Hacettepe University

Artificial Intelligence Engineering Department

Research Project - 2025 Summer

Dataset Analysis - PART I

August 30, 2025

Student:
Mohamed Yahya Mansouri

Supervisor:
Dr.Gülden Olgun

Contents

List of Figures

I Introduction

This document provides findings of data analysis for the three datasets on somatic mutations along with their phenotypes and survival time. This data was taken from National Common Institute GDC (Genomic Data Commons). The data can be accessed from [3], [4], [5].

II Mutation Dataset

This dataset represents different samples of somatic mutations primarily comprising of Single Nucleotide Polymorphisms (SNPs) and a smaller fraction of insertions/deletions (INDELs) mutations. This dataset is made of 11 columns and 3175929 entries. After performing data analysis, detailed below, this dataset was found to be biased which may negatively affect model performance for synthetic lethality (SL) prediction later on. It exhibits the same selection bias that was discussed in [12] and [13]: the over-representation of some genes and cancer types which can potentially introduce overfitting.

II.I Features' Values Understanding

Table 1: Feature Description Table with Sample Entries

Feature Name	**Description**	**Sample Entry**
Sample_ID	Unique identifier for the sample, a TCGA barcode	TCGA-XX-XXXX-XXX
gene	Name of the affected gene	TTN
chrom	Chromosome where the variant is located	chr2
start	Genomic start coordinate of the variant	178717193
end	Genomic end coordinate of the variant	178717193
ref	Reference allele at the variant position	G
alt	Alternate allele observed in the sample	A
Amino_Acid_Change	Protein-level change resulting from the variant (HGVS format)	p.T8197I
effect	Predicted functional consequence (e.g., missense, nonsense, silent)	missense$_variant$
filter	Quality or annotation filter applied to the variant (e.g., PASS, population frequency)	PASS
dna_vaf	Variant allele frequency in DNA sequencing data	0.500000

II.I.1 Amino_Acid_Change feature

This feature represents protein based mutations shown in the HGVS protein nomenclature format[8]. It has the highest number of missing values reaching to 575969 rows, 18% of the whole mutation dataset. According to [10] page 16 column 56, these missing values are not random or due to experimental errors, they are intentional. The values in this feature are given only if the mutation is protein-based otherwise will be left empty. During data analysis, this was verified through investigating the type of effects these variants have. Only 3.33% of the effects were attributable to protein-level or near protein-level changes, supporting the interpretation that the null values in the *Amino_Acid_Change* feature reflect a genuine absence of protein-level impact from the mutations. **These results suggest data imbalance since protein-level mutations represent 82% of the dataset.**

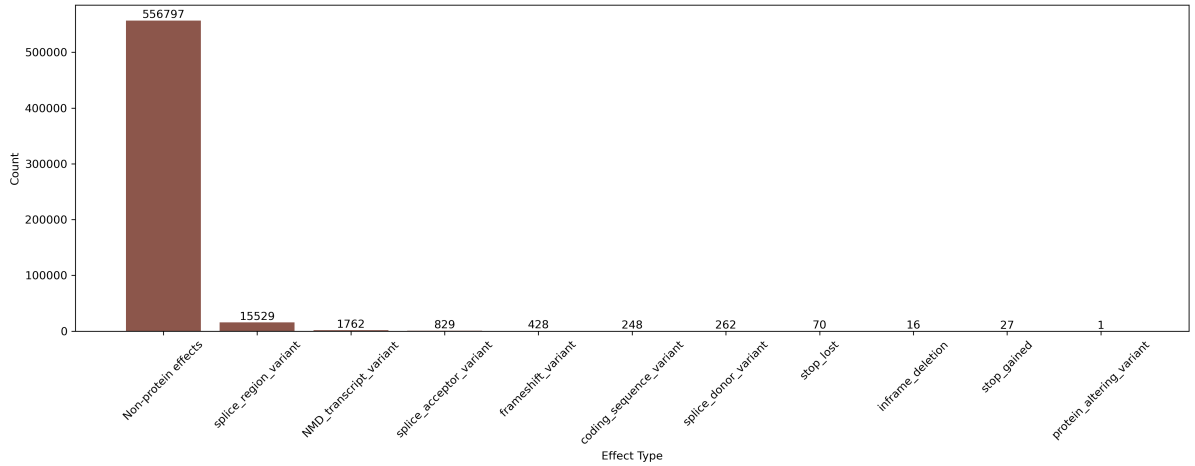


Figure 1: Characterizing Mutation Effect Profiles to Interpret Amino Acid Annotation Gaps.

II.I.2 Filter Feature

This feature indicates whether certain variants passed some quality control tests to evaluate if a mutation is genuine and reliable and not an illustration of technical artifacts or some other errors. A value of *"PASS"* means the variant is considered high-confidence and suitable for analysis. 3035380 samples, approximately 96% of the dataset, are considered reliable for the future work while the rest should be discarded since noise and uncertain variants can introduce bias.

II.I.3 dna_vaf Feature

According to [4], *dna_vaf* or DNA variant allele frequency is:

$$dna\_vaf = \frac{t\_alt\_count}{t\_depth} \tag{1}$$

It is equal to the number of reads with a variant divided by the total number of reads (with and without variants). It is used to estimate how common the mutation is in a sample. [1] discusses how VAF is a factor that estimates the confidence to trust the truthiness of a specific mutation and filter out false positives. A high VAF increases confidence that a mutation is real and clonal, meaning it is presented in large fraction of tumour cells and biologically significant. Conversely, a low VAF value decreases confidence indicating the emergence of passenger mutations[1] instead of driver mutations[2]. [15] **mentions that mutations with very low VAF is challenged by errors arising during PCR[3] amplification and sequencing as a result, they should be filtered out from the dataset.**

One row with missing *dna_vaf* was found and discarded as there is no meaning for the missing value.

II.II Encoding dataset

One of the main problems this dataset has, due to its biological nature, is high-cardinality categorical variables, meaning categorical variables have a huge number of unique values. Traditional encoding methods such as label encoding or one-hot-encoding are not effective. On the contrary, they exacerbate the problem through introducing numerical order (label encoding) or

[1] They are mutations occuring in cancer cells with no direct role in their progression[14]
[2] They directly enable cancer cells to proliferate and evolve[14]
[3] Polymerase Chain Reaction (PCR) to copy/amplify sepcific DNA segments[11]

the curse of dimensionality (one-hot-encoding). Other methods such as "feature aggregation" or more advanced encoding techniques such as entity embeddings can be employed to mitigate this problem.

ref and *alt* features can be grouped into a single feature called *type* describing the type of the observed variant. This feature aggregation is primarily backed up by [2] and the fact this dataset is a somatic mutation dataset composed of SNPs and small INDELs.[4].

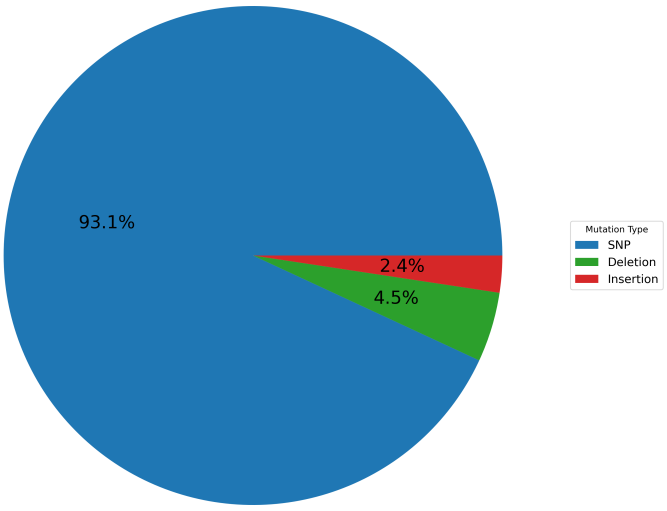


Figure 2: Mutation Type Frequency: Single-Nucleotide Polymorphism (SNP), Insertion, Deletion

II.III Cancer Feature

[12] discussed the problem of over-representation of some genes and cancer types as some remain under-studied for historical or academic reasons. For example, cancers with higher prevalence in the general population, such as breast or lung cancer dominate scientific attention and experiments. As a result, to assess the bias within the mutation dataset, a cancer feature was added that describe the type of cancer the mutations come from. The cancer names were taken from the phenotype dataset[3]. No missing values were found as all samples from the "mutation data" were found and covered in the "phenotype data". 33 cancer types were found.

These 10 cancer types are the most frequent ones.

Table 2: Distribution of Cancer Types by Percentage

Cancer Type	UCEC	SKCM	COAD	STAD	LUAD	LUSC	BLCA	BRCA	CESC	HNSC
Percentage	27.91%	12.36%	8.34%	6.71%	6.55%	5.70%	4.24%	3.81%	3.26%	3.22%

II.IV Bias Diagnostic

According to Figure 3a and 3b, the *gene* feature exhibits a long-tail distribution where the most frequent genes are only clustered in the head of graph Figure 3a. The head of the graph is made of the first four genes: TTN, MUC16, TP53, and SYNE1. While the difference between these genes is distinguishable, the observed percentages make them a rare event. Due to the huge number of the entries in the dataset, they are sparsely exhibited. As a result, predictive models

might not generalize well. **One hypothesis for this result of the frequent genes can be due to how vastly these genes are researched and especially how long they are to be easily accessed and affected by mutations.** Using the REST API for Ensembl to access vast genmoic data [9], gene length information were retrieved to test the hypothesis. Figure 4a represents the distribution of gene length. It exhibits how gene length follows the same behavior observed in Figure 3a as the most frequent genes possess the highest gene length. This suggests that gene length plays a significant role in mutation accumulation. Figure 4c shows gene length for the first 200 most frequent genes. This graph showed intrinsic deviations of some genes having longer lengths but relatively lower mutation counts and vice versa. For example TTN and MUC16 (the first two bars in the very left of Figure 4c) are the most frequent genes, yet other genes are much larger. This suggests that there are other biological factors influencing mutation.

Figure 4b shows the correlation between *gene length* and *mutation count*. A moderate positive correlation of value 0.38 supports the interpretations made from graph 4a, 4c, and 4d.

In conclusion, this dataset demonstrates gene bias. This is manifested in the high number of mutations affecting the dominating genes. This is partially due to the high length of genes. To account for these variations, a potential solution would be gene normalization. This solution will be investigated in Data Analysis - PART II.

dna_vaf, variant allele frequency, feature was also investigated. The value of this feature is in the range of $[0, 1]$. dna_vaf has a lot of mild outliers but none extreme ones[4]. Figure 5a shows a boxplot that exhibits a very large set of mild outliers. In fact, tohse points describe data points that have very high dna_vaf values which is only 1.8% of the dataset, 57406 entries. This is a sign of data imbalance for genes with high values in this feature.

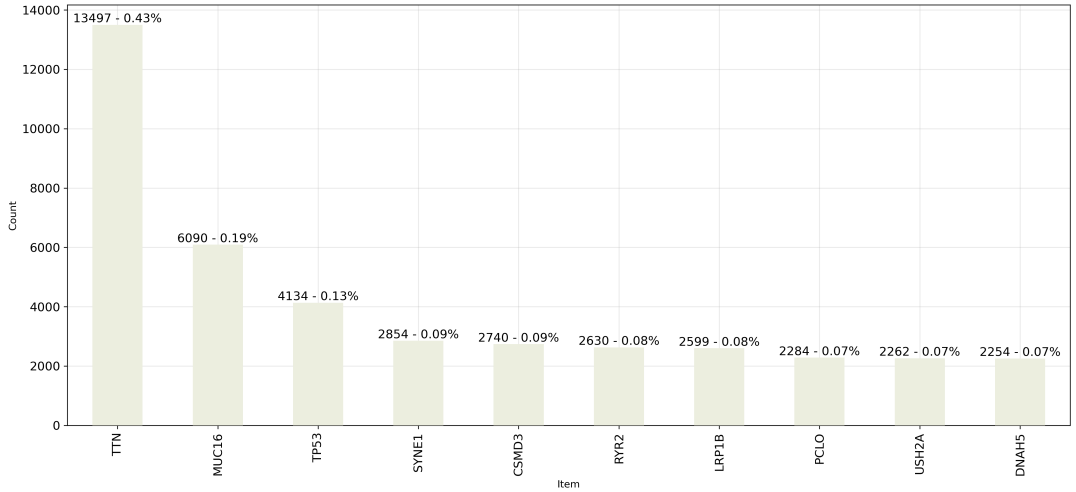
Another interesting observation is that most frequent genes with very low *dna_vaf* values **were not** the most frequent genes with very high *dna_vaf* values. Figure 5b and 5c illustrate these observations. **Interestingly, if the very low *dna_vaf* values should be discarded, then gene bias might decrease since some of the most frequent genes will be removed as well.** There are no further explanations or interpretations made from the obeserved phenomenen in Figure 5, but more will be made for Data Analysis - PART II.

III Other

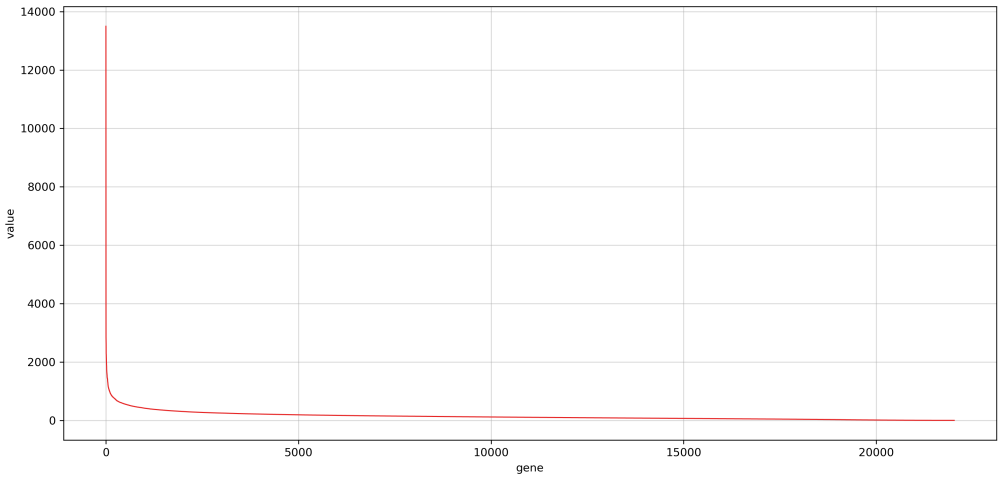
About 309567 rows, 9.7% of the dataset, are repeated observations of mutations in the same biological identifiers *(gene, chrom, start, end, ref, alt, Amino_Acid_Change, effect, filter)* with meaningful differences in *dna_vaf* and different *Sample_ID*. A possble approach to this would be keeping the rows with the highest *dna_vaf*. Since the number of genes would decrease, then so gene frequency, affecting frequent itemset mining later on. One concerning thought should be the removed *Sample_IDs* which can influence survival analysis later on. However, it turns out that there are only two removed Sample IDs from the whole dataset.

Some research was done concerning frequent itemset mining, for curiosity purposes only, and there were two algorithms I learned about: Apriori algorithm[6] and Frequent Pattern Growth (FP-Growth) algorithm[7]. Aside from this, I learned about Survival Analysis Kaplan-Meier curve along with Cox hazardous model along with their Python implementation and survival interpretations. **I am only mentioning this to account for the time spent on these topics instead of data analysis.**

[4]Mild outliers are data points outside the range of $1.5 \times IQR$ but less than $3 \times IQR$ while extreme outliers are data points outside the range of $3 \times IQR$

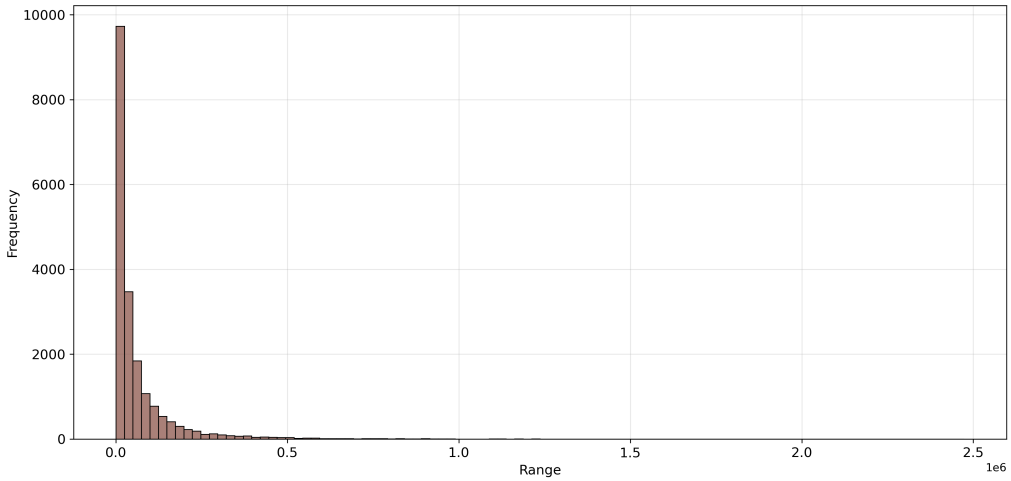


(a) The 10 Most Frequent Observed Genes in the Dataset

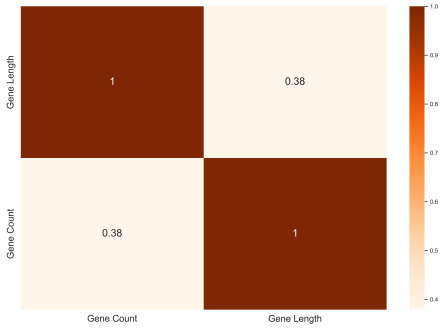


(b) Gene Distribution Across all Entries in the Dataset

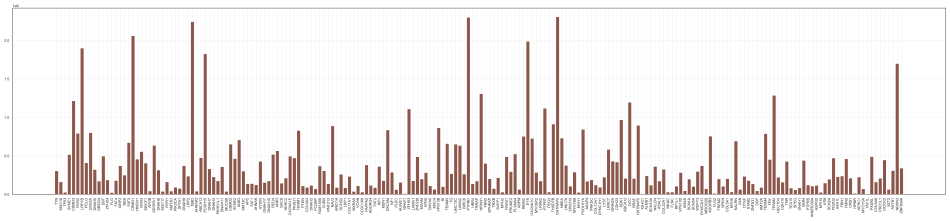
Figure 3: Distribution of Gene Occurrence in the Dataset: Frequency Spectrum and Long-Tail Characteristics



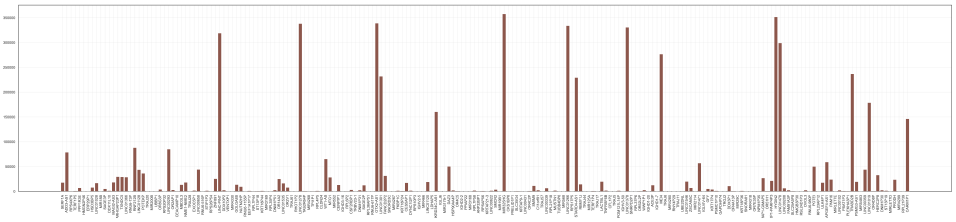
(a) Distribution of Gene Length Ranked by Mutation Count in Descending Order.



(b) Correlation Heatmap between Gene Length and Mutation Count



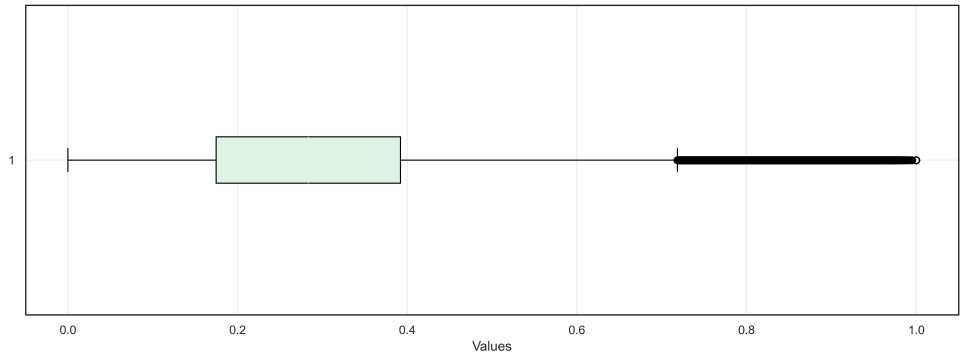
(c) Gene Length for the Most 200 Frequent Genes



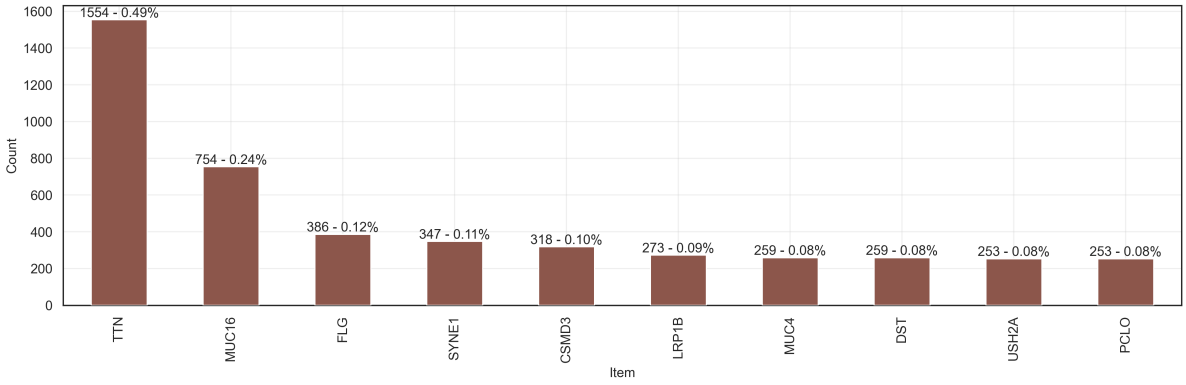
(d) Gene Length for the Least 200 Frequent Genes

While Figure c and d exhibits great deviations they still demonstrate the general pattern how gene length decreases with frequency.

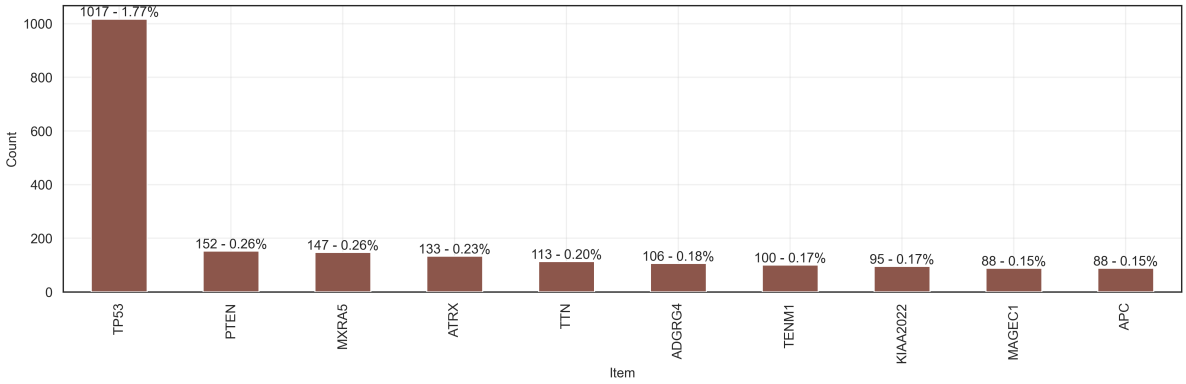
Figure 4: Investigating the Reason Behind the Result of Most Frequent Genes.



(a) *dna_vaf* Feature Boxplot



(b) Top 10 Most Frequent Fenes With Very Low *dna_vaf*



(c) Top 10 Most Frequent Genes With Very High *dna_vaf*

Figure 5: Comparsion of entries with very low *dna_vaf* and those with very high *dna_vaf*

References

[1] Luca Boscolo Bielo et al. "Variant allele frequency: a decision-making tool in precision oncology?" In: *Trends in Cancer* 9.12 (2023), pp. 1058–1068. ISSN: 2405-8033. DOI: https://doi.org/10.1016/j.trecan.2023.08.011. URL: https://www.sciencedirect.com/science/article/pii/S2405803323001711.

[2] Harvard Chan Bioinformatics Core. *Variant annotation with snpEff*. Accessed: 2025-08-30. 2016. URL: https://hbctraining.github.io/In-depth-NGS-Data-Analysis-Course/sessionVI/lessons/03_annotation-snpeff.html.

[3] *GDC Pancan Basic Phenotype Data*. Accessed: 2025-08-25. URL: https://xenabrowser.net/datapages/?dataset=GDC-PANCAN.basic_phenotype.tsv&host=https%3A%2F%2Fgdc.xenahubs.net.

[4] *GDC Pancan MuTect2 SNV Data*. Accessed: 2025-08-25. URL: https://xenabrowser.net/datapages/?dataset=GDC-PANCAN.mutect2_snv.tsv&host=https%3A%2F%2Fgdc.xenahubs.net.

[5] *GDC Pancan Survival Data*. Accessed: 2025-08-25. URL: https://xenabrowser.net/datapages/?dataset=GDC-PANCAN.survival.tsv&host=https%3A%2F%2Fgdc.xenahubs.net.

[6] GeeksforGeeks. *Apriori Algorithm*. https://www.geeksforgeeks.org/machine-learning/apriori-algorithm/. Last updated: 11 July 2025. 2025.

[7] GeeksforGeeks. *Frequent Pattern Growth Algorithm*. https://www.geeksforgeeks.org/machine-learning/frequent-pattern-growth-algorithm/. Last updated: 12 July 2025. 2025.

[8] "HGVS Nomenclature". In: (). URL: https://hgvs-nomenclature.org/stable/recommendations/general/.

[9] European Bioinformatics Institute. *Ensembl REST API*. https://rest.ensembl.org/. Accessed August 2025. 2025.

[10] National Cancer Institute. "GDC Data User's Guide". In: (). URL: https://docs.gdc.cancer.gov/Data/PDF/Data_UG.pdf.

[11] National Human Genome Research Institute. *Polymerase Chain Reaction (PCR)*. 2025. URL: https://www.genome.gov/genetics-glossary/Polymerase-Chain-Reaction-PCR.

[12] Colm Seale, Yasin Tepeli, and Joana Gonçalves. "Overcoming Selection Bias In Synthetic Lethality Prediction". In: *Bioinformatics* 38 (July 2022). DOI: 10.1093/bioinformatics/btac523.

[13] Yasin I Tepeli, Colm Seale, and Joana P Gonçalves. "ELISL: early–late integrated synthetic lethality prediction in cancer". In: *Bioinformatics* 40.1 (Dec. 2023), btad764. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad764. eprint: https://academic.oup.com/bioinformatics/article-pdf/40/1/btad764/55432096/btad764.pdf. URL: https://doi.org/10.1093/bioinformatics/btad764.

[14] Dominik Wodarz, A. Newell, and Natalia Komarova. *Passenger mutations can accelerate tumor suppressor gene inactivation in cancer evolution*. Oct. 2017. DOI: 10.1101/202531.

[15] Chang Xu et al. "Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller". In: *BMC Genomics* 18 (Jan. 2017). DOI: 10.1186/s12864-016-3425-4.