Analysis of the effects of mutations

Student name：Yichao Gao

Student ID:52428004

**1.Introduction**

Understanding the biological implications of genetic mutations is critical for advancing personalized medicine, disease modeling, and therapeutic design. In recent years, bioinformatics approaches have been instrumental in accelerating mutation discovery and prioritization pipelines, especially when dealing with large genomic datasets.

In genes, mutations in sequences can affect and alter the level of molecular expression and cell viability(**Jia & Zhao, 2017**). Mutations can not only alter gene structure and expression but also impact metabolic activity and cellular behavior. Systematic analysis of mutation data using computational biology tools helps identify potential pathogenic variants or functional regulatory elements(**Anon, 2015**).

The aim of this analysis is to determine which factor has the greatest effect on a biological system by analysing changes in mRNA,protein expression and cell viability. The largest mutations can be carried out. A set of mutated genes was evaluated through python computational methods and the ability to track and analyze the data to visualize the data to filter out target genes worthy of the next step in the experiment.

**2.Methods:**

**2.1. Data preparation**

The dataset includes 49 mutation sample files in tab-separated .txt format. Each file contains wild-type (WT) and mutant (Mut) data for three biological replicates across three biological measures: mRNA expression, protein expression, and cell viability. Using Python's pandas library, all files were batch-imported and merged into a single DataFrame for streamlined processing and analysis. It may contain multiple duplicate values.

**2.2 Classification of mutations**

For each gene:

Mutation type: identified as replacement, insertion or deletion

Mutation region: classified according to the location of the mutation, as promoter region (first 1000 bp) and coding region

**2.3 Expression analysis**

Average values were computed from replicates for WT and Mut conditions.Multiple iterations are needed to calculate the average expression values:

mRNA\_change = mRNA\_Mut\_mean - mRNA\_WT\_mean

Protein\_change = Protein\_Mut\_mean - Protein\_WT\_mean

Viability\_change = Viability\_Mut\_mean - Viability\_WT\_mean

These values represent the net effect of the mutation on transcription, translation, and phenotypic fitness. Changes were calculated for each gene and stored for scoring.

Then use the formula to calculate the score for gene filtering

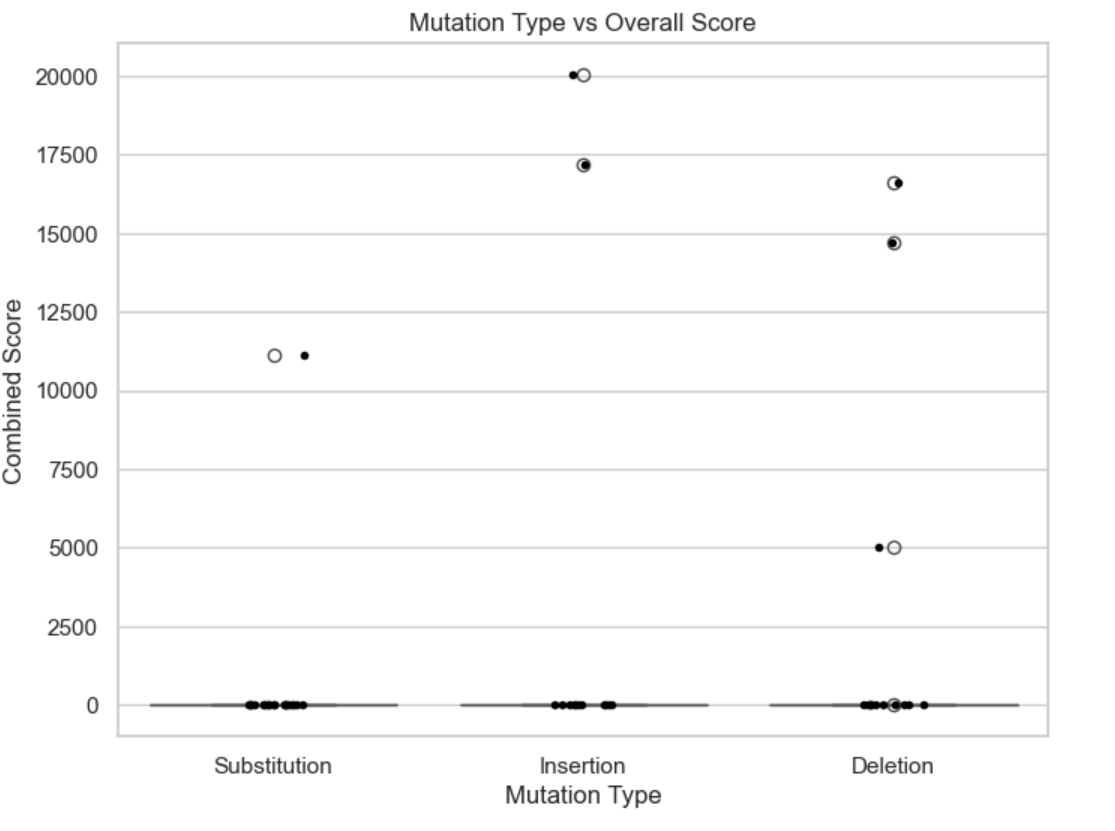
mRNA\_change + Protein\_change + Viability\_change

The five mutated genes with the highest score are filtered

**3.Results**

Python was used to create three images, respectively:

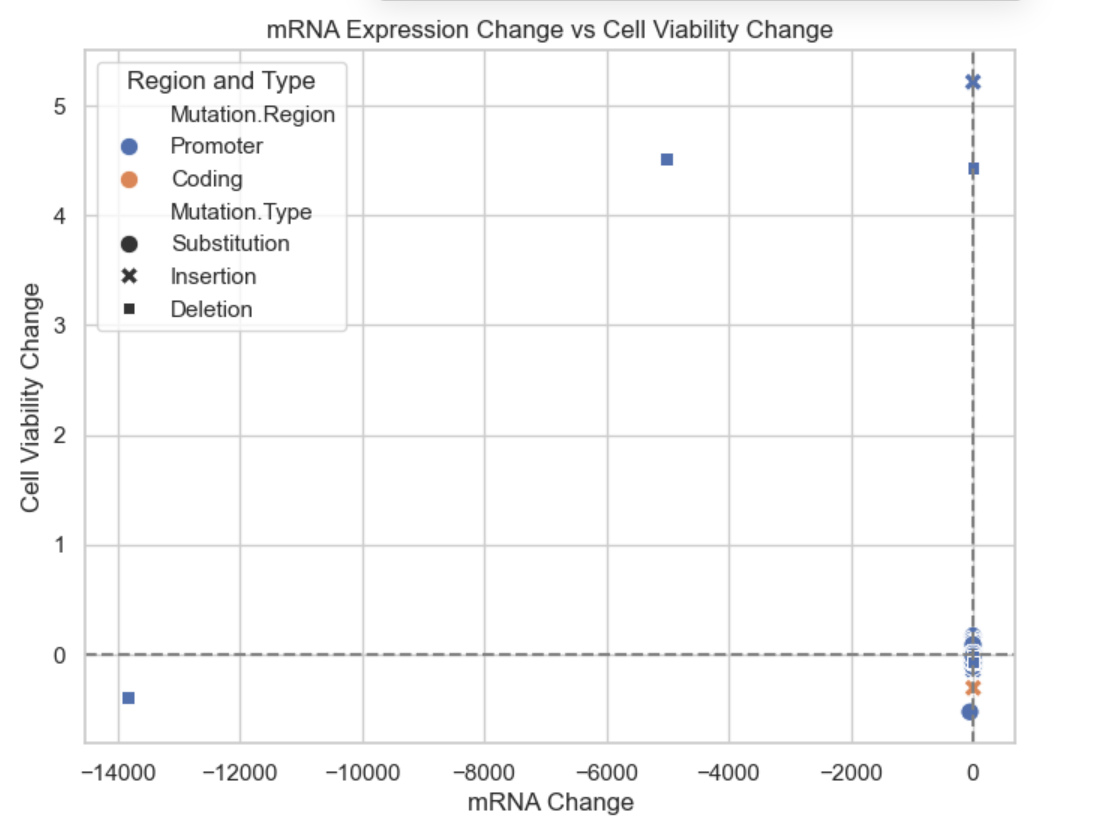
Figure 1: Scores for mutation types and complexes



Boxplot showing the distribution of composite scores by mutation type. Insertion and deletion mutations display greater variability and more high-scoring outliers compared to substitution mutations, suggesting they may have more pronounced biological effects.

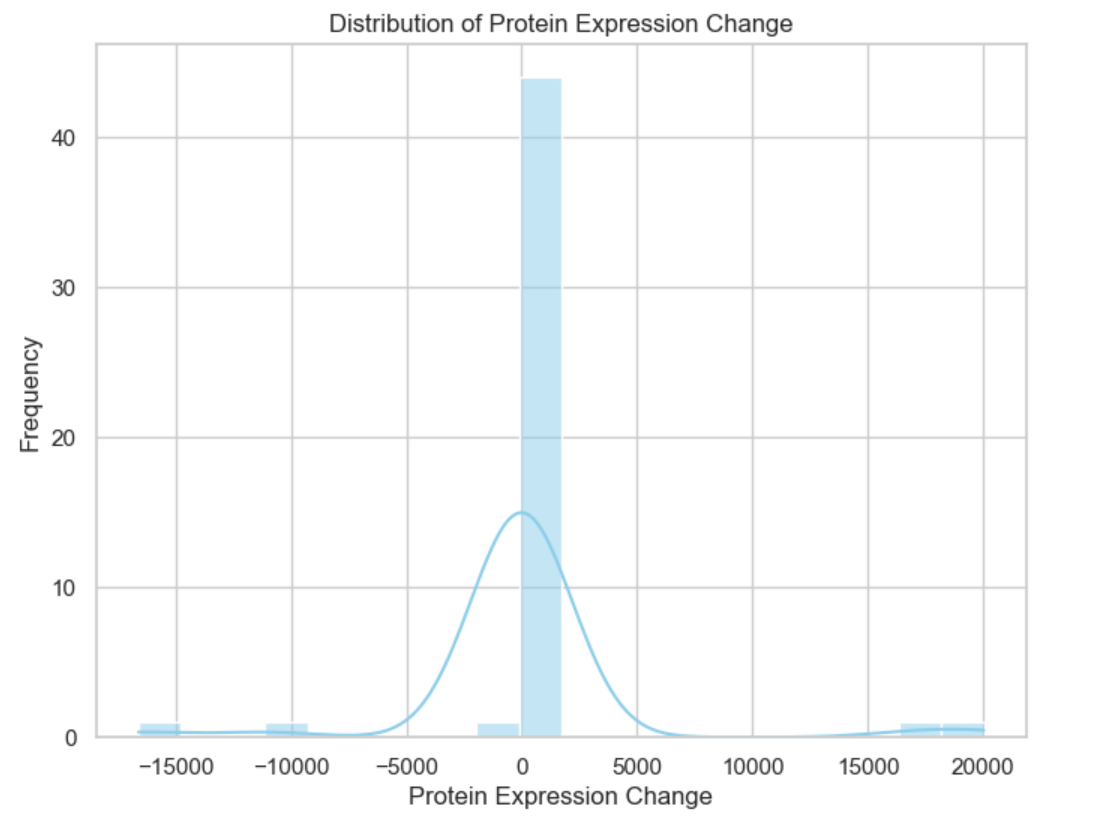
Box plot + scatter plot to show the combined scores for different mutation types (insertions, deletions and substitutions). As can be seen through this figure, most replacement type mutations have lower scores, while the outliers for insertion and deletion type mutations are much more pronounced and may cause a wider range of expression variation.

Figure 2: Changes in mRNA expression in relation to changes in cell viability



A scatter plot was used to show the relationship between changes in mRNA expression as the x-axis and changes in cell viability as the y-axis. Most points are clustered near the origin of the disclosure, suggesting that most mutated genes do not significantly express cell viability and change, but there are points away from the origin that represent mutations that would be strongly effective.A scatter plot illustrates the relationship between mRNA expression changes and cell viability. The clustering around the origin suggests most mutations are neutral.

Figure 3: Distribution of changes in protein expression



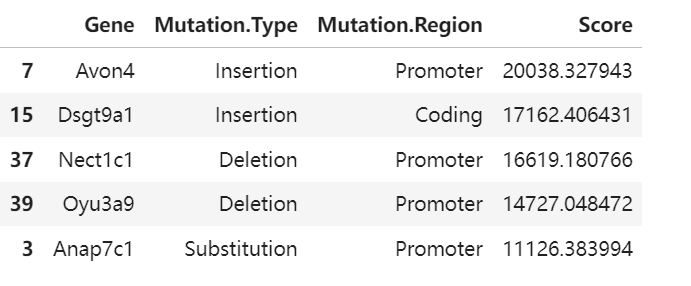
The pattern of the overall distribution of protein expression is shown using a cryptic curve histogram. The graph shows that most mutations have a strong central tendency with the protein change values clustered around 0. The plot shows a central tendency around zero with long tails, representing rare but extreme expression shifts caused by specific mutations.However, there are still distributions on both sides which are worth further investigation of these genes.

**4. Discussion**

In this project, we filtered a number of mutations that stand out at both the expression and functional level using a scoring system in python. However, it is worth noting that a high score does not mean that it is associated with a drastic change in a single expression type, but is usually the result of a synergy of all three. Integration of mutation type and region information reveals that insertion and deletion mutations tend to break the original coding sequence, leading to an overall expression imbalance. While some of the mutations in the promoter region affect mRNA levels more, mutations in the coding regions may affect protein translation and cell viability more. Extreme values are usually found in the coding region, indicating damage to protein structure or toxicity,which is consistent with previous findings in cancer-related studies(**Maturo et al., 2020**).

**5. Recommendations**

Code writing and computations were carried out using python to come up with the following five mutations as the most suitable topics for the study:



These mutations show dramatic changes in mRNA expression, protein expression and cell viability, have significant research value and can be used as a focus for subsequent experiments to verify whether they have a specific effect on phenotype and pathway.

1. **Conclusion**

This project successfully applied computational techniques to quantify and visualize the impact of gene mutations across multiple biological levels. By combining expression data with mutation classification and scoring, we developed a reproducible workflow for prioritizing potentially impactful mutations. The integration of Python-based data processing and visualization tools allowed for efficient analysis of complex datasets, facilitating deeper biological interpretation.

Our findings highlight the heterogeneity in mutation effects—while many mutations yield minimal changes, a select group exhibits substantial shifts in mRNA levels, protein expression, and cell viability. These results suggest that not all mutations are biologically equivalent, and certain mutation types and regions are more prone to functional disruption. In particular, insertions and deletions in coding regions frequently ranked among the top-scoring mutations, reinforcing their potential to alter gene function significantly.

Beyond individual scores, the consistency of certain gene responses across metrics strengthens the case for targeted experimental follow-up. These candidate mutations not only serve as strong leads for functional validation but also demonstrate the utility of integrative bioinformatics pipelines in early-phase research.

Overall, this analysis provides a foundational approach to mutation effect screening, combining clarity, scalability, and biological relevance. As genomic datasets expand and experimental resources remain limited, such data-driven prioritization strategies will be invaluable in guiding hypothesis generation and efficient experimental design.

**Instructions for the use of AI tools**

Portions of this report and python code were made using chatgpt to help the following:

Explaining the structure of certain elements of the code and helping to resolve reported bugs

Helping me create the code that generates the output results and stores the images

The vast majority of the report was done independently by myself.

**Reference lists：**

Anon (2015) Improving correlated mutation analysis: Bioinformatics. [Online] 12 (11), 1009–1009.

Jia, P. & Zhao, Z. (2017) Impacts of somatic mutations on gene expression: an association perspective. *Briefings in bioinformatics*. [Online] 18 (3), 413–425.

Maturo, M. G. et al. (2020) Coding and noncoding somatic mutations in candidate genes in basal cell carcinoma. *Scientific reports*. [Online] 10 (1), 8005–8005.