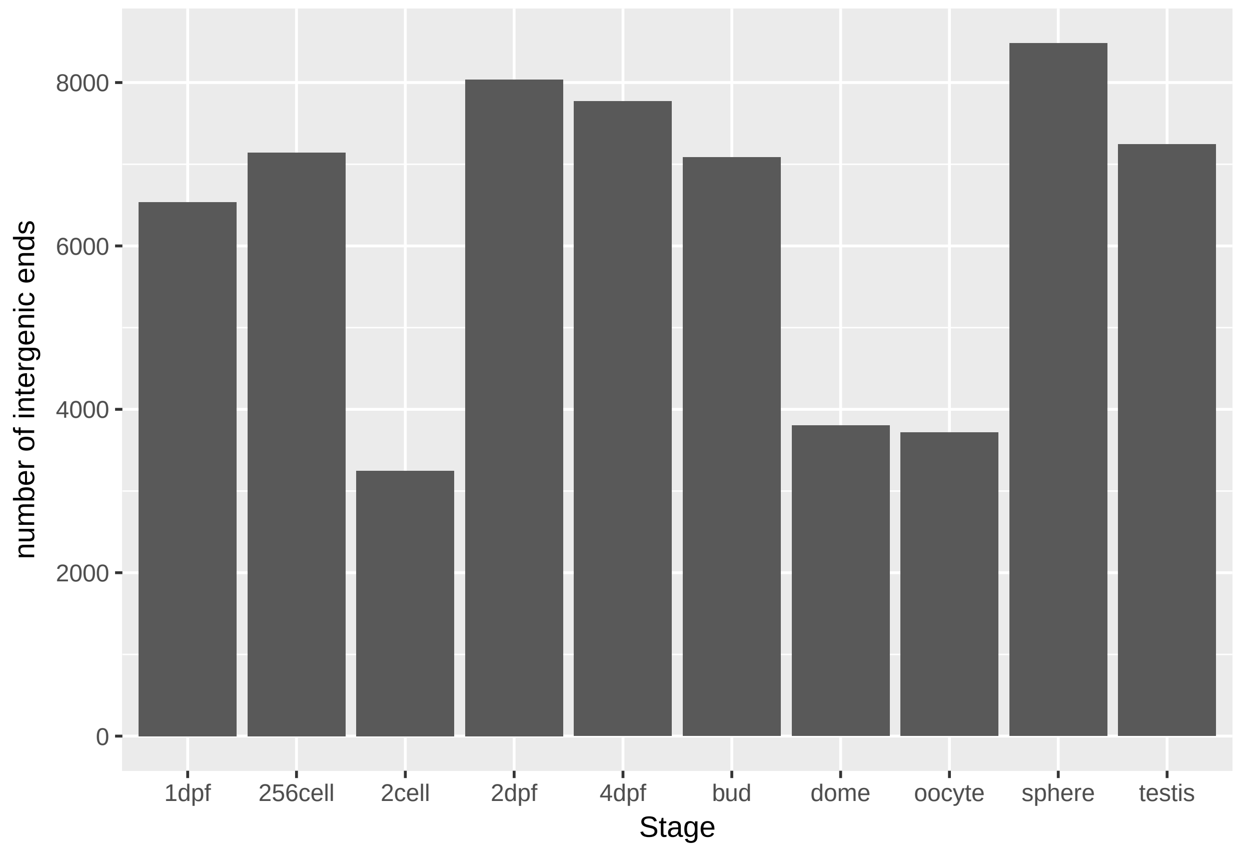
I wanted to enxt compare the high confidence intergenic ends that we identify in the different stages. This is still very exploratory and I want to see the deviation in the data we have.

```
completePath_intergenicEnds = paste0(completePath,"onlyIntergenic_90percent_n100.bed")
intergenicEnds = lapply(completePath_intergenicEnds,function(x) read.delim(x,stringsAsFactors = F,heade
names(intergenicEnds) = stages



library(reshape)
nuMberOfIntergenicEnds = melt(lapply(intergenicEnds,nrow))
colnames(nuMberOfIntergenicEnds) = c("nuMberOfIntergenicEnds","stage")
write.table(nuMberOfIntergenicEnds,"/Volumes/groups/ameres/Pooja/Projects/zebrafishAnnotation/zebrafish

library(ggplot2)
q  = ggplot(nuMberOfIntergenicEnds,aes(x=stage,y=nuMberOfIntergenicEnds)) + geom_bar(stat = "identity")
print(q)
```



looks like there is a relaitonship between the number of ends and the number of intergenic ends. This could just be a function of the number of polyA reads used in the samples or the sequencing depth. So the initial threshold to identify priming sites will probably have to be set differently for different samples.

```
path_preprocessing = paste0(path_allStages,stages,"/output/polyAmapping_allTimepoints/logs/")
preProcessingFile=c()
for(i in 1:length(path_preprocessing)){
  preProcessingFile =  c(preProcessingFile,paste0(path_preprocessing[i],list.files(path_preprocessing[i

}

preProcessingStats = lapply(preProcessingFile,function(x) read.table(x,stringsAsFactors = F))
names(preProcessingStats) = stages

preProcessingStats= lapply(preProcessingStats,function(x) x[1:5,1])

preProcessingStats_split = lapply(preProcessingStats,function(x) strsplit(x,":",T))
preProcessingStats_split = lapply(preProcessingStats_split,function(x) lapply(x,function(y) y[2]))

preProcessingStats_split_melt = melt(preProcessingStats_split)
preProcessingStats_split_melt$value = as.numeric(as.character(preProcessingStats_split_melt$value))
sampleNames = c("initialFile","adapterTrimmed","fivePrimeTrimming","polyAcontaining","finalFile")
preProcessingStats_split_melt$sample = sampleNames
#preProcessingStats_split_melt$stage = rep(stages,each = 5)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.1
```

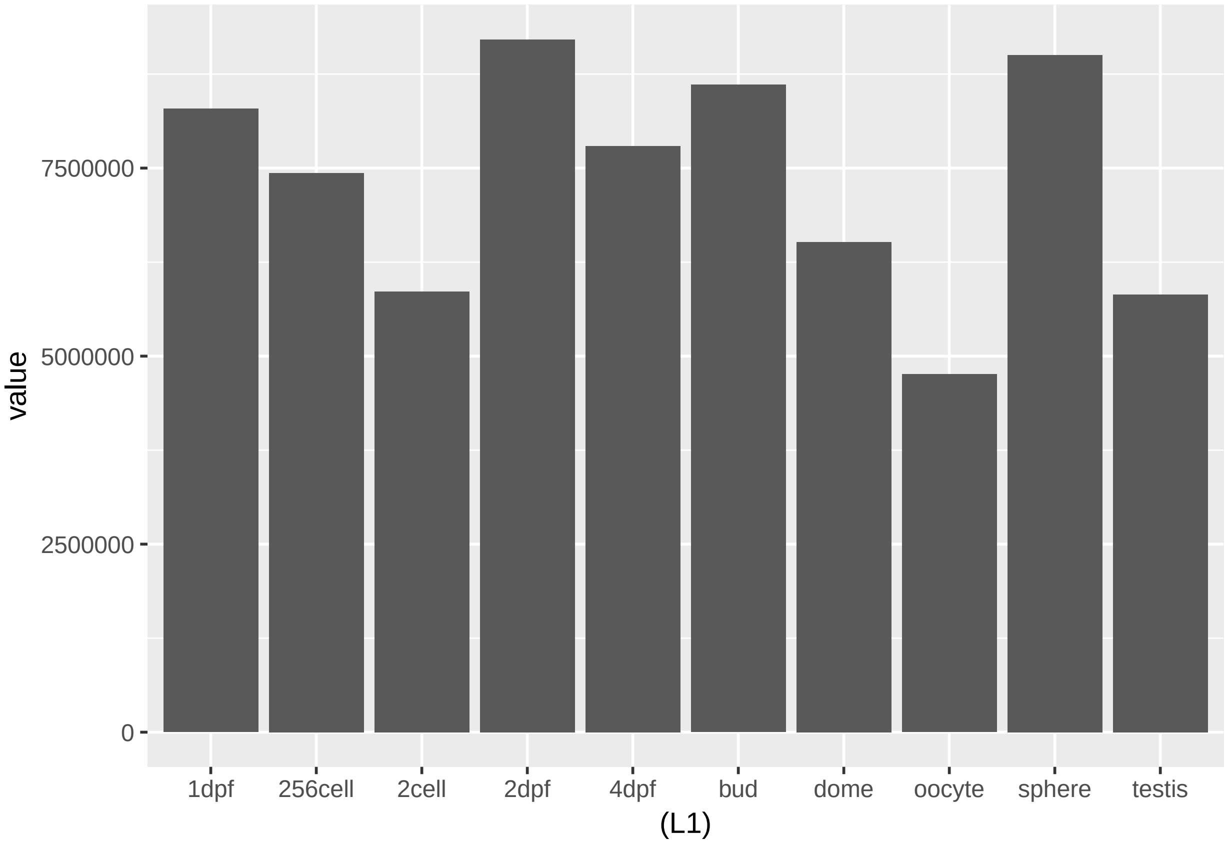
```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:reshape':
##
##     rename

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

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

```
filalFile = preProcessingStats_split_melt %>% filter(sample=="finalFile")
ggplot(filalFile,aes(x=(L1),y=value)) + geom_bar(stat="identity")
```



It looks like the number of ends we identify is based directly on the polyA readthrough. So we must either :

1. Normalize the cutoff to the number of polyA reads.
2. Consider even 1 read as indication of a priming site.