Salisbury
UNIVERSITY

Honor College at Salisbury University

Honors Thesis

An Honors Thesis Titled

Copy Number Alternations on Human Chromosome 21 Associated with Prostate Cancer.

Submitted in partial fulfillment of the requirements for the Honors Designation to the

Honors College

of

Salisbury University

in the Major Department of

Biological Science

by

Fiona Halloran

Date and Place of Oral Presentation: SUSRC, April 2018

Signatures of Honors Thesis Committee

Mentor:	Dr. Philip D. Anderson	Dr. Philip Anderson
Reader 1:		Dr. James Buss
Reader 2:		
Dean:		Dr. James Buss
	Signature	Print

Abstract

The goal of my thesis project was incorporate my skills of computational biology, genomics, computer science and mathematics into something that would contribute to the knowledge of the scientific community and public in a useful way. In pairing with Dr. Philip Anderson, I was able to accomplish this goal. The research I pursued was focused on identifying potential regions of importance in prostate tumorigenesis. I analyzed publically available from The Cancer Genome Atlas using the programs R, Perl, and Bash. The code I wrote mined the data of the patients' chromosome 21 DNA, specifically copy number variants. The output of interest was the sums of patients who had a deletion or insertion at every nucleotide. I plotted this to find regions on chromosome 21 where insertion or deletion was prevalent. The region of greatest deletion was in an intron of a gene, and I used this information to hypothesize possible explanations for the regions' involvement in prostate cancer growth. My findings will hopefully allow other scientists to better understand the mechanisms of prostate cancer and inspire them to continue what I just began to uncover.

Introduction

Prostate cancer is the second most common malignancy in men, with few treatment options in advanced cases. Previous research has shown that the behavior of the cancer is highly unpredictable ranging from rapidly spreading and fatal to slowly developing forms that are treatable with little therapy. Currently, there are very few reliable predictors of which outcome a patient will face. Genomics research is a rising field being used to uncover previously unknown predictors and causes of cancer. The goal of our research project is to identify potentially important regions in chromosome 21 involved in prostate tumorigenesis, or tumor growth. To do this, we are looking at copy number variants (CNV) in chromosome 21 of 477 prostate cancer patients. A CNV is any non-diploid region of a human chromosome, constituting amplifications (>2 copies) or deletions (<2 copies). We used publicly available data from The Cancer Genome Atlas (TCGA), a data set of tumor related genomics and data for normal and tumor tissue in cancer patients. We then analyzed the data to find areas of the genome that were amplified or deleted. This is significant because if a region is important to tumorigenesis then it would have recurring amplification or deletions across several patients, whereas unimportant regions to the tumor growth would not be expected to be amplified or deleted more than chance alone would predict. Areas of frequent deletion would suggest the region plays a role as a tumor suppressor, whereas regions of frequent amplification would suggest a role as an oncogene. For this analysis, we are assuming that tumor suppressors and oncogenes are found in the same regions on the human genome among individuals.

1 Loading TCGA Data

First we will need to load the TCGA CNV data into our workspace. Then we will look at how many different cancers are in the data set. Since we are specifically looking at prostate cancer patients, we will need to figure out exactly how many patients we have. Hereafter, all analyses will be done with prostate cancer patients only.

```
load("/home/hallorf/TCGA_CNV_Flat.Rdata")
with(TCGA_CNV_Flat,unique(Chromosome)) #Cancer Types in dataset

##  [1] 1  2  3  6  4  5  11 21 22 X  10 7  8  9  12 13 14 15 16 17 18 19 20
## Levels: 1 10 11 12 13 14 15 16 17 18 19 2 20 21 22 3 4 5 6 7 8 9 X

#Preparing a dataset of prostate cancer only.
PRAD<-subset(TCGA_CNV_Flat, subset = (CancerType== "PRAD"))
```

2 Quality Control

As we looked through TCGA_CNV_Flat, we saw a relationship between the number of probes and segment mean. As the segment mean increased, the number of probes decreased. This relationship can be seen in the plot below.

```
plot(Num_Probes~abs(Segment_Mean),
     type="n",
     xlim= c(0,3),
     ylim= c(0,5000),
     main = "Segment Mean vs. Number of Probes",
     bty="n",
    data=PRAD)
LM<-lm(Num_Probes~abs(Segment_Mean),data=PRAD)
abline(LM)
```

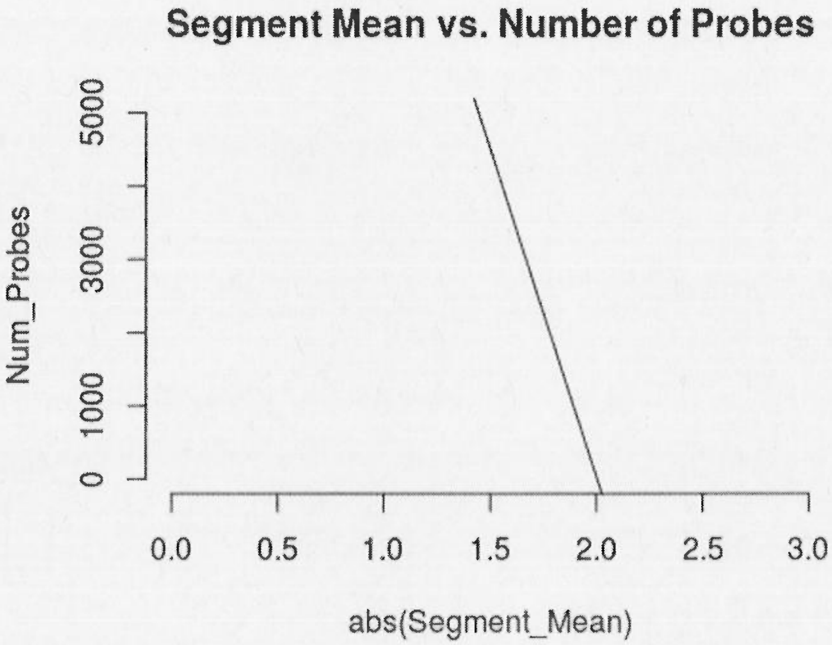

Figure 1. Due to this inverse relationship, we will be analyzing only the samples that have at least 20 probes, and have a segment mean that is either less than or equal to -0.2 (deletion), or greater than or equal to 0.2 (addition). These limits are in concordance with the literature by Laddha et al.

The code below will subset the data for the previously stated parameters.

```
TCGA_CNV_Flat <- subset(TCGA_CNV_Flat,
                        subset=((Num_Probes>=20 &(Segment_Mean<=(- .2)|Segmen
t_Mean>= .2))))
PRAD<-subset(TCGA_CNV_Flat, subset = (CancerType== "PRAD"))
length(with(PRAD,unique(ID)))  #Number of Patients
```

```
## [1] 477
```

3 Creating Patient Deletion Vectors

Here we survey for regions of amplification or deletion on chromosome 21. We will look at each nucleotide position individually and see if that nucleotide is inserted or deleted in each patient. We will then create two matrices of all of the nucleotides and whether they are inserted (score=1) or not (score=0) and deleted (score=1) or not (score=0) in each patient. Then we will add up all of the numbers of each row to see how many times a single nucleotide was amplified or deleted in all patients. Finally, we will graphically display this to see all of the nucleotides and identify regions of frequent amplification or deletion, which may be used as bio markers for prostate cancer.

We will write a loop that finds all nucleotides that were deleted in prostate cancer patients on one chromosome, starting with chromosome 21. We need to add a column in PRAD that give each patient a number, so they can be identified by that instead of by the ID. This will also allow us to cycle through the patients in the while loop using their number.

```
length(with(PRAD,(as.factor(ID))))
```

```
## [1] 26327
```

```
PRAD["Patient"] <-c(with(PRAD,(as.factor(ID))))
```

This code should make 477 vectors each of length for 48129895, one position for each nucleotide of chromosome 21.

```
del.PRADchr21 <- subset(PRAD, subset = ((Chromosome== 21) & Segment_Mean<=-0.
2))
i<-1
while (i<=477) {
  deletions.p<-rep(x=0, times=48129895)
    y <- 1
      while (y<=nrow(del.PRADchr21))        {
deletions.p[as.integer(del.PRADchr21[y,"Start"]):
              as.integer(del.PRADchr21[y,"End"])]<-1
    y <- y+1
      }
        write.table(deletions.p, file=paste(i, "del.Chrm21.txt"
                                            ,sep = "_"),
                    row.names = FALSE, col.names = F)
        i <- i+1
}
```

4 Making the Matrix of Deletions

Now that we have .txt files for each nucleotide and whether it was deleted in each patient, we need to write command that assembles a matrix of nucleotides (rows) and patients (columns).

```
paste *.txt > del.Chrm21.txt
```

5 Adding up Nucleotide Deletion Counts

Then we will write Perl code that will take the array of all 0 and 1s of Chromosome 21 deletions and add up each row. It will capture the sums in a file called del.Chrm21.Counts.txt.

```
use strict;


my $sum = 0;


while( my $line = <STDIN> ) {
        my $sum = 0;
    my @vals = split( /\t+/, $line );
    chomp(@vals);
        for ( @vals ) {
              $sum += $_;
                      }
        print $sum."\n";
                                }
exit;
```

```
more del.Chrm21.txt | perl ~/pasted/CountLines.pl > del.Chrm21.Counts.txt
```

6 Repeating the Process for Insertions

This code is to make a table for chromosome 21 insertions for all PRAD patients.

```
add.PRADchr21 <- subset(PRAD, subset = ((Chromosome== 21) & Segment_Mean>=0.2
))
i<-1
while (i<=477) {
 bypatient<- subset(add.PRADchr21,subset = (Patient == i))
 add.p<-as.numeric(logical(length=48129895L))
  y <- 1
   while (y<=nrow(bypatient))    {
add.p[as.integer(bypatient[y,"Start"]):
    as.integer(bypatient[y,"End"])]<-1
  y <- y+1
   }
   write.table(add.p, file=paste(i, "add.Chrm21.txt",sep = "_"),
         row.names = FALSE, col.names = F)
    i <- i+1
 }
```

Pasting them together into a file called add.Chrm21.txt.

```
paste *.txt > add.Chrm21.txt
```

Adding up the insertions for each nucleotide into a file called add.Chrm21.Counts.txt.

```
more add.Chrm21.txt | perl ~/pasted/CountLines.pl >add.Chrm21.Counts.txt
```

7 Displaying Insertion and Deletion Regions

We now have two .txt files that contains one number for each nucleotide in the chromosome, where the number in that spot is the total number of patients that had a deletion at that nucleotide and the same for insertions. Here we graphically display the number of deletions or insertions that occured at each chromosome for all patients. This code converts the txt files into vectors in our workspace which we can graph. It then makes three plots, of deletions only, insertions only and both on the chromosome. For the graph with both insertions and deletions, we replaced counts of zero on the insertion line with twos, so the line could be seen along with the deletions.

```
del.Chrm21.Counts<-read.table("del.Chrm21.Counts.txt", header = FALSE)
del21Vector<-as.numeric(del.Chrm21.Counts[,1])

add.Chrm21.Counts<-read.table("add.Chrm21.Counts.txt", header = FALSE)
add21Vector<-as.numeric(add.Chrm21.Counts[,1])
Edittedadd21Vector<-replace(add21Vector, add21Vector==0,values = 2)

#deletions
png(filename = "del21graph.png")
plot(del21Vector,col="blue",
     type="o",
     bty="n",
     axes=F,
     xlab="",
     ylab="Number of Patients")
axis(side = 2)
dev.off()

#insertions
png(filename = "add21graph.png")
plot(add21Vector,
     col="red",
     type="o",
     bty="n",
     axes=F,
     xlab="",
     ylab="Number of Patients")
axis(side = 2)
dev.off()

#insertions and deletions

png(filename = "Chr21graph.png")
plot(Edittedadd21Vector,
     col="red",
     type="o",
     ylim=c(-135,17),
     bty="n",
    axes=F, xlab="",
   ylab="Number of Patients")
axis(side = 2)
lines(negdel21, col="blue", type = "o")
dev.off()
```

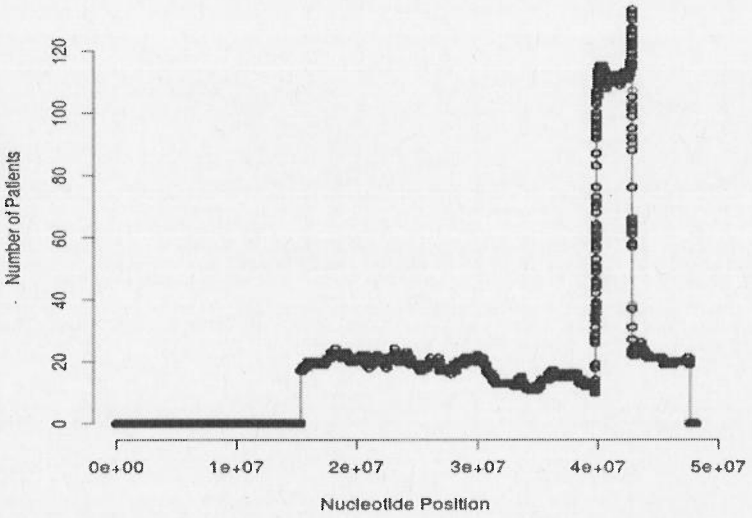

Figure 2. Deletions on Chromosome 21 by nucleotide on all patient data

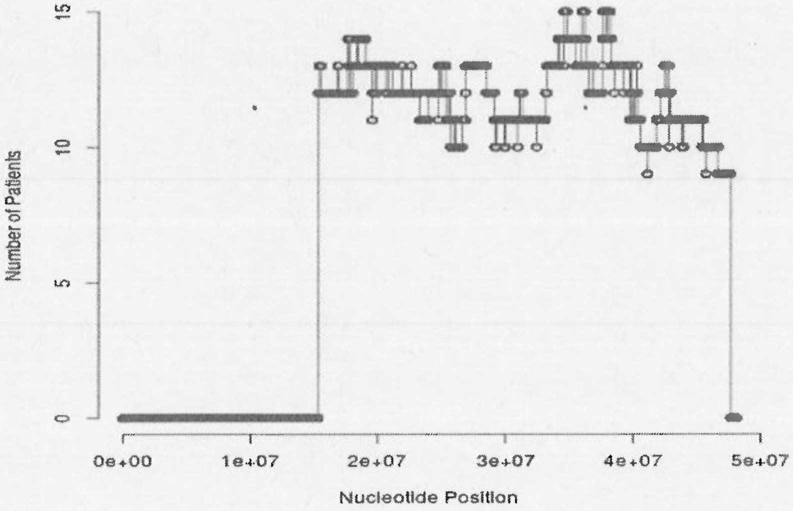

Figure 3. Insertions on Chromosome 21 by nucleotide on all patient data

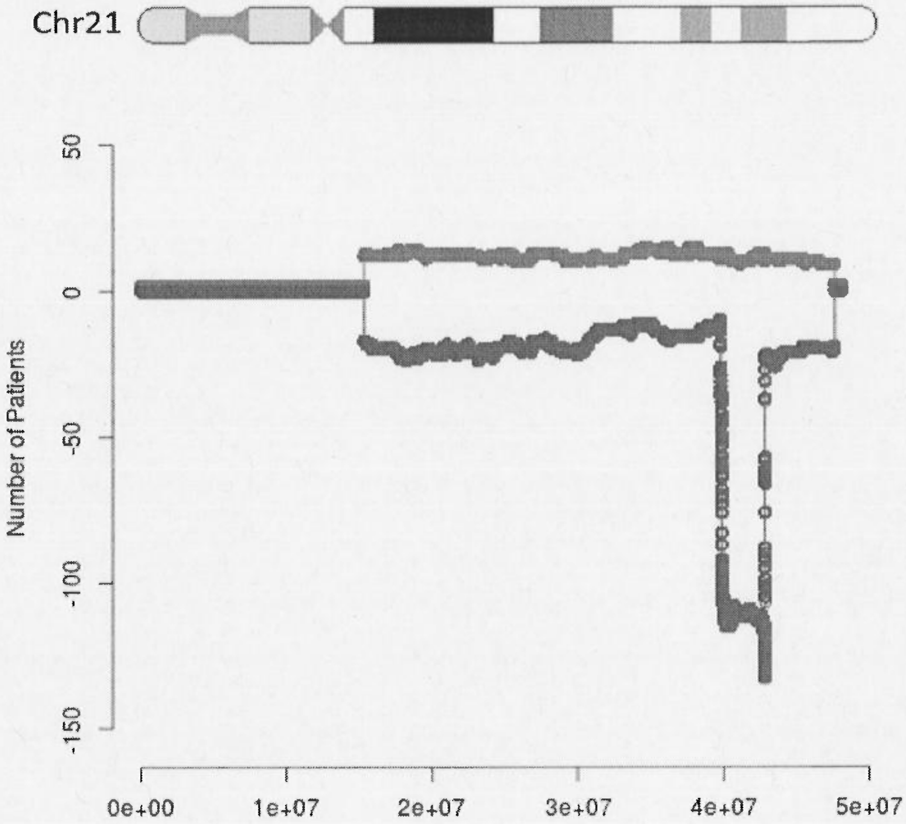

Figure 4. The results of our analysis show conservation among the first 15798741 nucleotides, where they are not deleted or inserted in any patient. We can also identify a dramatic increase in patients with deletions at approximately 42,000,000 nucleotides.

8 Identifying the Content of High Deletion Regions

Thirty-five nucleotides from position 42857981 to 42858015 were all deleted in 133 patients. We found that these nucleotides were found in intron 5 of the gene TMPRSS2 on chromosome 21. This gene is know to be related to prostate cancer, but the specificity of the mechanism involved in tumorigenesis to the nucleotide scale has not yet been identified.

Since the region of interest are found in an intron and therefore non-coding, we are going to look for transcription factor motifs among this high deletion region. This will allow us to identify possible reasons for such high rates of deletion among prostate cancer patients.

The code below queries a motif database for binding site matches.

```
require(TFBSTools)

require(Biostrings)

suppressMessages(library(JASPAR2016))

opts<-list()

opts[["type"]]<-"SELEX"

opts[["tax_group"]]<-"vertebrates"

AllVerts<-getMatrixSet(x=JASPAR2016,opts=opts)

names(AllVerts)

VertPWM<-toPWM(AllVerts)

MyDNA<-DNAString("GGCTTGTGGTTTCTACAGAAAACTCCAGGAGGTTA")

MySearch<-searchSeq(VertPWM,subject=MyDNA,strand="*")

SearchResults<-writeGFF3(MySearch)

SearchResults<-SearchResults[,-(1:2)]

SearchResults<-SearchResults[,-6]

SearchResults<- cbind(SearchResults,transcriptionfactor)
```

The query gave us 5 motifs that could be present in this 35 base pair sequence. Each motif matched a section of our sequence with at least 80% accuracy.

-Basic helix-loop-helix factors

-Fork head / winged helix factors

-Tryptophan cluster factors

-Rel homology region (RHR) factors

-C2H2 zinc finger factors

The transcription binding sites with the highest scores, or match to a fragment in our DNA sequence were a Rel homology region (RHR) factor and a C2H2 zinc finger factor. Now we will look up the actual DNA sequence that matched those two transcription factor binding sites and compare them to the sequence of those transcription factor binding sites. The code below identifies the reverse complement of both regions of our input sequence.

```
#Rel homology region (RHR)

reverseComplement(MyDNA[18:27])

seq: TGGAGTTTTC


#C2H2 zinc finger

reverseComplement(MyDNA[6:11])

seq: AACCAC
```

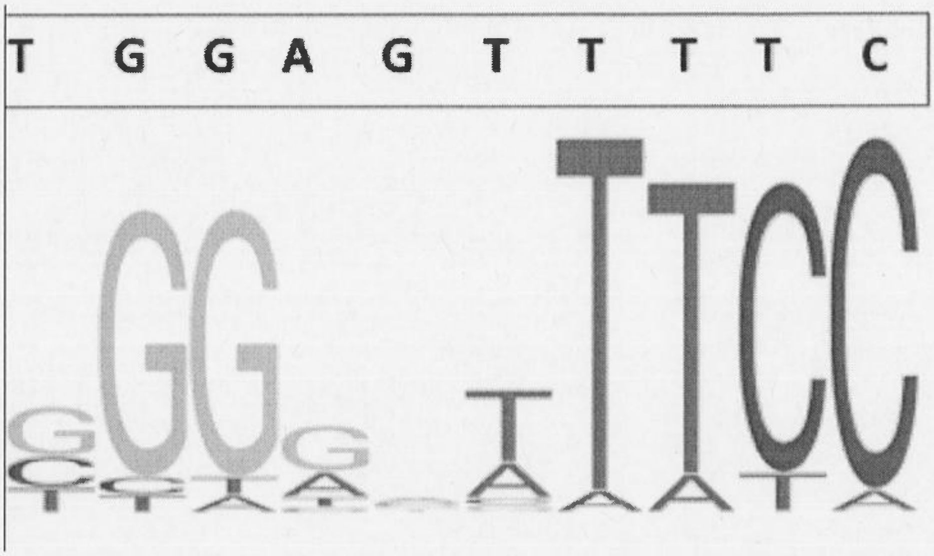

Figure 5. Transcription factor biding site Rel homology region (RHR) factor sequence compared to genome sequence.

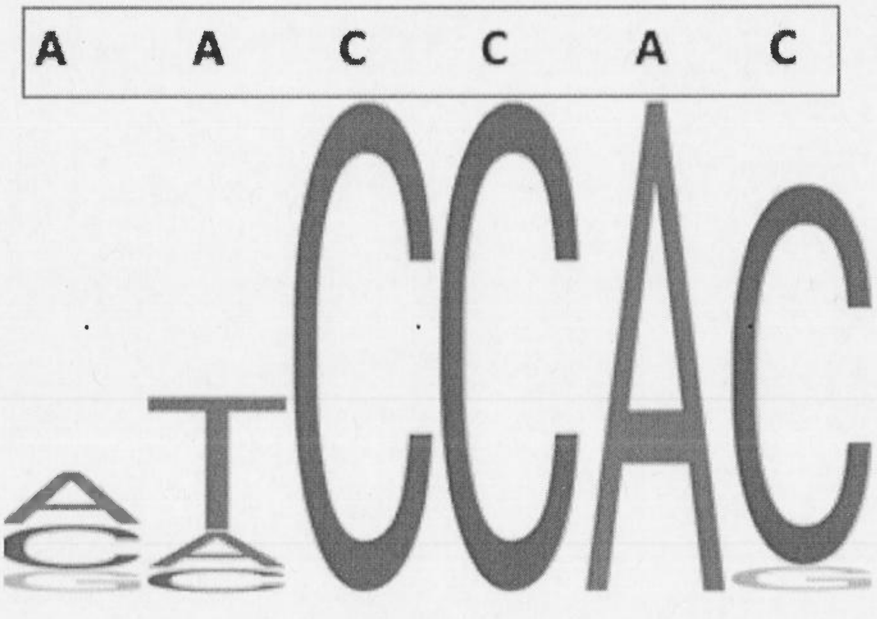

Figure 6. Transcription factor biding site C2H2 zinc finger factor sequence compared to genome sequence.

Summary

The CNVs on chromosome 21 appear to be nonrandomly distributed. The analysis showed areas of fluctuating frequency of insertion along the chromosome and an increase in deletion around the 42 million nucleotide region. We identified that the greatest number of deletions occurred in a region inside of intron 5 of the TMPRSS2 gene on chromosome 21. Though this gene has been previously shown to be involved in prostate tumorigenesis, the exact nucleotide location of deletion we found has not yet been reported. We then looked for possible motifs to explain the abnormal increase in CNV deletion in this particular intron sequence. Our search led us to five different types of transcription binding sites that may be in this sequence. Two of them were further investigated and compared by nucleotide to our sequence. The presence of transcription factor binding sites may explain the large number of patients have that region deleted, but it cannot be confirmed without further research. Future research should be done to test whether or not the region of greatest deletion is important in tumorigenesis This could be done using in vitro studies of cells that were mutated using CRIPR-Cas9 to delete this region. The effects of the deletion could confirm the sequence's role. This code should also be run on the entire genome of all patients in order to identify other regions of abnormally high copy number variation. The results could be used to identify bio markers and behavior of prostate cancer tumors.

References:

1. Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer.Cell.2015;163:1011–1025

2. "The Cancer Genome Atlas Home Page." National Institutes of Health, U.S. Department of Health and Human Services, cancergenome.nih.gov/.

3. R Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

4. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D.The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996-1006.

5. Laddha, Saurabh V et al. "Mutational Landscape of the Essential Autophagy Gene BECN1 in Human Cancers." Molecular cancer research: MCR 12.4 (2014): 485–490. PMC. Web. 8 Apr. 2018.

6. Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2018; 46:D260–D266, doi: 10.1093/nar/gkx1126

7. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996-1006.