Nabihah Momin, Alan Fletcher, and Kishan Bhakta
BME 335
5/4/2022

Written Report & R Code

Context:

CRISPR-Cas9 is a modern method of genome editing that is one of the more accurate, efficient, and cheaper ways of modifying DNA. This development is based on naturally occurring systems in which bacteria store a segment of viral DNA to disable viruses. The goal of this modification was to prevent and treat various human diseases and we wanted to test if the SpCas9-BME 335 variant was better than the standard wild type SpCas9 variant.

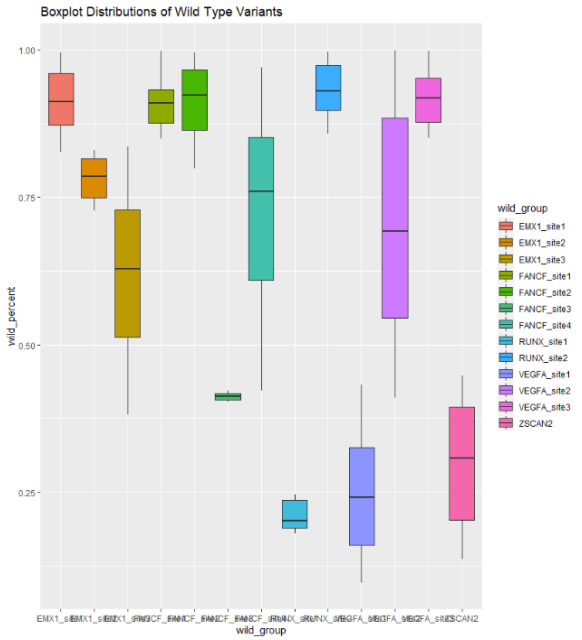
Overall Hypothesis:

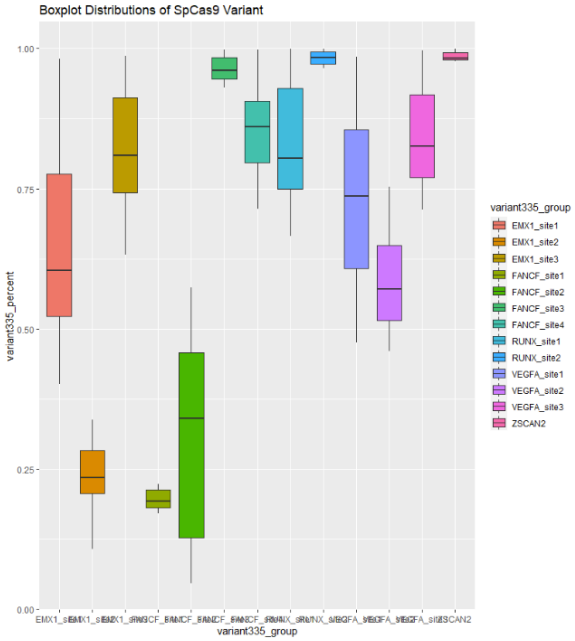
The overall null hypothesis for each of the three studies conducted was that the SpCas9-BME335 variant will not significantly improve specific on-target DNA contacts when compared to the wild type SpCas9 variant. The overall alternative hypothesis for each of the three studies conducted was that the SpCas9-BME335 variant will significantly improve specific on-target DNA contacts when compared to the wild type SpCas9 variant.

Study 1:

The purpose of Study 1 was to record the percent modified measurements of functionality with on-target sites for SpCas9-BME335 and wild-type SpCas9. Since we aren't trying to see if one variable influences or causes a change in another variable, this is going to be an observational study because we are only recording the measurements of each trial. The study population is all possible measurements for all possible site locations and the study sample is 35 and 33 percent modified measurements each at 13 different site locations for the original and wildtype variants, respectively.

The boxplot distribution for each site location for each variant appears to be normal and the central limit theorem is also upheld to perform our analysis given that in both instances, a sample size larger than 30 measurements were provided at each site location.





The explanatory variables are each of the 13 different site locations for both the original and wild type variants and the response variable is percent modified measurements for on-target sites for each nuclease. The null hypothesis is that the mean percent modified measurements of on-target sites for the SpCas9-BME335 and wild type variants are the same, respectively. The alternative hypothesis is that the mean percent modified measurements of on-target sites for the original and wildy type variants are significantly different, respectively. To conclude, we would reject the null hypothesis for both variants since our P value was less than our critical value of 0.05 meaning that the mean percent modified measurements of on-target sites for each separate variant are significantly different from each other.

SpCas9-BME335													
	EMX1_site1	EMX1_site2	EMX1_site3	FANCF_site1	FANCF_site2	FANCF_site3	FANCF_site4	RUNX1_site1	RUNX1_site2	ZSCAN2	VEGFA_site1	VEGFA_site2	VEGFA_site3
Mean	0.6514453	0.2349272	0.820571	0.1956846	0.3096781	0.9642124	0.8584348	0.8341721	0.9830805	0.9856472	0.7426902	0.5860548	0.8381035
Medain	0.6045826	0.2346527	0.809351	0.1928666	0.3409625	0.9610121	0.8605764	0.8042041	0.9841508	0.9828217	0.7372391	0.5710056	0.8252061
Standard Deviaiton	0.1634585	0.06381431	0.1083748	0.01698834	0.167979	0.02168579	0.08210258	0.1095251	0.0117479	0.006941237	0.1632339	0.08980979	0.09121664

```
> SpCas9AOV <- aov(variant335_group ~ variant335_percent, data = my_data_variant335)
> summary(SpCas9AOV)
                   Df Sum Sq Mean Sq F value Pr(>F)
variant335_percent   1   1229  1229.0   108.3 <2e-16 ***
Residuals          453   5141    11.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

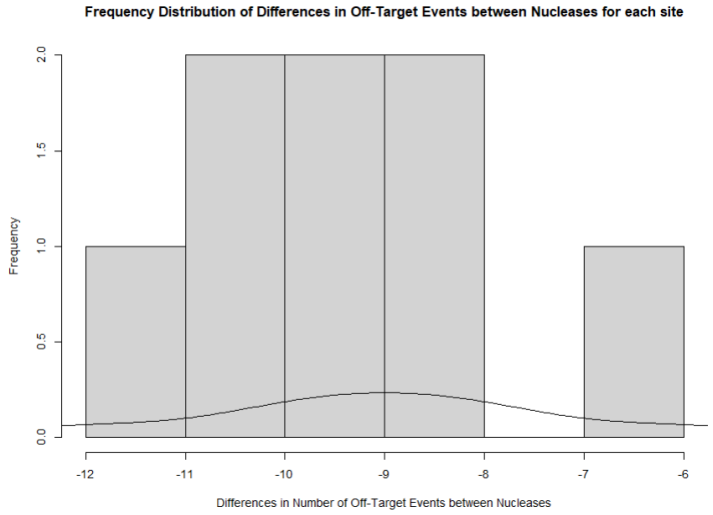
Wild Type Variant													
	EMX1_site1	EMX1_site2	EMX1_site3	FANCF_site1	FANCF_site2	FANCF_site3	FANCF_site4	RUNX1_site1	RUNX1_site2	ZSCAN2	VEGFA_site1	VEGFA_site2	VEGFA_site3
Mean	0.916467	0.7806065	0.6247764	0.90807	0.9105898	0.4128735	0.728144	0.2118074	0.9322326	0.3041003	0.2469917	0.7111821	0.9191446
Medain	0.9120843	0.7848081	0.629081	0.9092427	0.9227942	0.4132032	0.7601295	0.2017236	0.9295625	0.3084445	0.2415919	0.692281	0.9183595
Standard Deviaiton	0.04854839	0.03210051	0.1316697	0.03893801	0.05963183	0.006192917	0.1667144	0.02415465	0.04178635	0.1031114	0.1100564	0.1945408	0.9183595

```
> SpCas9AOV_WildType <- aov(wild_group ~ wild_percent, data = my_data_wild)
> summary(SpCas9AOV_WildType)
             Df Sum Sq Mean Sq F value   Pr(>F)
wild_percent   1    466   466.2   35.93 4.35e-09 ***
Residuals    427   5540    13.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Study 2:

The purpose of Study 2 is to test if the SpCas9 BME 335 variant significantly reduces the number of off-target events in cells at each site when compared to the wild-type SpCas9 variant. This study is classified as an experimental study since the treatments, represented by the type of nuclease (SpCas9 BME 335 and the wild type SpCas9), were randomly assigned to each of the eight sgRNA sites. The explanatory variables for this study are the type of nuclease (SpCas9-BME335 variant or the wild type SpCas9 variant) and eight sgRNA targeted sites (EMX1 Site1, EMX1 Site 2, FANCF Site 1,FANCF Site 2, FANCF Site 3, FANCF Site 4, RUNX1 Site1, and ZSCAN2). The response variable for this study is the number of off-target events. The study population consisted of the eight different sgRNAs targeted sites in the endogenous human genes. The study sample consisted of the number of off-target events that were observed for these sgRNAs at each site for both the SpCas9-BME335 variant and the wild type SpCas9 variant.

After examining the data and understanding what the data represents, my group first decided to conduct a paired-t-test. A paired t-test was first chosen since both types of nucleases were assigned to each of the eight sgRNA targeted sites. Therefore, this means that for each sgRNA site there was a pair of the number of off-target events for both nucleases. In order to test the assumptions of a paired-t-test, my group decided to create a histogram.



As seen by the histogram above, the x axis of the histogram represents differences in the number of off-target events which was calculated by taking the difference between the number of off-target events recorded for the SpCas9-BME335 variant and the wild type SpCas9 variant. The y axis of the histogram represents the frequency at which each of the differences in the number of off-target events occurred.

From the histogram above we can see that the data does not have a normal distribution and is skewed. Due to this conclusion, my group decided to conduct a non-parametric sign test since the assumption of normality needed for a paired t-test was not met. The null hypothesis for our sign test was that the median value of the difference between the SpCas9-BME335 variant and the wild type SpCas9 variant is equal to zero while our alternative hypothesis states that the difference between the SpCas9-BME335 variant and the wild type SpCas9 variant is not equal to zero. After conducting a sign test in R, my group found that the p value for the sign test was equal to 0.007812 which leads us to reject the null hypothesis.

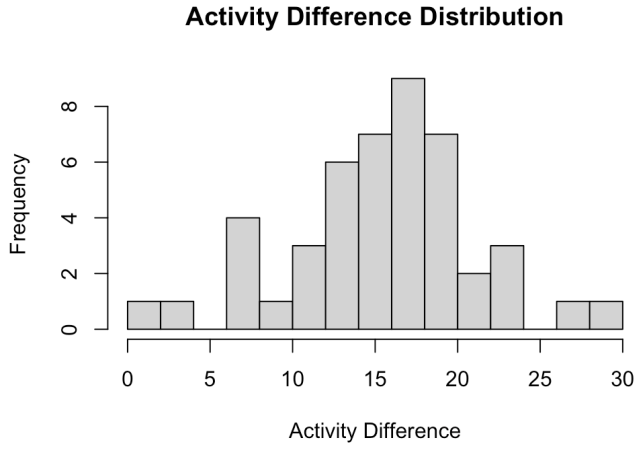
Descriptive Statistics for Study 2

	SpCas9-BME335	Wild Type SpCas9
Mean	1.25	10.25
Median	1.50	10.50
Standard Deviation	0.89	2.38
Variance	0.79	5.64

This rejection of the null hypothesis was further supported by the fact that the median value for the SpCas9-BME335 variant is significantly lower than the median value for the wild type SpCas9 variant as seen in the table above. Therefore, we safely reject the null hypothesis and can conclude that the SpCas9-BME335 variant significantly reduces the number of off-target events in cells at each site.

Study 4:

Study 4 compares the off target activity between a newly modified SpCas9 BME335-2 variant and the standard wild type SpCas9 protein. The given data was the raw difference between the off target activity of the SpCas9 BME335-2 variant and the wild type SpCas9 variant across the 45 sgRNA sites that they affected. We interpreted this raw difference as: (SpCas BME335-2) - (Wild type SpCas9). In this experimental study, the population was the sites altered by the either protein being compared and the sample was the off target activity observed for both proteins. The explanatory variables are the nuclease type and sgRNA site; and our response variable is the off target activity at each site.



Descriptive Statistics

diff_mean	15.2839239446674
diff_standard_de…	5.54633102024879
median	15.8701608714982
n	46
SE	0.817761895108068

Preliminary Visualization: Histogram (generated in R) of the given data to determine the normality of the distribution

Based on our initial inspection of the data, it is roughly gaussian. However, given that our mean and median are equal, we can cite the central limit theorem to bolster our assumption of normality. Because the data is comparing the off target activity of each nuclease type at the same sgRNA sites, we felt it appropriate to perform a paired test on the data. Our null hypothesis for this test was that the mean of the off target activity of the SpCas9 BME335 variant would be the same as the mean of the off target activity of the wild type SpCas9 protein. Our alternative hypothesis is that there is a significant difference between the means of the off target activity difference of the SpCas9 BME 335-2 and wild type SpCas9 nucleases.

After running the test in R, the data yielded a test statistic of t = 16.884 and a p-value of 2.2e-16. Because the p-value was less than 0.05, we can reject the null hypothesis and conclude that the means of the off target activity of the two groups were significantly different. Since the means were significantly different, and the differences between the activity at each site was positive, we can further conclude that the off target activity of the SpCas9 BME335-2 variant was greater than the off target activity of the wild type SpCas9 protein.

Overall Conclusion:

Based on the conclusions drawn from each of the studies, we can draw a conclusion about the overall performance of the SpCas9 BME335 variant. In Study 1, we reject our null hypothesis that the within group means are the same, meaning that the within group means at each site is different. If we wanted to draw a full conclusion about which nuclease has a higher mean and determine whether or not the BME335 variant improves upon the wild type at each site, we would have to perform a Tukey-Kramer test to determine which off-target activity means are greater at each site. If the BME335 variant has more sites with less off-target activity than the wild type, we can support our main alternative hypothesis. In Study 2, we reject our null hypothesis which means that the median value of the difference between the SpCas9-BME335 and the wild type SpCas9 variant is not equal to zero. From this, we can conclude that the BME335 variant did improve upon the wild type SpCas9 protein at the targeted sites because the BME335 variant lowered the off target activity. In Study 4, we rejected our null hypothesis which means that the means of the off target activity between the two proteins was significantly different. We further concluded that the off target activity of the BME335-2 variant was greater than the wild type SpCas9 protein, which would not be an improvement upon the wild type SpCas9 nuclease because our goal is to reduce off target activity. Because of all of these factors, our test is inconclusive and will remain so until a Tukey-Kramer test is performed.