CRISPR CNV Analysis

Keith Dunaway

March 7, 2016

Assignment

Using CRISPR library data supplied and Copy Number Variation (CNV) data of your choice, identify relationships between CNV of target site and CRISPR guide activity. This CRISPR library was conducted in the A375 melanoma cell line, so CNV and genotype information should be obtained for this cell line using published literature or cell line databases. The data source should be justified.

Background

The data provided came from the GeCKO experiment found in Shalem et al. 2014. Briefly, 64,751 unique sgRNAs (on Lentiviral expression vectors) targeting 18,080 coding genes were inserted into a population of A375 melanoma cells. The concentration of exposure was such that it optimized adding only one Lentiviral expression vector per cell. The population was then allow to grow for 7 and 14 days after transduction in the presense and absense of PLX treatment. There were two biological replicates for each condition except the starting condition (which was only one). Once harvested, the sgRNAs were isolated and sequenced.

Since the sgRNAs would target coding regions which would create indels, those corresponding genes would become disfunctional, leading to a knock out of the gene. The number of reads corresponding to essential genes decreased in the control day 7 and 14 populations compared to other genes. This was because since the essential genes were knocked out, those cells would die, decreasing the population of cells with the corresponding sgRNAs. In the PTX experiment, if the number of reads that aligned to a gene increased over time, the assumption would be that if you knock out those genes, the cells would become resistance to the PTX drug.

My job was to determine if there was a Copy Number Variation effect on the experiment. The A375 melanoma cell has 62 chromosomes (normal human cells have 46) with a vast array of large scale CNVs. I found the locations of these CNVs through **C**atalogue **o**f **S**omatic **M**utations **i**n **C**ancer (COSMIC) which can be found at cancer.sanger.ac.uk. It is a website maintained by Wellcome Trust Sanger Institute, a research institute governed by the registered charity and legal entity Genome Research Limited (GRL).

Libraries, Functions, and Loading Data

After getting data through the accompanying pipeline, I loaded and processed the final file using the following code:

```
library(reshape2)
library(plotrix)
library(plyr)
library(ggplot2)
library(vegan)
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

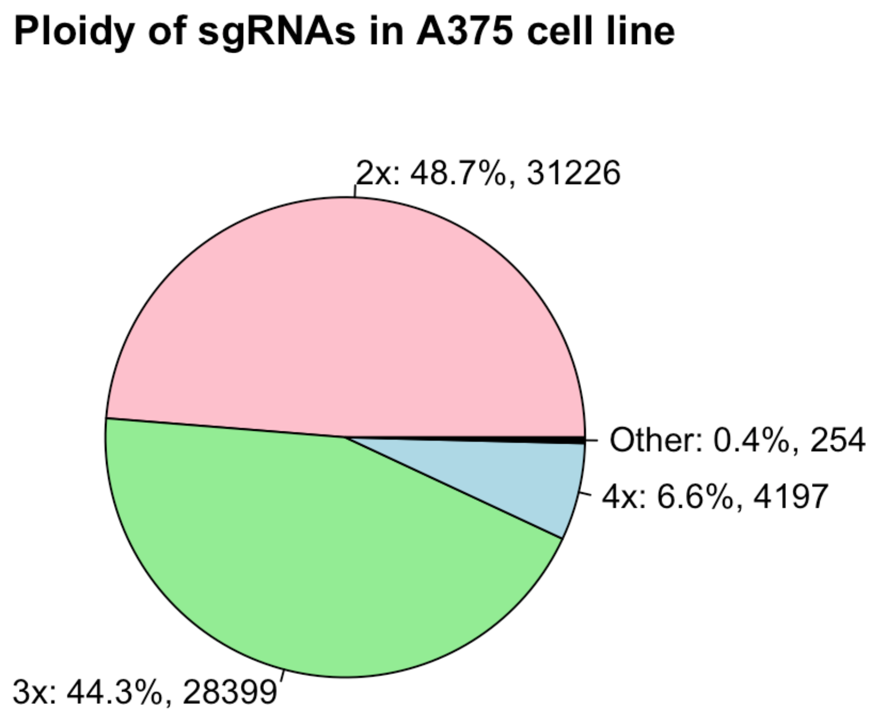
```
## This is vegan 2.3-4
```

```
CNVload = read.table("data/CNV_guide_table.tsv", header = TRUE, sep = "\t")

process_CNV_table = function(df){
  df$D7R1 = df$norm_count_D7_Rep1/df$norm_count_plasmid
  df$D7R2 = df$norm_count_D7_Rep2/df$norm_count_plasmid
  df$D14R1 = df$norm_count_D14_Rep1/df$norm_count_plasmid
  df$D14R2 = df$norm_count_D14_Rep2/df$norm_count_plasmid
  df$PLX7R1 = df$norm_count_PLX7_Rep1/df$norm_count_plasmid
  df$PLX7R2 = df$norm_count_PLX7_Rep2/df$norm_count_plasmid
  df$PLX14R1 = df$norm_count_PLX14_Rep1/df$norm_count_plasmid
  df$PLX14R2 = df$norm_count_PLX14_Rep2/df$norm_count_plasmid
  df$D7 = (df$norm_count_D7_Rep1+df$norm_count_D7_Rep2)/(2*df$norm_count_plasmid)
  df$D14 = (df$norm_count_D14_Rep1+df$norm_count_D14_Rep2)/(2*df$norm_count_plasmid)
  df$PLX7 = (df$norm_count_PLX7_Rep1+df$norm_count_PLX7_Rep2)/(2*df$norm_count_plasmid)
  df$PLX14 = (df$norm_count_PLX14_Rep1+df$norm_count_PLX14_Rep2)/(2*df$norm_count_plasmid)
  return(df)
}

CNVtable = process_CNV_table(CNVload)
OStable = read.table("data/Shalem_2014Table_S1.tsv", header = TRUE, sep = "\t")
```

Results

CNV probe distribution

The following is a pie chart describing the percentage of probes found in CNVs:

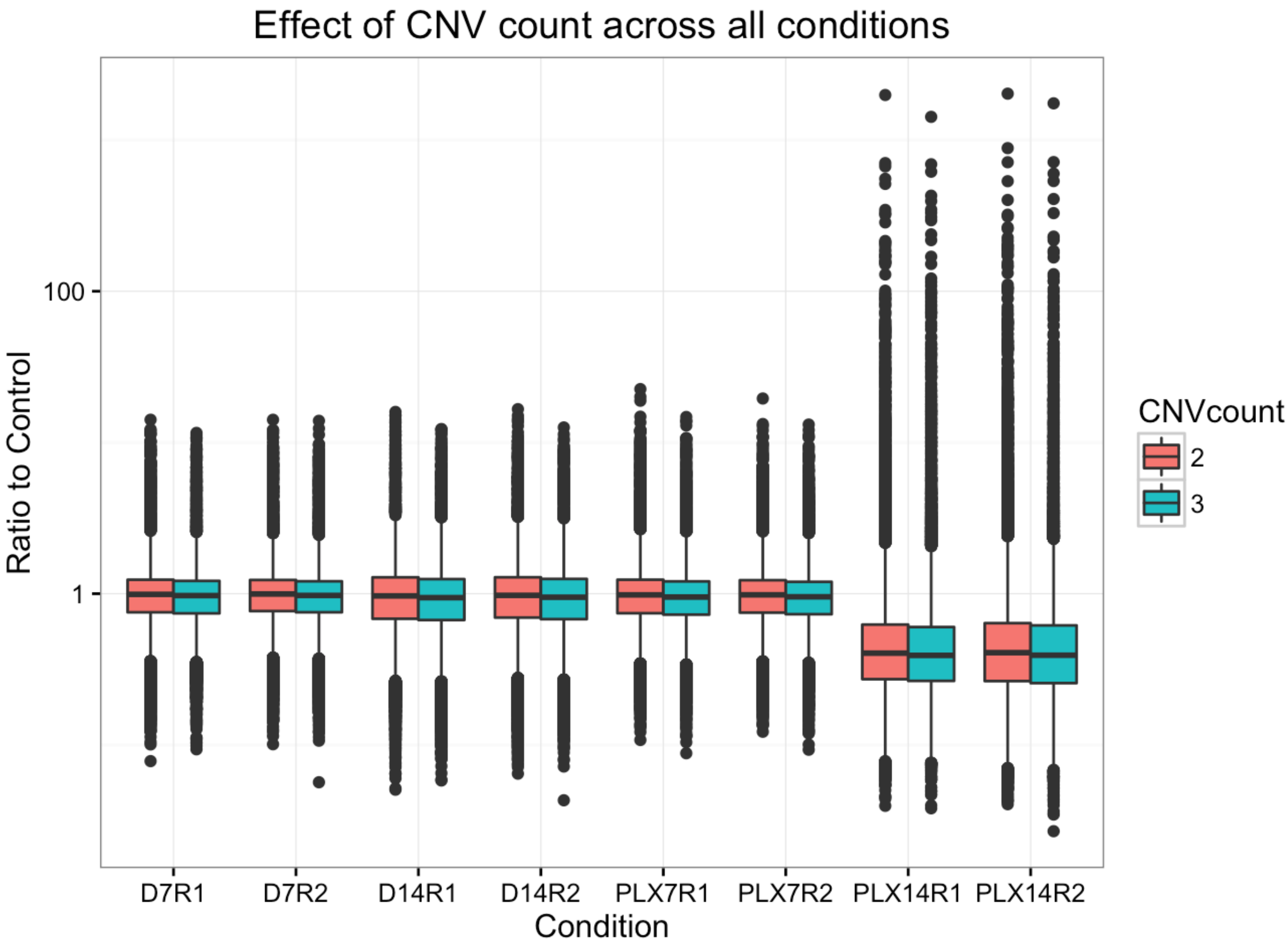


Since almost half of the target sites are within 2x genes and another 44% are 3x, I focused the remainder of my analyses on just those two ploidy states. While 4,197 target sites were in 4x copies of the genome, I felt that was still too few sites in relation to 2x and 3x and could skew my results.

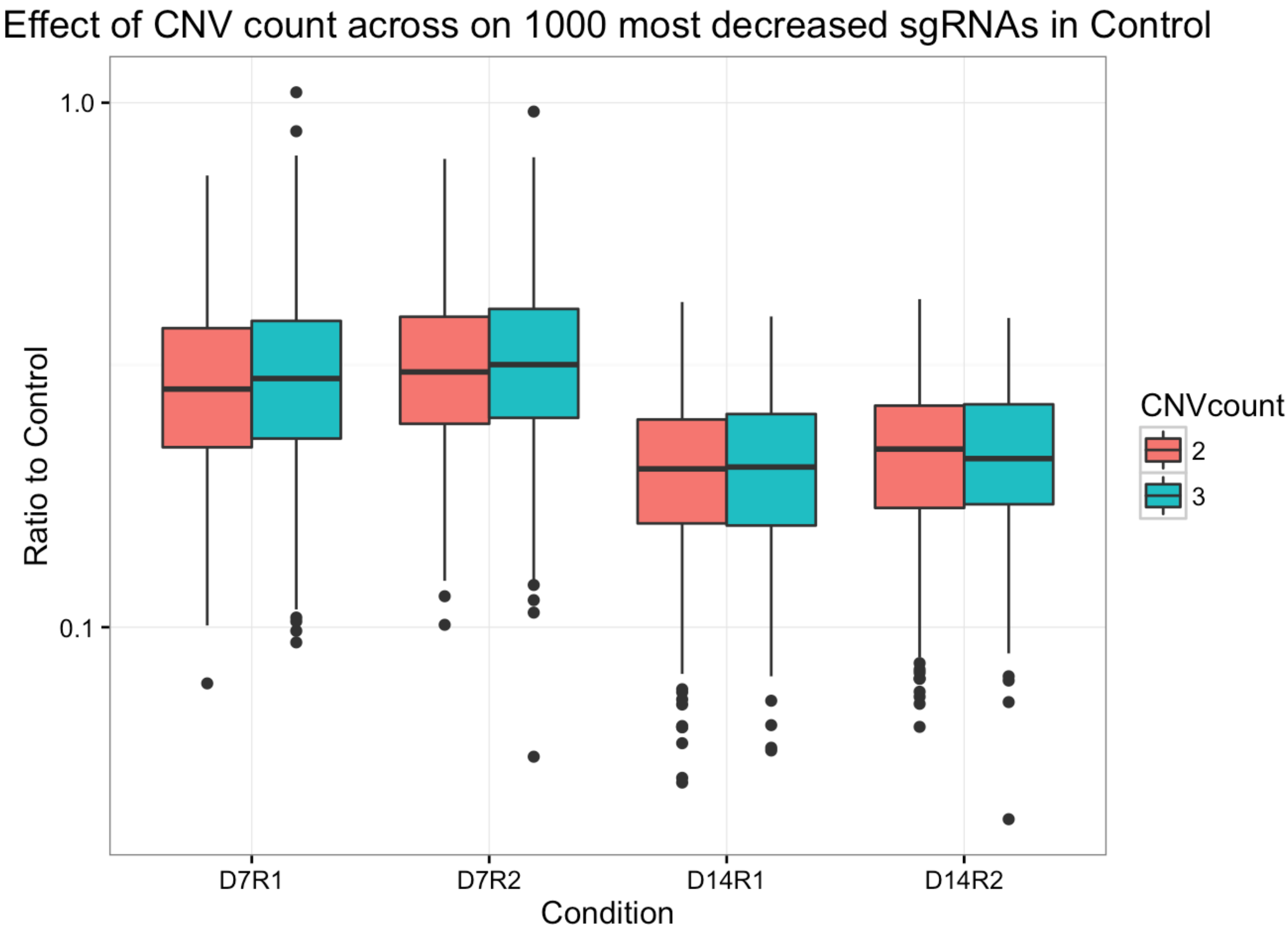
```
CNVtable_23x = subset(CNVtable, CNVcount > 1 & CNVcount < 4)
```

Effect of CNVs on CRISPR guide activity

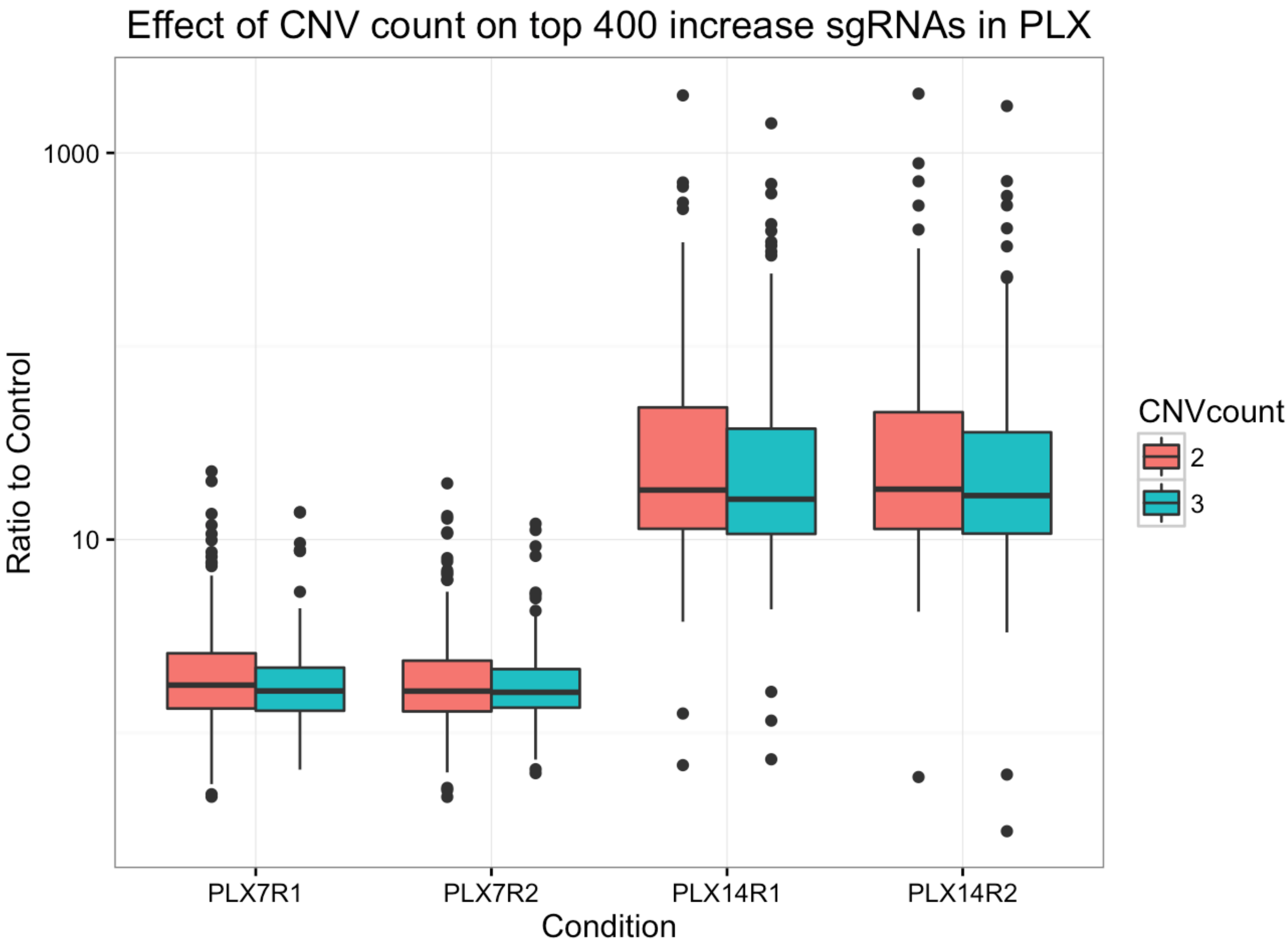
To get an idea of what the relationship looked like between CNVs on CRISPR guide activity, I created boxplots looking at the ratio of probe concentration change based on condition and CNV count. A hypothesis of the effect of CNV would be *“If there are more copies of a gene, it would take longer to see the effects in the GeCKO experiment”*. If this hypothesis were true, you would see 3x be closer to a ratio of 1 than 2x in the 7 day, but would become more similar to the 2x levels in day 14. However, this boxplot show no striking difference between 2x and 3x across all conditions.



While the previous plot did not show a significant difference between 2x and 3x, there is the possiblity that the majority of genes which are not affected by the sgRNAs could mask the few that are being affected. So, I took the top 1000 with the greatest decrease in Day 14 of control condition. These probes are significantly enriched for genes that are essential for survivability in cell culture. When this is done, it appears that could **possibly** be an effect of CNVs that fits with the previously described hypothesis.



We can also take the top 400 sgRNAs in the PLX experiment. While in day 7 the 3x sgRNAs are closer to a ratio of 1 than the 2x, this trend is maintained in day 14. So, it could be that those genes are have less of a selection effect on PLX treatment.



Statistical Analysis using MANOVA

While graphing the data can show us striking differences, a statistical test will determine if there are actually differences due to ploidy. Since there is a relatively large amount of data points and the data is not expected to be normally distributed, I chose to use a multivariate analysis of variance (MANOVA).

```
fit <- manova(cbind(D7R1,D7R2,D14R1,D14R2,PLX7R1,PLX7R2,PLX14R1,PLX14R2) ~ as.factor(CNVcoun
t), data = CNVtable_23x)
summary(fit, test="Pillai")
```

```
##                          Df    Pillai approx F num Df den Df    Pr(>F)
## as.factor(CNVcount)      1 0.0013976    10.43      8  59616 1.017e-14 ***
## Residuals            59623
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The MANOVA resulted in **Pr(>F) = 1.017e-14** which is extremely significant, indicating there is an effect of CNV on the system. I tried this multiple times with different sections of the data, and it always came out significant to some degree. I then tried randomizing the data to see if the significant is still there. Expectedly, when randomizing the data, the significance goes away:

```
flips=sample(0:1, length(CNVtable_23x[,1]), replace=T)
fit <- manova(cbind(D7R1,D7R2,D14R1,D14R2,PLX7R1,PLX7R2,PLX14R1,PLX14R2) ~ as.factor(flips),
data = CNVtable_23x)
summary(fit, test="Pillai")
```

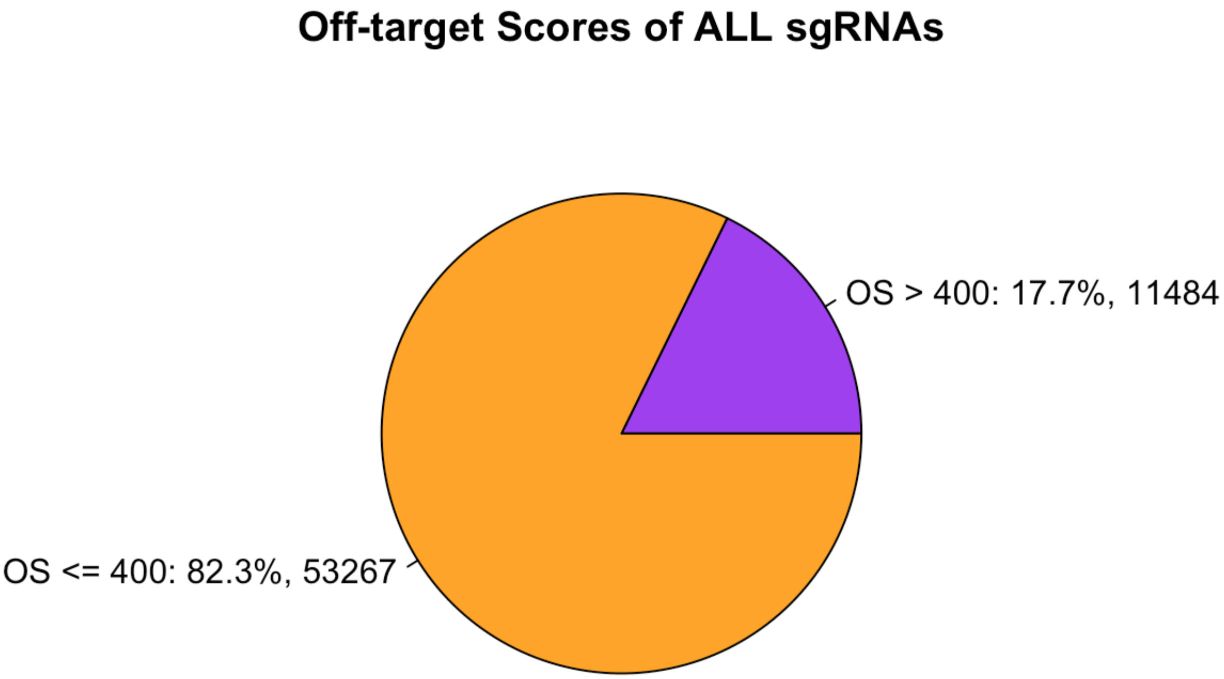
```
##                       Df     Pillai approx F num Df den Df Pr(>F)
## as.factor(flips)       1 0.00010951  0.81614      8  59616 0.5882
## Residuals          59623
```

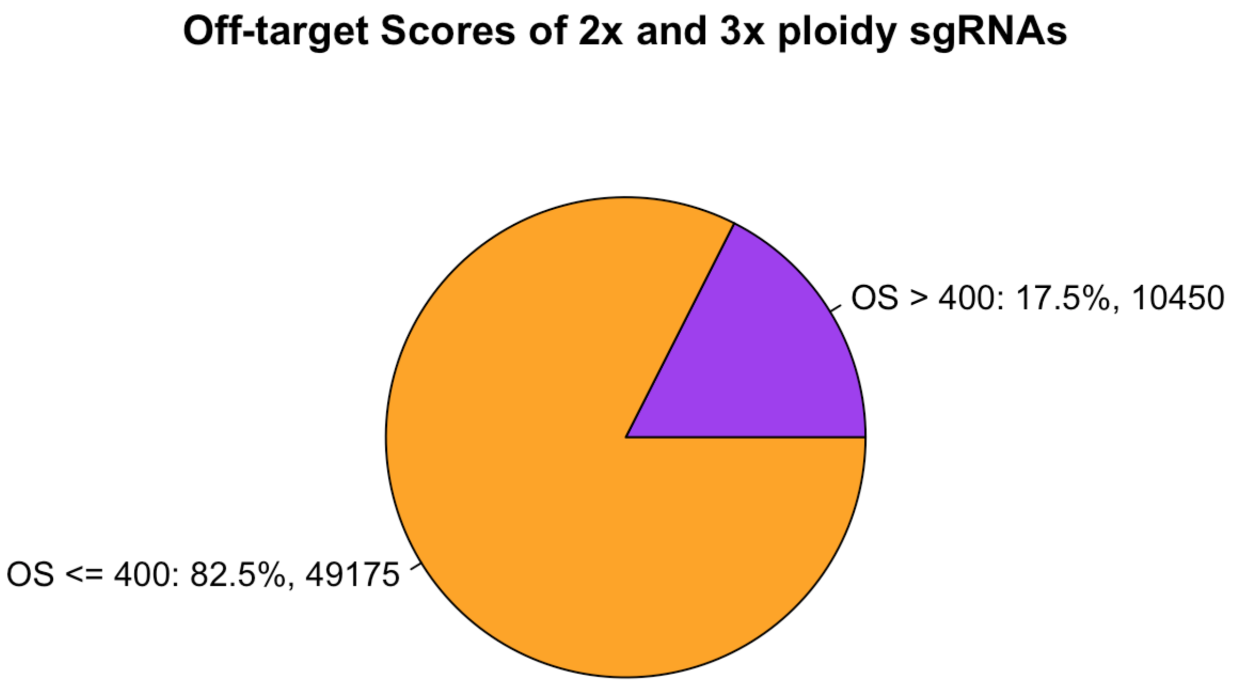
Off-target Scores

In researching this project, I went back to the original paper by Shalem et al. 2014. In their Supplemental Materials *GeCKO library design* section, they stated:

For each gene, the best (lowest OS) sgRNAs were chosen with the constraint that no sgRNAs have a OS>400. This resulted in a library of 64,751 unique sgRNAs targeting 18,080 coding genes with an average of 3-4 sgRNAs per gene.

However, when I downloaded Supplemental Table 1 from the paper, I found a significant amount of probes had an Off-target score greater than 400.




Off-target Scores of 2x and 3x ploidy sgRNAs

Rerun MANOVA taking Off-target scores into account

Knowing that a large portion of the sgRNAs have a high OS, I reran the MANOVA after removing those sgRNAs. Once filtering, We get an even more significant effect **Pr(>F) = 8.223e-16**.

```
CNVtable_23x_OSg400 = subset(CNVtable_23x, OS <= 400)
fit <- manova(cbind(D7R1,D7R2,D14R1,D14R2,PLX7R1,PLX7R2,PLX14R1,PLX14R2) ~ as.factor(CNVcount), data = CNVtable_23x_OSg400)
summary(fit, test="Pillai")
```

```
##                        Df   Pillai approx F num Df den Df    Pr(>F)
## as.factor(CNVcount)     1 0.001804   11.107      8  49166 8.223e-16 ***
## Residuals           49173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Furthermore, if we run MANOVA on those sgRNAs with OS > 400, there is no significance.

```
CNVtable_23x_OSg400 = subset(CNVtable_23x, OS > 400)
fit <- manova(cbind(D7R1,D7R2,D14R1,D14R2,PLX7R1,PLX7R2,PLX14R1,PLX14R2) ~ as.factor(CNVcoun
t), data = CNVtable_23x_OSg400)
summary(fit, test="Pillai")
```

```
##                         Df     Pillai approx F num Df den Df Pr(>F)
## as.factor(CNVcount)     1 0.00087461   1.1425      8  10441 0.3307
## Residuals           10448
```

Conclusion

While the graphs show a small amount of difference between ploidy and CRISPR guide activity, the MANOVA indicates there is a significant effect (p-value = 1.017e-14). You get an even further significant effect of you filter out the CRISPR guides with a high Off-target score.