Sample 1

Matt Bixley

2020-06-24

# Background

rawcount <- read\_csv("../data/sample1.csv")  
head(rawcount)

## # A tibble: 6 x 4  
## id sgRNA T0 T20  
## <chr> <chr> <dbl> <dbl>  
## 1 sample1 mcf7\_unique:TSS100024\_-\_17662045.23-CUFF.46609.1 1383 1  
## 2 sample1 mcf7\_unique:TSS100024\_-\_17662051.23-CUFF.46609.1 1229 304  
## 3 sample1 mcf7\_unique:TSS100024\_-\_17662250.23-CUFF.46609.1 1265 602  
## 4 sample1 mcf7\_unique:TSS100024\_-\_17662267.23-CUFF.46609.1 2897 10  
## 5 sample1 mcf7\_unique:TSS100024\_+\_17661984.23-CUFF.46609.1 67 264  
## 6 sample1 mcf7\_unique:TSS100024\_+\_17662123.23-CUFF.46609.1 90 0

we sent 10 samples to XYZ lab for expression profiles from NUMBER2 genes. the file is returned with the following column names id, sgRNA, T0, T20. We used the methods from (Breitling et al. 2004) in our analysis and maybe use the bioconductor package (Gentleman et al. 2004)

# Summary

## Tidying  
gene\_dat <- rawcount %>%   
 # separate the sgRNA column into multiple new columns based on the "\_" character as a delimiter   
 separate(col = sgRNA, into = c("something","TSS", "strand", "probe\_gene"), sep = "\_") %>%  
   
 # separate the column that has the probe and gene info into two columns based on the "-" character  
 separate(col = probe\_gene, into = c("probe", "name"), sep = "-") %>%   
   
 # remove wording from TSS  
 mutate(TSS = str\_remove(TSS, pattern = "unique:")) %>%   
   
 ## mutate  
 # Calculate fold change/ probe by using T20/T0  
 mutate(fold\_change = T20 / T0)  
  
head(gene\_dat)

## # A tibble: 6 x 9  
## id something TSS strand probe name T0 T20 fold\_change  
## <chr> <chr> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 sample1 mcf7 TSS1000~ - 17662045~ CUFF.4660~ 1383 1 0.000723  
## 2 sample1 mcf7 TSS1000~ - 17662051~ CUFF.4660~ 1229 304 0.247   
## 3 sample1 mcf7 TSS1000~ - 17662250~ CUFF.4660~ 1265 602 0.476   
## 4 sample1 mcf7 TSS1000~ - 17662267~ CUFF.4660~ 2897 10 0.00345   
## 5 sample1 mcf7 TSS1000~ + 17661984~ CUFF.4660~ 67 264 3.94   
## 6 sample1 mcf7 TSS1000~ + 17662123~ CUFF.4660~ 90 0 0

# join our probe data to the output  
probe <- read\_delim("../data/probes.csv", delim = "\t")  
  
#head(probe)  
  
gene\_dat <- gene\_dat %>%   
 left\_join(.,probe) %>%   
 select(-something)  
  
head(gene\_dat)

## # A tibble: 6 x 12  
## id TSS strand probe name T0 T20 fold\_change chr start end  
## <chr> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 samp~ TSS1~ - 1766~ CUFF~ 1383 1 0.000723 4 8.81e7 8.82e7  
## 2 samp~ TSS1~ - 1766~ CUFF~ 1229 304 0.247 4 8.81e7 8.82e7  
## 3 samp~ TSS1~ - 1766~ CUFF~ 1265 602 0.476 4 8.81e7 8.82e7  
## 4 samp~ TSS1~ - 1766~ CUFF~ 2897 10 0.00345 4 8.81e7 8.82e7  
## 5 samp~ TSS1~ + 1766~ CUFF~ 67 264 3.94 4 8.81e7 8.82e7  
## 6 samp~ TSS1~ + 1766~ CUFF~ 90 0 0 4 8.81e7 8.82e7  
## # ... with 1 more variable: gene <chr>

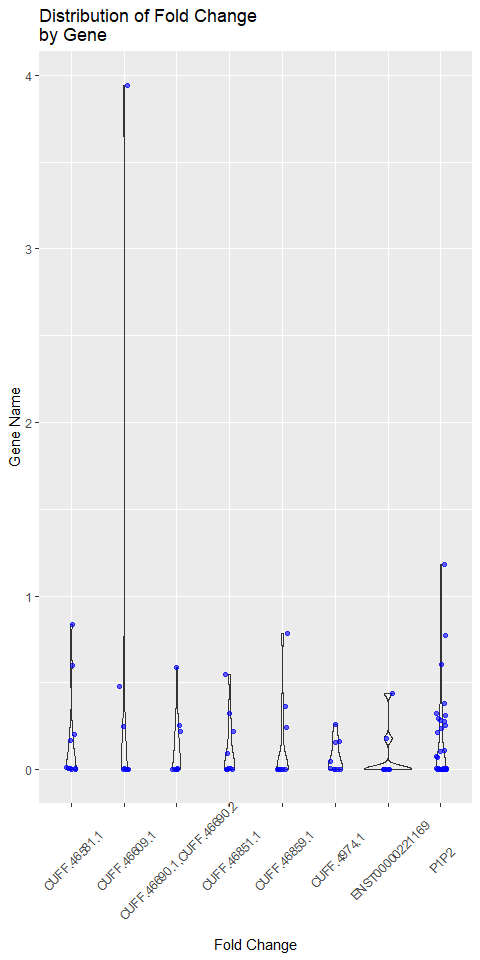
Table of Gene Summary Data

## gene summary  
gene\_summary <- gene\_dat %>%  
 group\_by(name) %>%  
 summarise(mean = mean(fold\_change, na.rm = T),   
 sd = sd(fold\_change, na.rm = T),  
 n\_probes = n())  
  
knitr::kable(gene\_summary,   
 digits = 3, # number of digits  
 align = "lccc", # column alignment  
 caption = "Summary fold change"  
 )

Summary fold change

|  |  |  |  |
| --- | --- | --- | --- |
| name | mean | sd | n\_probes |
| CUFF.46581.1 | 0.183 | 0.297 | 10 |
| CUFF.46609.1 | 0.584 | 1.367 | 8 |
| CUFF.46690.1,CUFF.46690.2 | 0.133 | 0.213 | 8 |
| CUFF.46851.1 | 0.120 | 0.188 | 10 |
| CUFF.46859.1 | 0.140 | 0.261 | 10 |
| CUFF.4974.1 | 0.070 | 0.098 | 9 |
| ENST00000221169 | 0.079 | 0.159 | 8 |
| P1P2 | 0.149 | 0.254 | 37 |

Plot Fold Change by Gene



# Discussion

best <- gene\_summary %>%   
 filter(n\_probes == max(n\_probes)) # filter to the gene with most probes

For no better reason than it’s a tidy looking name and has the most probes (n = 37), we are interested in the gene \*\*\*\*\*\*

# References

Breitling, Rainer, Patrick Armengaud, Anna Amtmann, and Pawel Herzyk. 2004. “Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.” *FEBS Lett.* 573 (1-3): 83–92. <https://doi.org/10.1016/j.febslet.2004.07.055>.

Gentleman, Robert C., Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, et al. 2004. “Bioconductor: open software development for computational biology and bioinformatics.” *Genome Biol.* 5 (10). <https://doi.org/10.1186/gb-2004-5-10-r80>.