Heuristics In Genome Danger Identification

William Li

Phoebe Au

AP Statistics

Ms. Sinatra

June 12, 2024

Abstract

Molecular genetics is the scientific discipline that studies the various aspects of genes in living organisms. Part of molecular genetics is to understand the effects of random events occurring in the DNA to the well being of the organism. Throughout the course of a human life, trillions upon trillions of cells within the body divide. Sometimes random events could happen during division which cause variations in the cell's DNA. These variations can aid an organism to better adapt to their environment, or lead to debilitating diseases that prevent an individual from living their lives to the fullest. In this paper, we specifically investigate what heuristics we could use to help separate those dangerous gene mutations from benign ones. We have found that the danger level of a mutation, as defined by various scores such CADD, LoFtool, and SIFT, is heavily related to the impact and consequences of that mutation and correlates with the position of the mutation within a certain subset of the data. On the last point, more research is needed to determine the meaning of that subset.

Introduction

Key Words:

- ➔ DNA: Genetic information used in the process of protein synthesis within living organisms
- ➔ Genome: The entire set of genetic information of an living organism
- ➔ Gene: Sections of a DNA strand that contributes to the formation of certain proteins. The created protein manifests certain traits of the human body or contribute to various bodily functions, such as metabolism.

- ➔ Alleles: The different variations of a gene is called an allele, it is what gives variation to each individual's DNA. For example, a person that possesses the allele, RR, can have brown eyes, while another person with the allele, rr, can have blue eyes.

The data catalogues different types of alleles within humans. There are a variety of alleles within the human body, this is what gives each individual their own unique bodily characteristics, and allows individuals to inherit diseases from previous generations. The process of DNA replication for gametes, sex cells, is called meiosis. This process is responsible for shuffling the genes of the parents' sex cells and passing it on to their young. Under regular circumstances, through genetic shuffling, gene variation is created. However sometimes genes can mutate into new types of allele. This in turn gives rise to more gene variation and is the focus of this data set. Sometimes allele mutations are safe, as they produce the correct protein as instructed despite the mutation. However, certain allele mutations are dangerous, as they produce a defective protein that does not function as intended causing harm to the body. To measure how dangerous these mutations are, the data set also shows various scores (e.g. SIFT score, PolyPhen score) that determine whether a variation (allele combination) is safe or not. This is done automatically by genone-analyzing algorithms. With safe variations being denoted as benign or tolerated and dangerous combinations are denoted by deleterious, damaging, or pathogenic.

Hypothesis

If statistical analysis is used to compare the different variables in gene expression, then it can predict whether or not a mutation is dangerous or not as defined by various scoring systems.

Data

Source

Genetic Variant Classifications, Link

Columns (renamed):

A. Chromosome
   a. There are 22 chromosomes in the human genome . This column shows which chromosome the variant is located on.
B. Position
   a. The position on the chromone that the allele variation is located on, based on a metric system called Locus.
C. Reference Allele
   a. A reference allele from a normalized reference genome.
D. Alternate Allele
   a. An alternate allele that was observed.

E. Allele frequencies (GO-ESP)
   a. The frequency of allele taken from GO-ESP
F. Allele frequencies (ExAC)
   a. The frequency of allele taken from ExAC
G. Allele frequencies (1000 genomes project)
   a. The frequency of allele taken from 1000 genomes project
H. Disease database name and identifier
   a. Tag-value pairs of disease database name and identifier, e.g. OMIM:NNNNNN
I. ClinVar's preferred disease name
   a. ClinVar's preferred disease name for the concept specified by disease identifiers
J. HGVS expression
   a. Top-level (primary assembly, alt, or patch) HGVS expression.
K. Variant Type
   a. Whether the variance is caused by a chromosomal mutation or variance due to meiosis or single nucleotide mutation
      i. For chromosomal mutation, they are denoted by the type of mutation:
         1. Deletion
         2. Duplication
         3. Inversion
         4. Reciprocal Translocation
      ii. For the rest, they are denoted by: single_nucleotide_variant
L. Variant sources
   a. The variant's clinical sources reported as tag-value pairs of database and variant identifier
M. Sequence Ontology ID|molecular_consequence
   a. Sequence Ontology is a form of vocabulary for genome annotation.
   b. Molecular consequence describes the different types of results in single nucleotide mutation
N. Allele Origin
   a. Cause of mutation that causes variation in alleles, number coding as follows:
      i. 0 - unknown;
      ii. 1 - germline;
      iii. 2 - somatic;
      iv. 4 - inherited;
      v. 8 - paternal;
      vi. 16 - maternal;
      vii. 32 - de-novo;
      viii. 64 - biparental;
      ix. 128 - uniparental;
      x. 256 - not-tested;
      xi. 512 - tested-inconclusive;
      xii. 1073741824 - other

O. Has conflicting submissions
    a. The original purpose of this dataset was to classify this variable.

P. Allele
    a. The allele observed.

Q. Consequence
    a. Describes which type of mutation has occurred.

R. IMPACT
    a. Descriptive words that describe how impactful the mutation was to the overall chromosome. How much does this mutation affect the function of the chromosome? These words includes:
        i. LOW
        ii. MODERATE
        iii. HIGH
        iv. MODIFIER
            1. Modifies the original function of chromosome

S. SYMBOL
    a. A name for the gene/allele.

T. Feature_type
    a. Type of feature. Currently one of Transcript, RegulatoryFeature, MotifFeature.

U. Feature (Ensembl stable ID)
    a. An ID for a gene.

V. BIOTYPE
    a. A form of gene classification, all alleles from the data set are denoted protein_coding.
        i. Protein coding means that this portion of the chromosome is used to transcribe into mRNA in order to produce proteins.
        ii. mRNA is akin to an instruction manual in the process of creating new proteins.
        iii. Transcription is the process of mRNA synthesis.

W. EXON
    a. Exon are the parts of the mRNA that are used for translation, the process of protein synthesis.
    b. Within the data set column, EXON, is denoted by a fraction, the numerator represents the exon portions and the denominator represents the entire mRNA strand.

X. cDNA Position
    a. Relative position of the complementary DNA (cDNA) strand.
        i. During transcription, the process for mRNA synthesis, the DNA double strand is split into 2. The strand directly used for the synthesis of mRNA is called the complementary strand. The other unused strand is called the coding DNA strand.

Y. CDS position
    a. Relative position of the coding DNA strand (CDS).

Z. Amino acid location in protein

a. Amino acids are the primary building blocks of protein. They are first built into chains before being turned into a mature protein through various steps of folding.

AA. Amino acids

a. The type of amino acid produced by this part of the allele, specifically a specific codon.

BB. Codons

a. Codons are the small units that create an allele or gene. Codons are defined by groups of 3 nucleic bases, the primary building blocks of DNA. All different sequences of codons correspond to 1 type of amino acid.

b. In the data set, Codons are denoted by a fraction-like form. The "numerator" represents the alternative codon observed from the mutated allele, and the "denominator" represents the reference codon from the reference allele.

CC. STRAND

a. Defines the direction of the DNA, as DNA strands are in complementary pairs in order to create the helix structure.

b. +1 signifies the "forward" direction, which is 5' to 3'.

i. The number 5' and 3' are based on the numbering system for carbon molecules within a carbon chain from Organic Chemistry.

c. -1 signifies the "reverse" direction, which is 3' to 5'.

DD. SIFT score

a. Method of determining how dangerous the mutation is to the function of the allele. .

EE. PolyPhen score

a. Method of determining how dangerous the mutation is to the function of the allele.

FF. LoFtool score

a. Method of determining how dangerous the mutation is to the function of the allele.

GG. CADD score (Phred-scaled)

a. Method of determining how dangerous the mutation is to the function of the allele.

HH. Raw CADD score

a. Method of determining how dangerous the mutation is to the function of the allele.

II. BLOSUM62 score

a. A score that deals with the alignment of amino acids in protein.

b. This is useful in evolutionary biology, as it can highlight change in the gene as natural selection takes place.

c. Not useful for protecting danger in mutations.

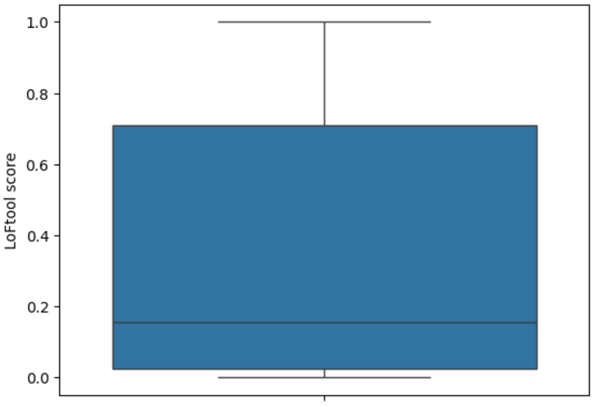
Analysis

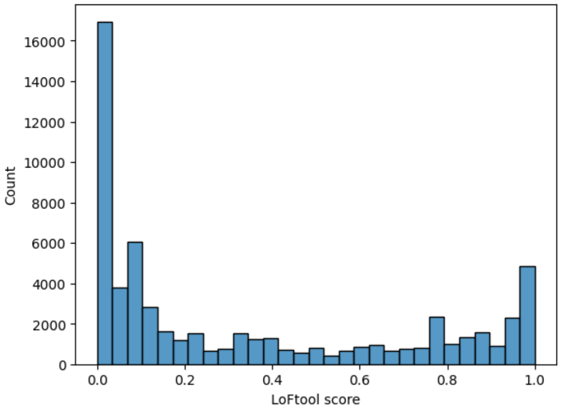
Github link for code, math, and graphs: Link

Single variable analysis

The first analyses to be done are on the dependent variables, the scores. Due to differing algorithms and methods, the scores have a different distribution. Some mutations marked as dangerous by one scoring method may not be true for another. Although, the general trend for quantitative variables seems to be: the higher the value of the numeric score, the more dangerous the mutation is to the overall function of the human body. While the categorical scores generally label dangerous mutations via descriptive words such as “deleterious” or “damaging”, therefore they are relatively easy to spot.

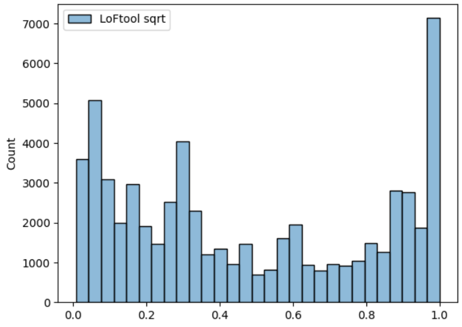
LoFtool Scores



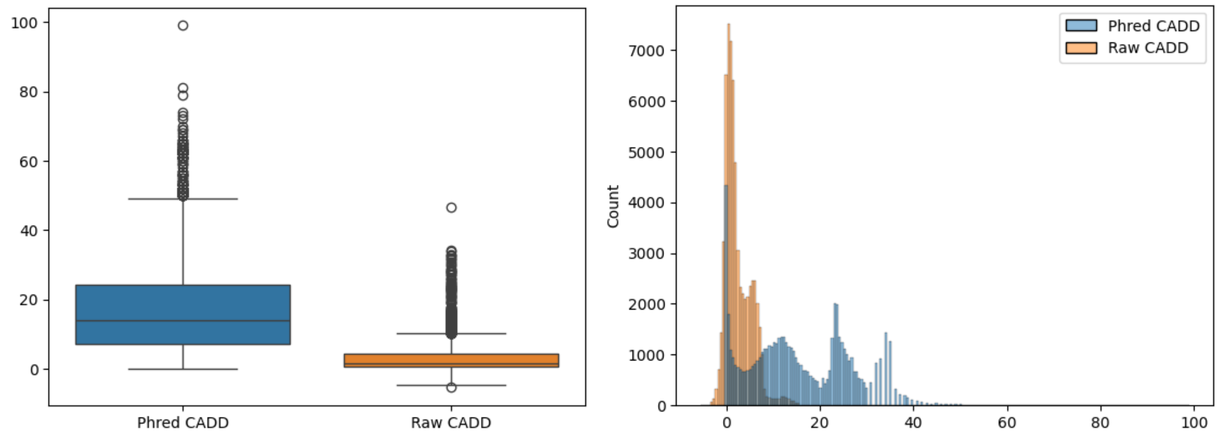


```
count    60975.000000
mean         0.345058
std          0.361238
min          0.000069
25%          0.024300
50%          0.157000
75%          0.710000
max          1.000000
```

For the Loftool scores, the score distribution is bimodal and relatively right skewed. This is true because the mean, 0.346 is far to the right relative to the median, 0.157. In terms of spread, the data ranges strictly from approach zero to one, the standard deviation is approximately 0.361, and the interquartile range is approximately 0.686. Since a majority of the data can be found within 1.4 standard deviations, the data is relatively close to each other, which is expected because of the right skewed bimodal. To expand on the graph seen and since a large number of values are situated around 0.0, another graph was made to better present the data in a more compressed form using a function. A square root function was used to present a less extreme result:



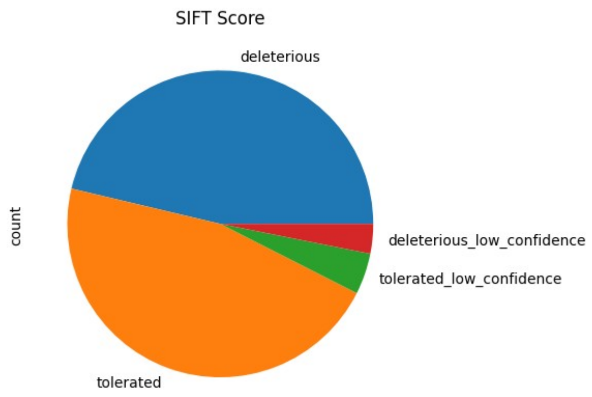
CADD scores

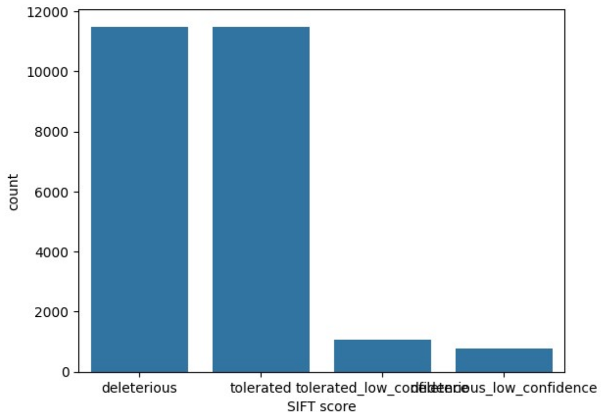


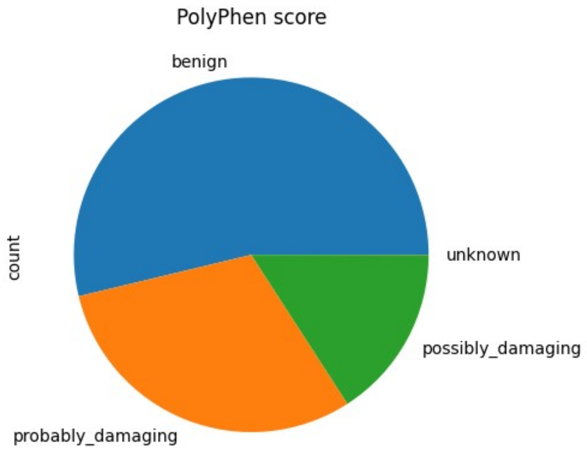
	Phred CADD	Raw CADD
count	64096.000000	64096.000000
mean	15.685616	2.554131
std	10.836350	2.961553
min	0.001000	-5.477391
25%	7.141000	0.462951
50%	14.090000	1.642948
75%	24.100000	4.381392
max	99.000000	46.556261

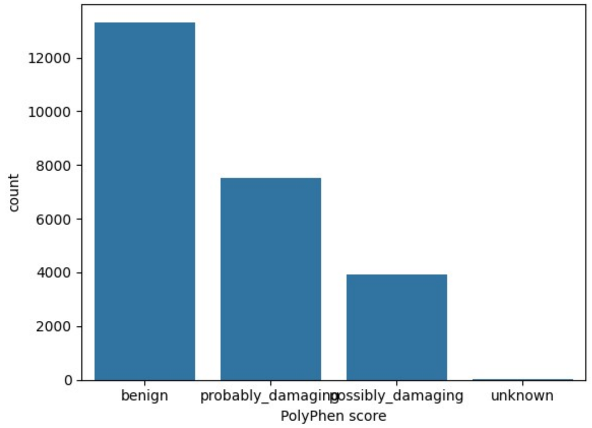
From the original dataset, it contains the CADD scores in both its raw form and with the Pred scale, a transformed log scale, applied. In its raw form, the distribution appears as bimodal and skewed strongly to the right, while the Phred distribution looks to be tetramodal with only a slight right skew. Since the Phred scaled scores are more spread apart, it has a standard deviation of 10.8 rather than one closer to that of 2.96. An observation of note is that the Phred scaled distribution ranges almost exactly from 0 to 100, which usually would imply that this was the unscaled distribution. A further investigation is necessary to determine the cause of this observation.

SIFT and PolyPhen scores









There is not much to be analyzed numerically, especially when compared to the quantitative variables, for these graphs, because both SIFT and PolyPhen scores are categorical. However, the SIFT score seems to be more decisive with its labeling, as there are less mutations labeled as “low confidence” in comparison to the PolyPhen’s “possibly”. In addition, the SIFT score seems to just label more mutations as dangerous than the PolyPhen score.

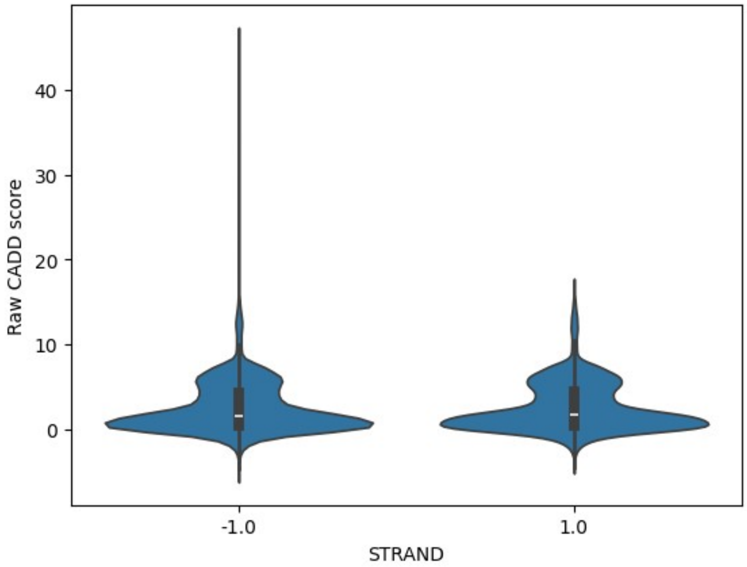
Two variable hypothesis tests

CADD vs Strand

Although not mandatory, hypothesis testing better supports the hypothesis. Therefore, the first test conducted was to measure whether the strand sign, representing which of the 2 DNA strands the mutation is on, had any effect on the danger of a mutation. Using a t-test, the difference in mean between the scores of the strand types was determined to be statistically significant. Alpha value is defined to be 5% as is usually done.

```
T-test Result: (statistic = -2.4368447524183057, p value =
0.014818745327453307, df = 63986.061936506856)
```

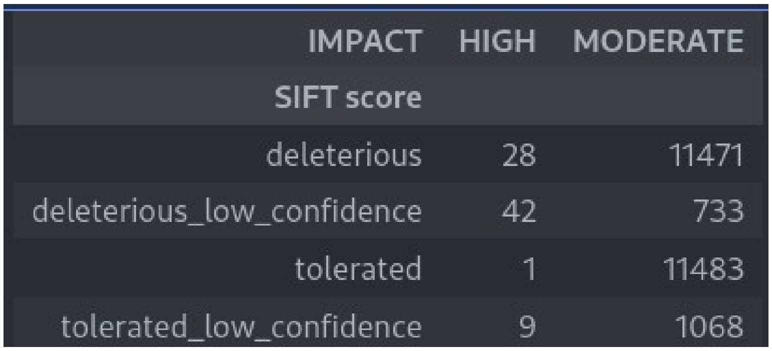
The means turned out to be around 2.44 standard deviations apart. Since the P-value, 1.48% was significantly less than the alpha, 5%, the null hypothesis could be safely rejected. That is until observing the distribution of data:



As seen, the distributions are essentially identical and the difference in mean is most likely just due to extreme outliers. Due to this, it can be reasonably concluded that this was not statistically significant. This is to be expected, because from a biological perspective, RNA transcription is not direction sensitive. Meaning regardless whether a mutation appears on the 5’-3’ strand or the 3’-5’ strand, transcription would still be carried out in the same way, and therefore the effects of the mutation would still be felt by gene expression in the long run.

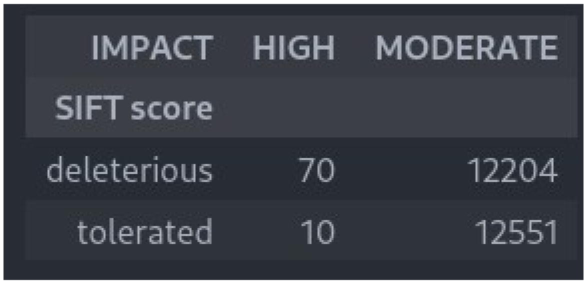
SIFT score vs Impact

Due to both, SIFT score and Impact being categorical variables, originally a chi-square test would be done. However, when the value counts were tallied up, it was observed that there was only one mutation with high impact that’s tolerated by the SIFT algorithm:



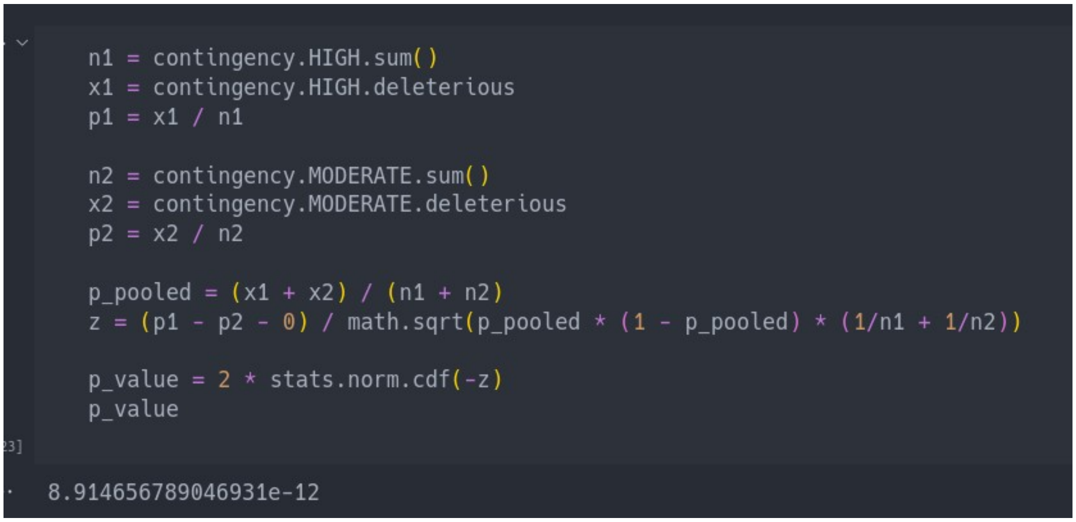
IMPACT	HIGH	MODERATE
SIFT score		
deleterious	28	11471
deleterious_low_confidence	42	733
tolerated	1	11483
tolerated_low_confidence	9	1068

Since a chi-square test requires a minimum count of 5 for each cell, a decision was made to merge the low confidence rows and do a z-test instead:



IMPACT	HIGH	MODERATE
SIFT score		
deleterious	70	12204
tolerated	10	12551

A new line of code was written, because of a lack of function provided by the python libraries to do a proportion z-test:

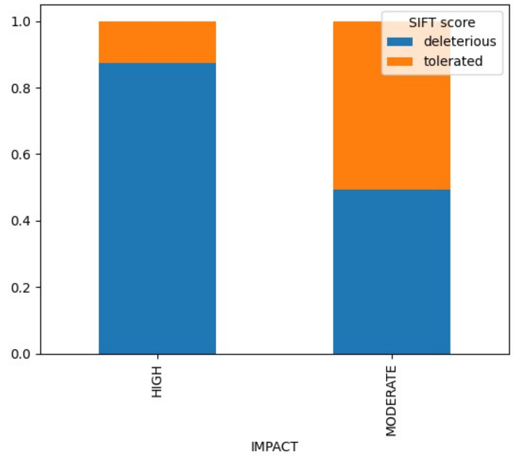


```
n1 = contingency.HIGH.sum()
x1 = contingency.HIGH.deleterious
p1 = x1 / n1

n2 = contingency.MODERATE.sum()
x2 = contingency.MODERATE.deleterious
p2 = x2 / n2

p_pooled = (x1 + x2) / (n1 + n2)
z = (p1 - p2 - 0) / math.sqrt(p_pooled * (1 - p_pooled) * (1/n1 + 1/n2))

p_value = 2 * stats.norm.cdf(-z)
p_value
```

```
8.914656789046931e-12
```

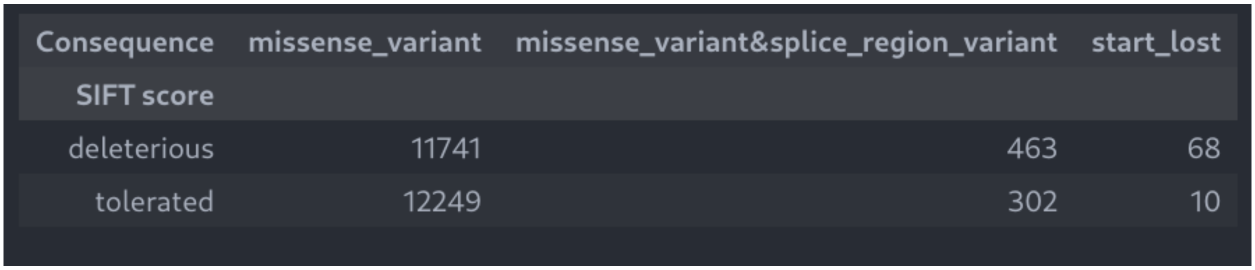
The p-value was 8.91×10^{-12}, demonstrating that there is an abysmal chance that this was due to pure randomness. Since the p-value was infinitesimally smaller than the chosen alpha value of 5%, this result is deemed statistically significant. Impact being a variable that affects the result of the scores, specifically SIFT score here, is expected. Impact in the data refers to the amount of change a chromosome has undergone due to mutation relative to a reference chromosome. If a chromosome has a high impact, it means the mutation has greatly affected the overall structure of the chromosome. If the structure of the chromosome is greatly affected then transcription and translation would likely not function as intended. This causes unexpected events to occur during these processes of gene expression, which in turn produce a dangerous result, relating back to the scores.

The bar graph is used to illustrate the difference in proportion. The x-axis represents frequency of SIFT Score given Impact. A mosaic plot may have been more appropriate given the difference between the sample sizes of the two impact types, but that would have been difficult to do within python.



SIFT score vs Consequence

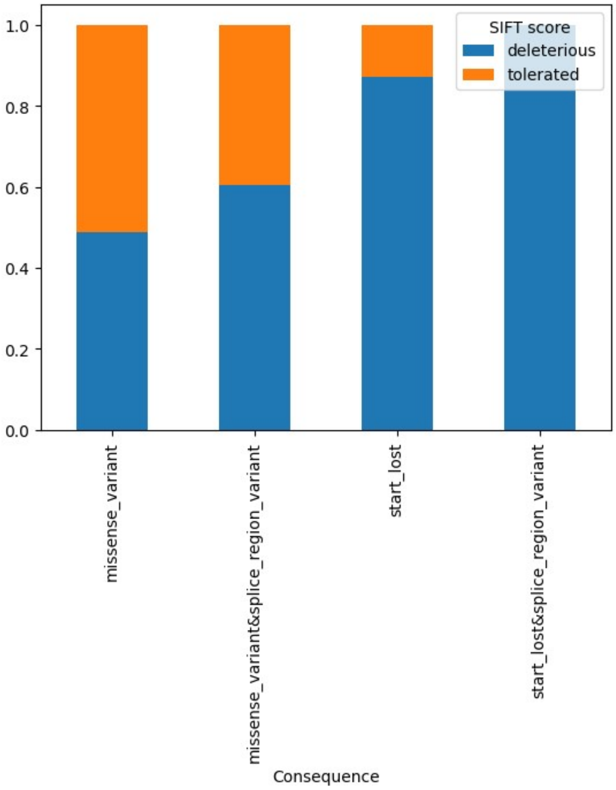
A chi-square test is decided to be done for the SIFT score and the consequence, in order to test whether these categorical variables are related. Consequence in the data refers to what type of mutation has occurred. Seen below is the result:



Consequence	missense_variant	missense_variant&splice_region_variant	start_lost
SIFT score			
deleterious	11741	463	68
tolerated	12249	302	10

```
Chi-2-Contingency Result: (statistic = 84.41714038470077, p value =
4.667140514138873e-19, dof = 2, expected freq = array
([[11855.40530745, 378.04856441, 38.54612814], [12134.59469255,
386.95143559, 39.45387186]]))
```

The p-value obtained from the chi-square test is approximately 4.66×10^{-19}, meaning it is extremely unlikely for the relationship to be based on random chance. Since the p-value obtained is smaller than the established alpha value of 5% by a wide margin, therefore the results are seen as statistically significant. It is reasonable in a biological perspective that the consequence had a correlation with the scores, SIFT score in this case. Mutations are not created equally, different mutations have different impacts on gene expression. For example, a missense mutation can cause an amino acid to be replaced in a protein, causing the protein to be defective and gene expression to be altered as a result. Since different mutations cause different forms of changes, these changes range in severity which in turn alters how dangerous a mutation can be. This connects back to how consequence is related to the scores, SIFT score in this case.



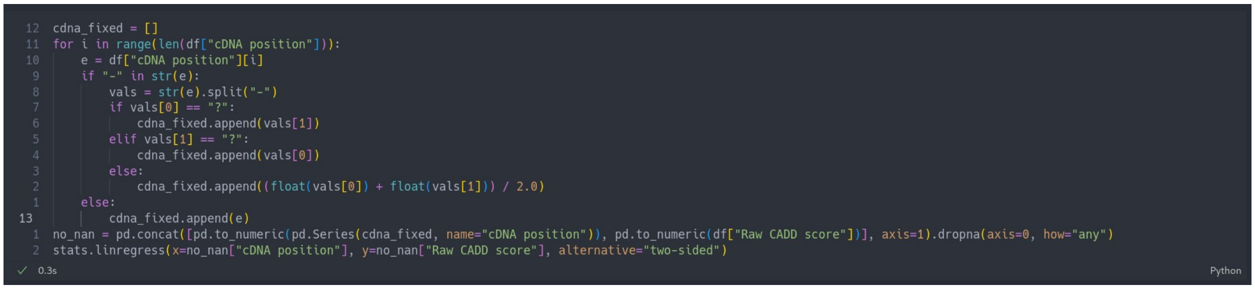
The bar graph to the side illustrates the difference in proportion, with the x-axis representing frequency of SIFT Score given consequence:

Note, since only one mutation had the ”start_lost&splice_region_variant”, this category was excluded from the chi-square test. However, it is still present in the bar chart because removing it would be too troublesome.

Regression Analysis

CADD vs cDNA

For the two variable analysis, it is decided to investigate if there was a correlation between cDNA position and CADD score. A problem arises during the processing of the data, the issue is that the positions are a range of values rather than a single position. In order to build a regression model which tries to predict CADD from cDNA and to display the data here, a new code is written which converted all cells with ranges into the midpoint of those ranges:

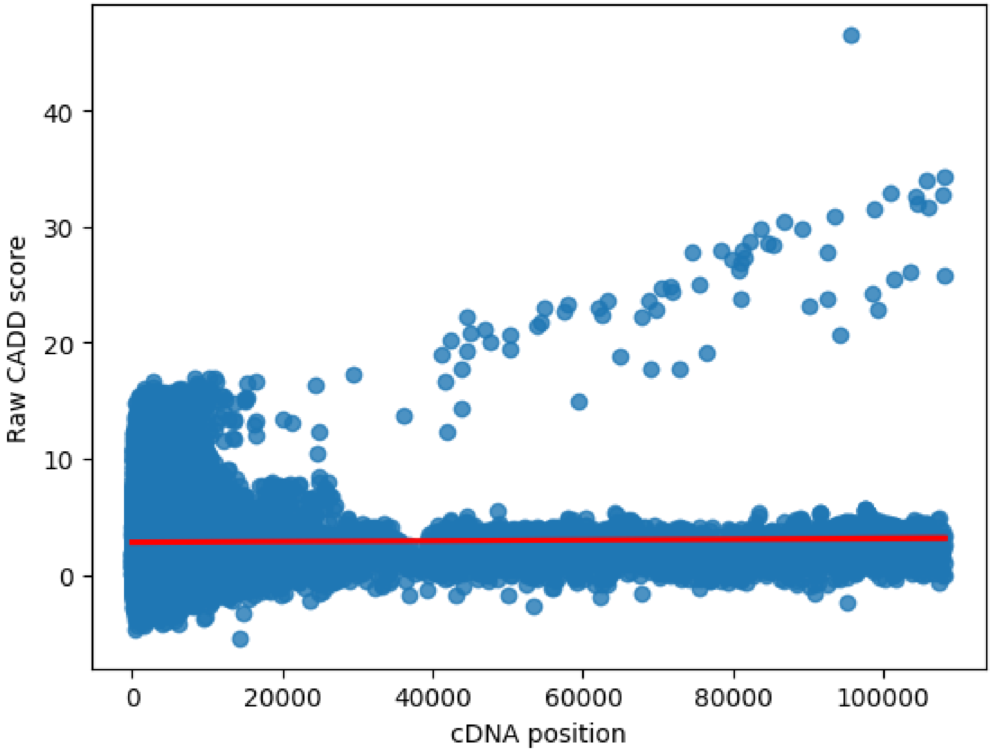


```python
cdna_fixed = []
for i in range(len(df["cDNA position"])):
    e = df["cDNA position"][i]
    if "-" in str(e):
        vals = str(e).split("-")
        if vals[0] == "?":
            cdna_fixed.append(vals[1])
        elif vals[1] == "?":
            cdna_fixed.append(vals[0])
        else:
            cdna_fixed.append((float(vals[0]) + float(vals[1])) / 2.0)
    else:
        cdna_fixed.append(e)
no_nan = pd.concat([pd.to_numeric(pd.Series(cdna_fixed, name="cDNA position")), pd.to_numeric(df["Raw CADD score"])], axis=1).dropna(axis=0, how="any")
stats.linregress(x=no_nan["cDNA position"], y=no_nan["Raw CADD score"], alternative="two-sided")
```

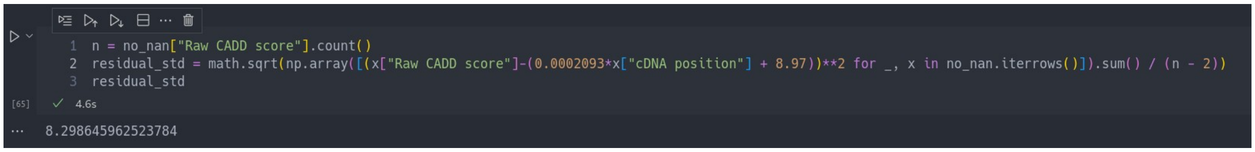
This was the result of the regression:

```
Line Regression Result: (slope = 3.2935755550719284e-06, intercept =
2.783026864305054, r value = 0.014102473034924606, p value =
0.0009101226162724868, std err = 9.929050140318108e-07, intercept std
err = 0.013930960840785314)
```

$$\widehat{CADD}=3.294*10^{-6}*cDNA+2.783$$



The standard deviation of residuals had to be calculated manually:

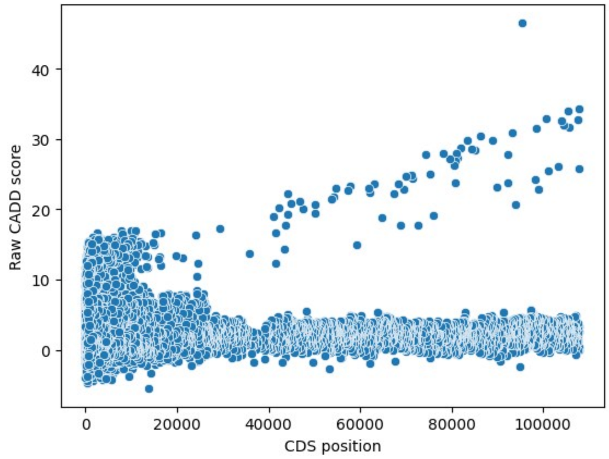


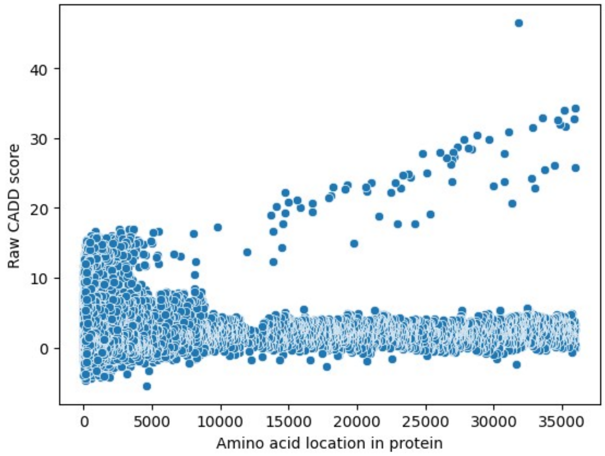
```python
n = no_nan["Raw CADD score"].count()
residual_std = math.sqrt(np.array([(x["Raw CADD score"]-(0.0002093*x["cDNA position"] + 8.97))**2 for _, x in no_nan.iterrows()]).sum() / (n - 2))
residual_std
```

```
8.298645962523784
```

As can be seen here, even though the correlation has a very low p-value, meaning that the relationship between the variables cannot be due to chance, it also has an incredibly low slope which implies an incredibly weak correlation. According to the R^2 value of 0.020%, which has been calculated from the r-value of 0.0141, our linear model only accounts for around a fiftieth of a percent of the variation in the data. After graphing the two variables, an observation can be made that even though the vast majority of data points had no relationship between its CADD score and its cDNA position, a few had really high CADD scores which formed a strong linear relationship between the two variables. It was because of this heteroscedasticity, failure of the "does the plot thicken" condition, that caused the low p-value despite the weak slope-of-best-fit. Usually in this situation, the residual plot should be graphed to prove homoscedasticity. However, since the slope is virtually nonexistent, and the plot clearly thickens, it is not deemed necessary.

Other variables that were thought to be related to the cDNA position were also graphed. Turns out they were near-exact proxies for the cDNA position:





Why does the relationship exist?

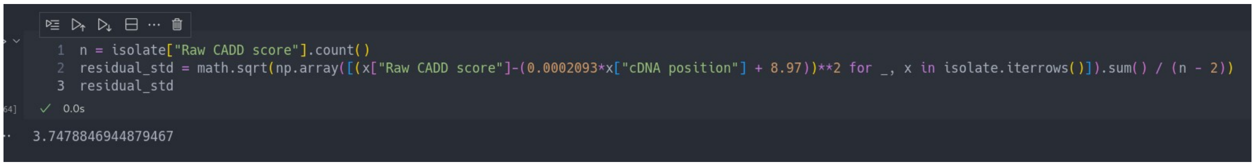
Since all the data points with the correlation have abnormally high CADD scores, the relationship might be due to a bug within the CADD scoring algorithm. To investigate this, all the values with a CADD score of 10 or higher and a cDNA position of 20k or higher were isolated to help determine the cause of this abnormality. This is the result of the regression on the isolated values:

```
Line Regression Result: (slope = 0.0002093087211034773, intercept =
8.970439791124365, r value = 0.8048599887512006, p value =
1.6089504085661705e-17, std err = 1.8446453753029453e-05, intercept
std err = 1.3804458691555022)
```

$$\widehat{CADD\,isolated} = 0.0002093 * cDNA\,isolated + 8.97$$

This doesn't include the R^2 and the standard deviation of residual, so it is calculated below:

$R^2 = r^2 = 64.78\%$

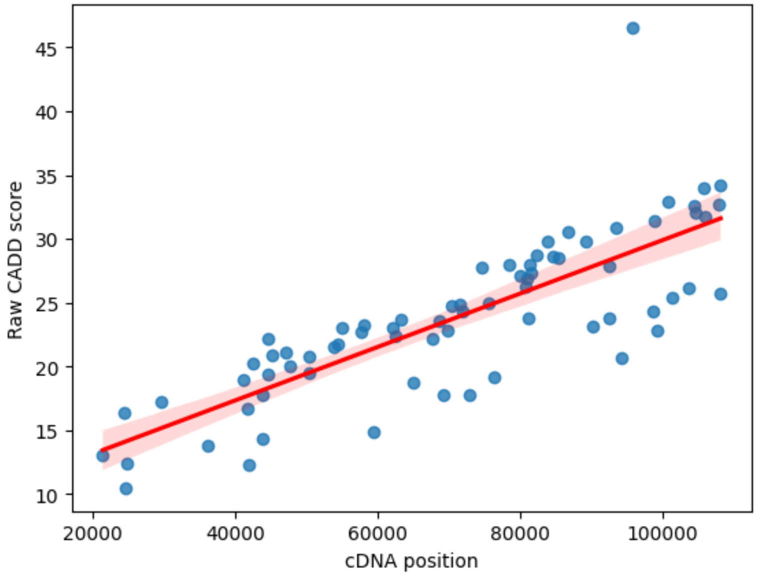


```
n = isolate["Raw CADD score"].count()
residual_std = math.sqrt(np.array([(x["Raw CADD score"]-(0.0002093*x["cDNA position"] + 8.97))**2 for _, x in isolate.iterrows()]).sum() / (n - 2))
residual_std
```

✓ 0.0s

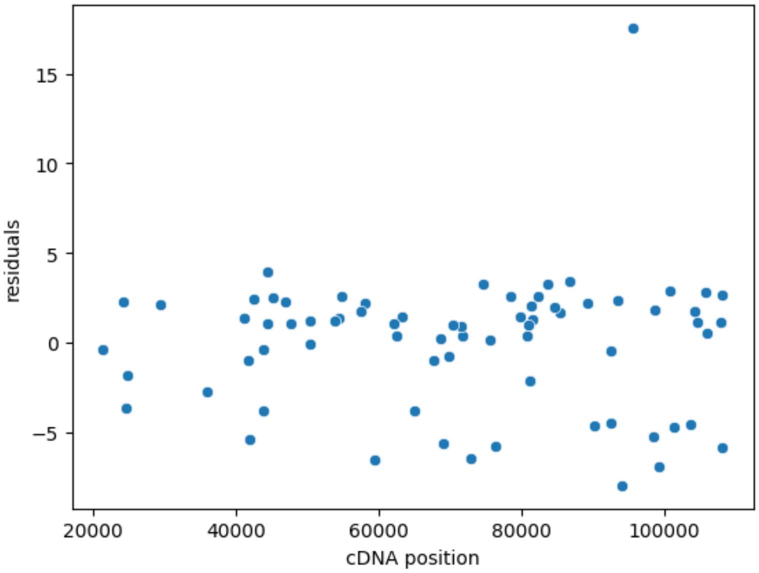
```
3.7478846944879467
```

This is the resulting graph:



Using the linear model, 64.78% of the variation is taken into account in this part of the graph, and the standard deviation of the residuals were halved. The correlation between the data is even less likely to be due to chance, as seen with p-value = 1.6×10^{-17}, and the slope of the best fit line actually means something after it grew by 2 orders of magnitude. For this specific area of data, a slope of 0.0002093 means that for every increase of one position in the cDNA position of the mutation, the raw CADD score increases by 0.0002093. This may seem like a small amount, but since the CADD position within this graph ranges from 20k to 100k, it adds up to a score that's anomalously high.

The residuals of this subset were also graphed:



Horizontally, the residuals seem pretty randomly distributed. Vertically, it seems that the data is more clustered above the trendline and an outlier exists at the top right of the graph. Because there is a decent sample size, even in this small subset, the outlier doesn't seem to have too much of an impact. Even though this data point was a rather extreme outlier, there doesn't seem to be much unusual about it other than its unfathomably high CADD score.

There are a few more things of note regarding this subset of data: First, knowing that the CADD scores of these data points are extremely high, it is not surprising that the LoFtool scores have a mean of 0.972 with a standard deviation of merely 0.0048. As a reminder, LoFtool scores range from 0 to 1, with the danger level increasing as it approaches 1. The mean of the LoFtool scores in the dataset is 0.345 so the scores in this subset are abnormally high. All the other scores (e.g. SIFT, PolyPhen), however, lacks even the values for these data points. Perhaps the lack of values are due to the same factors which lead to such high CADD and LoFtool scores. The following is a list of observations, which could be reasons for these abnormal values, about this subset in no particular order:

- All these mutations are on chromosome 2
- They all have a negative STRAND type
- They all have either `stop_gained` or `frameshift_variant` consequence, which was present in the consequence vs SIFT hypothesis test as none of these consequences have SIFT scores.
- They all have HIGH impact
- Most have the symbol TNN, with three having NBB
- The diseases which these data points are related to (nemaline myopathy, muscular dystrophy, dilated cardiomyopathy) all primarily affect the muscle.

More statistical analysis and biology research is needed to understand these specific values. As it stands currently, it is not suggested to use these features in order to predict anything; whether it be predicting how dangerous a mutation is or whether a data point was going to be in this subset or not.

Conclusion

In conclusion, we fail to reject the null hypothesis. Throughout the various tests conducted, there have been correlations between certain gene expression variables with the danger scores provided by different algorithms, such as SIFT and CADD scores. Although there are also tests that prove the opposite—that being, no correlation—, those tests notably the Regression Analysis could simply be due to a difference in algorithm used. The difference between algorithms can be seen in the single variable analysis, where each score gave different stat summaries to each other, with varying means, standard deviations, and five number summary. Tests that do support the hypothesis are mostly seen in the two variable hypothesis tests, specifically with the SIFT score vs Impact test and the SIFT score vs Consequence test. Within those tests, the p values of , 8.91×10^{-12} and 4.66×10^{-19} respectively, far subceed the alpha value of 5%, showing that the relationship between the two variables to be far too unlikely to be caused by pure random chances. This gives a statistically significant correlation, therefore supporting the hypothesis. Although, there is a test within the bunch of 3 that goes against the hypothesis. The CADD vs Strand test did not show a statistically significant conclusion due to outliers skewing the p value. However, this can simply be seen as a red herring, given not every variable within the data has to be correlated to the scores.

Moving on from this project, further analysis could be done in regards to the findings of the regression analysis between CADD vs cDNA. Within the initial test done of the two, it is found that the regression line has a very flat slope of $\widehat{CADD} = 3.294 * 10^{-6} * cDNA + 2.783$, signifying a weak correlation. However, upon further analysis via graphing, it is found that there are 2 distinct lines that can be seen, one of which seems to demonstrate a strong linear relationship between the variables of CADD and cDNA. Coupled with the fact that the original slope only takes into account 2% of the data variation, another regression analysis was done. The second regression line includes only a subset of the original, that subset includes only the values with a CADD score of 10 or higher and a cDNA position of 20k or higher, values that are in the "strong linear relationship line". This time the analysis gave a regression line of $\widehat{CADD\,isolated} = 0.0002093 * cDNA\,isolated + 8.97$ and takes into account 64.78% of *that* data. There are many observations that can be made with the second test, examples include: the slope being less flat which suggests a better correlation, but more notably the values of the subset all include abnormally high CADD scores. This begs the question of why? To answer this question more analysis must be done between the variables. In addition, by conducting more hypothesis tests it can also be used to predict whether a data point was going to be in this subset or not.