Title
EXPLORATORY DATA ANALYSIS AND CLUSTERING METHODS OF TP53 FUNCTIONAL TRANSACTIVATION PROFILES

Author
Palmukhopadhyay, Tista

Publication Date
2026-06-15

EXPLORATORY DATA ANALYSIS AND CLUSTERING METHODS OF TP53 FUNCTIONAL TRANSACTIVATION PROFILES

By

Tista Palmukhopadhyay

A capstone project submitted for Graduation with University Honors

May 15, 2026

University Honors

University of California, Riverside

APPROVED

Dr. Xiaoqian Liu

Department of Statistics

Dr. Begoña Echeverria, Howard H Hays, Jr. Chair

University Honors

TABLE OF CONTENTS

ABSTRACT

TP53 is one of the most commonly mutated genes in human cancers and is linked to major body regulation roles like cell cycle arrest, DNA repair, apoptosis, and oxidative stress responses. Because TP53 mutations display diverse functional behavior, researchers have increasingly relied on clustering analyses to identify the complexity of its behavior. Recent studies have clustered TP53 mutations into functional groups using different analyses and clustering methods. This study, however, applies exploratory data analysis (EDA), principal component analysis (PCA), and unsupervised clustering methods to analyze and comprehend functional transactivation patterns across eight TP53 scores that closely relate to the major regulation roles being studied (WAF1, H1433S, MDM2, BAX, AIP1, GADD45, p53R2, NOX). These scores are measured using a yeast transactivation assay (YTA) (Kato et al.). Clustering techniques and validation, such as k-means clustering, hierarchical clustering, and silhouette analysis, helped identify mutation groups that follow loss-of-function, intermediate, and gain-of-function activity patterns. Results revealed significant variability across mutations, while correlation analyses and PCA found coordinated relationships. Overall, this study values data-driven analysis in the comprehension of complexities in mutational behavior and the significance of utilizing clustering methods to improve mutation classification.

ACKNOWLEDGEMENTS

I cannot express enough gratitude to my faculty mentor, Dr. Xiaoqian Liu, for her support and guidance throughout this project. Her dedication towards harbouring a flexible and insightful journey in my understanding of TP53 and the findings that statistical methods can uncover. I would also like to thank Dr. Begoña Echeverria and the entire University Honors team for their support throughout the past years and for providing resources that helped complete this capstone.

INTRODUCTION & BACKGROUND

The Tumor Protein p53 gene (TP53) is widely recognized as "the guardian of the genome" due to its crucial role in maintaining genomic stability. In 1989, several years after the protein's discovery, TP53 was identified as the target of the common 17p chromosomal deletion in human colorectal cancer, providing significant evidence of its tumor-suppressor functions. Furthermore, in the same year, it was proven that the activation of p53 could block cells from transforming into cancer cells, even when triggered by cancer-causing genes (oncogenes) (Aubrey et al.). TP53 provides instructions for generating the tumor protein p53 to regulate cell division, repair damaged DNA, and initiate programmed cell death.

To detect cellular stress effectively, the p53 protein is primarily located in the nucleus of cells throughout the human body, where it directly binds to DNA. When DNA is exposed to damaging agents such as toxic chemicals, radiation, or ultraviolet (UV) rays, p53 plays a key decision-making role in determining whether the damage is fixable or if cell death (apoptosis) will be initiated.

If the DNA can be repaired, p53 activates a network of genes that possess genetic repair functions. These repairing genes scan the DNA for errors in the genetic code, remove damaged sections, and insert correct nucleotides. However, if the DNA damage is irreversible, p53 can signal the cell to undergo apoptosis to prevent rapid cellular division. These processes help prevent tumor formation and suppress cancerous growth ("TP53 Gene").

When functioning properly, TP53 ensures that damaged or dangerous cells are not divided or replicated and suppresses cancer development. However, mutated TP53 genes are among the most frequently observed genetic alterations in human cancers, often reversing the functions of p53. These mutations account for 29% of global cancer diagnoses, with over 50% of human tumors having the TP53 mutation. The 29 percent in global cases seems minuscule, but it also represents 1 in 3 of all malignant tumors across the global population (Richardson).

Although the tumor suppressive function of TP53 is well-established, there are still considerable implications in determining the clinical impact of its mutations. TP53 mutations are naturally highly heterogeneous; they vary greatly in how they corrupt the gene's function. Types of TP53 mutations include missense, truncating, and splice-site mutations, each affecting p53 in different ways. Missense mutations, the most commonly found, change one amino acid and can weaken or halt the protein from binding to DNA for gene regulation (Aubrey et al.).

In reality, two patients with unique TP53 mutations and the same cancer would have variable outcomes. Grouping TP53 mutations as binary classifications (whether being present or absent) does not provide significant information for cancer risk assessment.

To better understand TP53 mutation behavior, Kato et al.'s team developed a large-scale yeast transactivation assay (YTA) to assess the functional consequences of mutations in the DNA sequences of TP53. The YTA technique is capable of evaluating mutated and normal p53 proteins to determine whether they can activate transcription across several promoters responding to TP53. For reference, transcription is the process of copying DNA instructions into RNA, the single-stranded version of DNA, so a gene can be activated and used by the transcribing cell. The YTA assay enabled researchers to find a transcription measurement in evaluating the activity levels of mutant proteins, as well as to examine the influence of certain mutations on TP53 tumor-suppressing functions (Kato et al.).

In precision oncology, separating mutations into categories like Loss-of-Function (LoF), Gain-of-Function (GoF), and partial activity mutations can create insightful clusters that analyze their implications in the human body. LoF mutations disable p53's ability to suppress tumors, reversing its life-saving function, while GoF mutations also encourage cancerous growths by proliferating cells uncontrollably. Partial activity mutations lie in between these effects, allowing the organism to maintain some level of original function. However, they promote cancer spread by transmitting delayed cellular safeguards, such as apoptosis, against harmful agents. Clinical consequences of these mutations lead to unresponsive tumor therapies, chemotherapy resistance, degenerating outcomes, and exposure to more aggressive cancers (Miller et al.). Clustering methods allow mutations to be behaviorally classified at their molecular level, providing clinical interpretation and improved treatment (Sabapathy and Lane).

The functional transactivation scores used in this study show how mutated p53 proteins retain their ability to activate transcription across responsive promoters. Mutations with relatively low functional scores are associated with Loss-of-Function behavior, while mutations with higher functional scores are reflected with Gain-of-Function behavior. Uniquely, functional transactivation scores measure the actual functional behavior of mutated p53 proteins. TP53 mutations are traditionally classified by the location at which the mutation occurs in the DNA sequences and the type of amino acid change. However, transactivation scores allow researchers to observe the degree of impaired or sustained tumor suppressor functions and find biologically meaningful classifications in cancer research and precision oncology (Kato et al.).

This study aims to implement an exploratory data analysis and unsupervised clustering approaches to classify and understand TP53 mutations based on their transcriptional activity characteristics across eight functional transactivation scores regulated by the p53 protein. These scores are involved in cell cycle arrest, DNA repair, oxidative stress response, and apoptosis, regulating genetic resistance against cancers. This study's purpose is to solve the following research questions: identify which functional transactivation scores have the most variable activity across mutations, which mutations consistently increase or decrease in activity across functional transactivation scores, and which mutations have the most varying activity characteristics from other mutations. In addition to this, by applying k-means and hierarchical clustering approaches, we will be able to reveal how many clusters of functionally similar mutations there are in our dataset, and what kind of biological meanings each cluster corresponds to.

DATA & METHODS

The dataset used in this analysis consists of 2,314 observations, including one categorical variable representing the mutation type, while the other eight numeric variables denote the functional transactivation score measurements regulated by the p53 protein.

The eight selected functional transactivation scores reflect key biological roles in TP53 functions, specifically tumor suppression and regulatory mechanisms. WAF1 and H1433S are associated with regulating the cell cycle to monitor progression and preventing damaged cells from dividing. MDM2 reflects negative regulators of p53, targeting p53 levels for degradation and maintaining feedback control of protein levels. BAX and AIP1 are linked to apoptosis to eliminate damaged or abnormal cells in response to cancerous threats. AIP1 specifically signals proteins of programmed cell death. In contrast, GADD45 and p53R2 are associated with DNA repair by coordinating specific repair enzymes and ensuring genomic stability, with p53R2 specifically supporting nucleotide synthesis. Lastly, NOX is connected to oxidative stress, triggering apoptosis through reactive oxygen species production (Kato et al.).

To ensure data quality and accuracy, the dataset was cleaned by checking missing values and duplicates to avoid bias and redundancy. The collection was then structured accordingly, so that each row corresponds to a single mutation and each column represents the transcriptional defense of the mutated TP53 DNA sequence against the T53-responsive promoter. Given the wide range of activity values (0 to over 450), mutations displayed significant functional variability, from Loss-of-Function, strong activation, and high transcriptional activity.

An exploratory data analysis (EDA) was used as a statistical approach to examine patterns and interpret the intricacy of TP53 genetic mutations. It observed the distributions of gene activity, gene variability across mutations, and correlation relationships among functional transactivation features. These visualizations identify patterns such as skewness, outliers, and differences in transcriptional responsiveness across mutations.

This analysis revealed functional transactivation score distributions that were strongly right-skewed, with a large proportion of low activity mutations and a smaller subset of hyperactive outliers. To address this variability and skewness, a logarithmic transformation was applied to better study transactivation scores for downstream analysis.

After transforming the data, scaling techniques normalized all features to ensure consistent contribution across variables for clustering. This step was crucial in eliminating bias in clustering algorithms, since techniques such as K-means are sensitive to skewed features.

Two primary unsupervised clustering algorithms were used to identify groups of functionally similar TP53 mutations:

- K-means clustering: a technique that partitions data into a specified number of clusters by optimizing the minimum sum of squared errors of each cluster. The K-means method is the most suitable for this dataset as it can efficiently handle large amounts of data due to its high performance (Ikotun et al.).
- Hierarchical clustering: a method of building a tree structure (dendrogram), which reveals relationships between mutations based on similarity criteria. Hierarchical clustering is a

convenient tool for analyzing complex data because of its ability to identify hidden patterns in a dataset and validate clustering results (Murtagh and Contreras).

These two clustering methods provide two different perspectives on clustering patterns and their features. For obtaining the optimal number of clusters, the Silhouette Score can be applied. This method is based on the measurement of the quality of clustering results, more specifically, how each datapoint of its own cluster compares to other clusters. According to this approach, when there comes a moment when the addition of one more cluster will bring minimal change to the sum, we should consider such a point optimal for obtaining the most accurate clustering results (Rousseeuw).

EXPLORATORY DATA ANALYSIS

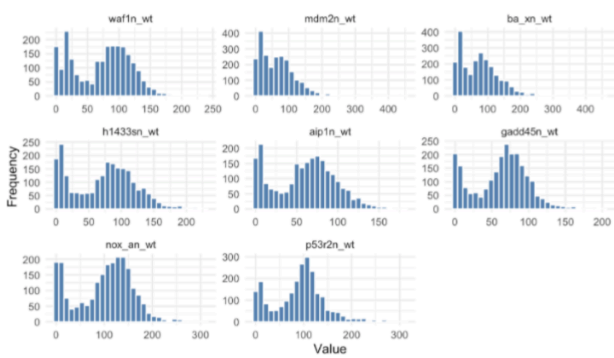

Figure 1. Histograms of TP53 functional transactivation scores distribution showing skewness, variability, and presence of high-activity outliers

The distributions of p53 functional transactivation scores in Figure 1 reveal a range in activity values from 0 to over 250, zero indicating no measurable activity, and 250 representing extreme gene activation. In addition to these distribution patterns, most TP53 mutations appear to reduce or slightly affect activities of the functional transactivation scores, whereas only a small proportion lead to unusually high activity levels.

Across the eight functional transactivation scores, activity levels vary noticeably, implying that TP53 mutations affect transcriptional functionality differently across responsive elements. For example, BAX and MDM2 present the most obvious right tails with a significantly wide range (0-250), suggesting the presence of unusually high-activity mutations. Whereas functional transactivation scores like GADD45 and AIP1 appear with a more concentrated distribution around their median, but still display right-tailed skewness and a few extreme values. Additionally, a large number of lower values suggest that mutations relate to loss-of-function effects. This functional diversity is further evidenced by functional transactivation scores such as H1433S and p53R2, as they help indicate that there are distinct groups of mutations affecting functional transactivation scores in different ways by showing signs of multiple peaks, rather than having a single central pattern.

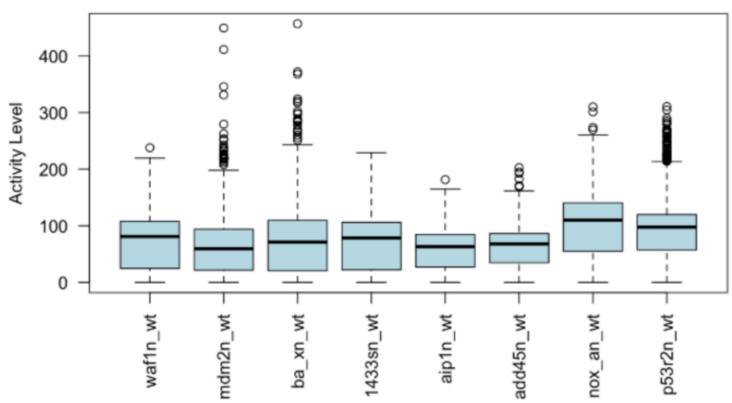

Figure 2. Boxplots of Functional Transactivation Scores displaying distribution, mean activity, and outliers of scores across mutations

Further analyzing the distinct TP53 target functional transactivation scores' activities across mutations, Figure 2 shows boxplots that represent the middle 50% of activity values (typical level of gene activity). The whiskers display the range, excluding extreme outliers or mutations that produce values that are usually high or low in activity. These plots provide a clear distinction among all transactivation scores, as some scores exhibit unique distributions.

The highest median activity levels are presented in the NOX and p53R2 functional transactivation scores, indicating that these functional transactivation scores typically remain relatively active even when TP53 mutations are present. TP53-related transcriptional responses in stress response and DNA repair reflect elevated levels of activity (Vousden and Prives). In contrast, functional transactivation scores like MDM2 and AIP1 show relatively lower median activity. This observation indicates that these functional transactivation scores, influenced by regulatory mechanisms, are less consistently activated and more likely to be suppressed in the presence of TP53 mutations. The limited access to DNA and reduced transactivation score suggest that functional transactivation scores similar to MDM2 and AIP1 are more controlled under mutation conditions, a protective mechanism where the organisms restrict abnormality in the score (Vousden and Prives).

The numerous outliers, as seen in MDM2, BAX, p53R2, and NOX, cause unusually high functional transactivation scores, which may reflect gain-of-function effects (Mantovani et al.). It appears that this extreme behavior is especially relevant for functional transactivation scores involved in apoptosis and cellular stress response functions (Vousden and Prives). The heterogeneity of mutation effects suggests that most mutations do not experience consistent gene activity.

The boxplots are seen to be right-skewed, supporting Figure 1's histograms. The skewness further justifies the need for logarithmic scaling to stabilize the data for clustering analysis.

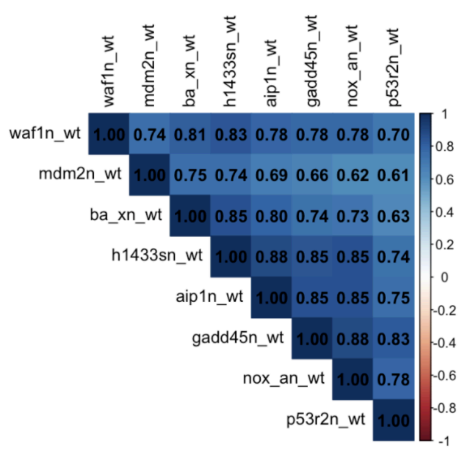

Figure 3. Correlation heatmap displaying positive relationships among TP53 functional transactivation scores

The correlation matrix in Figure 3 shows positive correlations among all TP53 functional transactivation scores, ranging from 0.61 to 0.88. Multiple functional transactivation scores, such as H1433S, AIP1, GADD45, and NOX, reflect coordinated biological responses within a tightly knit group of strongly correlated relations. However, functional transactivation scores that present weaker correlations, like MDM2, may be explained by their negative regulator roles, introducing an abnormal interaction that contrasts with the other functional transactivation scores.

The prevalent positive correlations among these functional transactivation scores indicate that TP53 mutations exert a consistent response across multiple transcriptional responses, even as individual transcriptional responses maintain heterogeneity in feedback (Hassin et al.). The observed biological processes involve apoptosis, DNA repair, and stress-response. The lack of negative correlations among H1433S, AIP1, GADD45, and NOX implies that they may share closely similar regulatory responses to TP53 mutated functions. Moreover, TP53 mutations preserve transcriptional functionality across multiple functional networks, supporting dimensional reduction and clustering methods, as this behavior indicates underlying functional patterns among the dataset (Vousden and Prives)

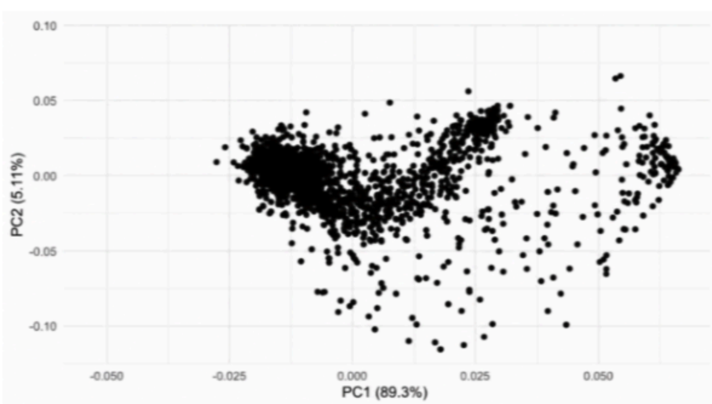

Figure 4. Principal Component Analysis (PCA) of TP53 functional transactivation scores

To simplify complex, high-dimensional data, a principal component analysis (PCA) is used to identify underlying structures and patterns in the TP53 data. Figure 4 splits the high-dimensional gene activity data into two dimensions: PC1 explaining 89.3% of the total variance and PC2 explaining 5.1%. The dominance of PC1 suggests that TP53 mutations have diverse impacts on gene activity levels. These mutations are separated based on overall transcriptional competency, with reduced activity corresponding to lower values and strong activation corresponding to higher values.

The comparably small contribution of PC2 suggests the remaining variation among mutations has subtle nuances and may reflect variation in specific transcriptional responses. Rather than global shifts (large-scale, often evolutionary-driven changes in the distribution of genetic changes within a genome) in gene activity, PC2 may capture subtle distinctions between biological processes of different mutations, supporting the dominant variation in PC1.

The functional transactivation scores from the loadings play a similar role in the creation of the principal components, indicating that their behavior is correlated. Validating the conclusions drawn from the correlation analysis, strong positive correlations are revealed between the functional transactivation scores.

PCA helped establish a clear pattern in the data and supported the presence of clusters of mutations with common functionality (Jolliffe and Cadima). These PCA results also support the use of clustering analysis to classify similar mutations.

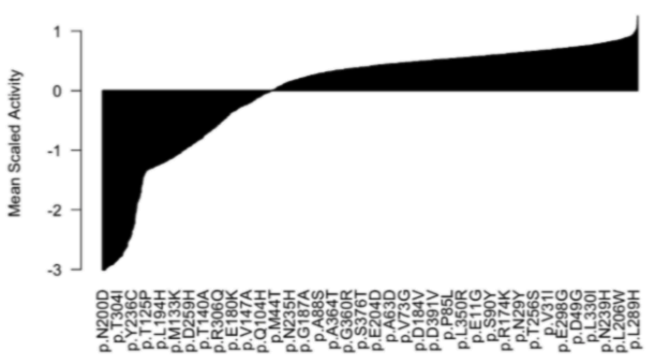

Figure 5. Average Mutation Activity Distribution showcasing the spectrum from extreme function impairment to retained transcriptional functionality

Following log transformation and normalization, Figure 5 shows the average mutation transactivation functionality and displays a hierarchy of TP53 mutations by their impact on the target gene's function. Due to the large set of TP53 mutations in the dataset, only a representative subset is labeled across the x-axis. The specific subset of mutations was selected by ranking observed examples of the low, middle, and high regions in the average scaled activity distribution to represent an accurate range of the dataset.

This figure shows the extreme function impairment, partial activity, and retained transcriptional functionality mutations throughout the dataset. From negative values (low activity) to positive values (high activity), mutations are most frequently observed at positive values, indicating retained transactivation functionality. However, the smaller subset of mutations falls above zero, suggesting increased transcriptional function. This distribution supports the biological heterogeneity of TP53 mutations through the gradual transition in scaled transactivation functionality. These patterns suggest that TP53 mutations exist along a spectrum of functional impact.

Mutations such as p.N200D, C242, and G244 have been observed to exhibit strong loss-of-function effects, with the lowest activity among all observed mutations. p.P322A, p.E221V, and p.L323R suggest high activity values, indicating that they abnormally activate the gene.

CLUSTERING RESULTS

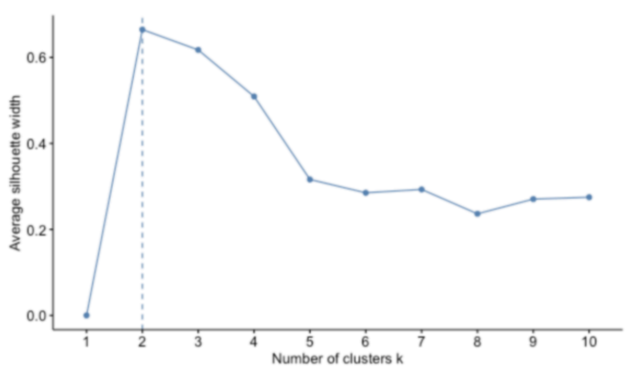

Figure 6. Silhouette Analysis to identify the optimal number of mutation clusters

The silhouette analysis in Figure 6 measures how well-separated and cohesive the TP53 mutation data points are based on gene activity profiles. Although the diagram shows that the highest average silhouette width occurs at k = 2, indicating that two primary clusters provide the most distinct separation of the data, the relatively high silhouette score at k = 3 implies that a third intermediate group may exist.

While this result aligns with the preliminary exploratory data analysis, where mutations appeared to break off into low- and high-activity groups, three clusters better capture the functional spectrum of TP53 mutation behavior. The partial activity mutations from LoF or GoF patterns display a meaningful insight that TP53 mutations often exhibit a spectrum of functional effects rather than restricted binary behavior.

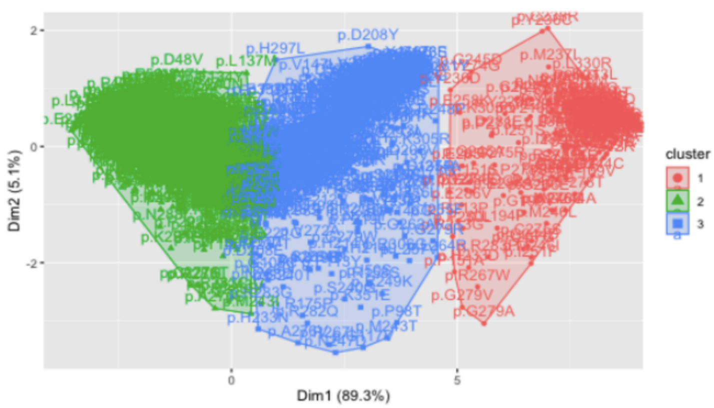

Figure 7. K-Means Clustering revealing distinct functional groups of TP53 mutations

The K-means clustering results in Figure 7 show three distinct clusters that indicate that TP53 mutations are largely differentiated by their collective transcriptional transactivation functionality. Looking back, the visuals from PCA space, where clusters are separated along the first principal component, capture the primary variation in the data. From this comparison, mutation effects follow a structure and are not random.

Upon further analysis, distinct differences in the three clusters of activity levels exhibit one cluster having very low activity across all functional transactivation scores. This cluster's mutations suggest an extremely impaired transactivation activity through consistently low functional transactivation scores across TP53 mutations. These mutations cause irregular tumor-suppressing processes, disrupting the body's ability to properly regulate functions such as apoptosis, DNA repair, and cell cycle arrest. On the opposite extreme, another cluster suggests higher transactivation functionality, suggesting that the mutation still keeps a high level of their original ability to activate gene transcription (Kato et al.). In simpler terms, this cluster may have mutations that keep proteins working just as well as unmutated proteins. In between these two extremes, the intermediate cluster represents mutations with partial transcriptional capability.

The clustering structure in Figure 7, along with the first principal component, reveals that TP53 mutations form functionally relevant groups based on their transcriptional competency and similarities. The tight outer clusters suggest that the mutations either lack or maintain high levels of transactivation activity, while the intermediate cluster shows a more assorted distribution among mutations that maintain partial transactivation functionality. The significant overlap between adjacent clusters reinforces the hypothesis that mutation functionality changes gradually through the dataset of functional scores (Kato et al.).

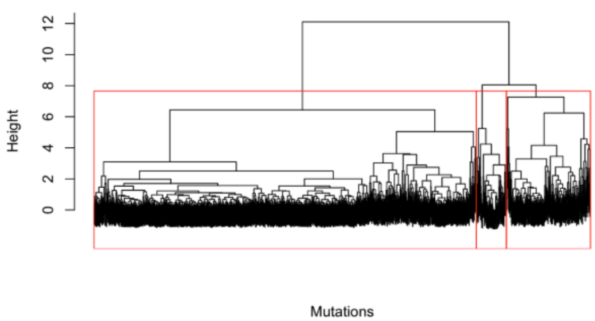

Figure 8. Hierarchical Clustering Dendrogram displaying relationships and functional groupings that share similarities among TP53 mutations

Figure 8 displays a dendrogram that provides a hierarchical structure of additional insights into the clusters of similarity between the TP53 mutations according to their functional transactivation profiles. Each TP53 mutation can be interpreted as a “leaf,” with multiple “branches” that show how closely related the mutations are. Lower branch heights indicate greater similarity, while high heights indicate vast differences. By partitioning the dendrogram based on the distance at which groups of mutations merge, the three clusters are identified in the highlighted red boxes.

The first cluster is relatively large, containing the majority of branches merging at low heights as they connect mutations with highly similar activity patterns. This means that TP53 mutations, for the most part, generally exhibit similar functions, which may be related to lower levels of gene expression or loss of function. On the other hand, two relatively smaller clusters situated on the right side of the tree bifurcate at greater heights, thus demonstrating relatively different functions from one another.

There is more similarity within the smaller groups than within the larger group; therefore, a pattern to the mutations is present, and they are arranged in a hierarchy according to their effects. Consequently, a spectrum of behavior regarding mutations is supported by the clustering method, which is evident by incremental differences in activities.

DISCUSSION & CONCLUSION

The exploratory data analysis and clustering show that the behavior in mutations of TP53 follows functional patterns. Rather than acting as an "off switch" to functional transactivation scores, these mutations present varying biological functions and have been identified to have distinct effects on different groups based on similarity among functional transactivation score patterns.

Based on the skewness and multiple outliers observed across multiple scores, implementing log transformation and normalization was necessary. Without this transformation, clustering distance calculations, variance measurements, and meaningful structure in calculated figures would be disproportionately influenced, resulting in an inaccurate data analysis.

One important finding was that a single TP53 mutation does not affect all downstream target functional transactivation scores equally. The box plots distributions among apoptotic, DNA repair, and oxidative stress scores (specifically NOX, BAX, and p53R2) show direct evidence of disproportionate patterns. These functional transactivation scores' sensitivity implies that TP53 mutations impair the body's ability to repair cellular damage and halt tumor progression, which may contribute to cancerous growths (Aubrey et al.). However, functional transactivation scores with safety functions in the initial stages of tumor development evidently display resistance to the mutation. Functional transactivation scores such as AIP1 and GADD45 demonstrate stable behavior that resists mutational consequences, implying that certain genetic functions are more regulated than others.

The distribution of average mutation activity shows definite contrasts in TP53 mutations' functional impacts. The distribution has most mutations displaying low activity and confirms the hypothesis that loss-of-function mutations are most prevalent (Miller et al.). In contrast, a small subset of mutations shows gain-of-function through exceptionally high activity levels. The study demonstrates that TP53 mutation behavior is a spectrum ranging from loss of function to gain of function.

Correlation and PCA analyses confirmed structured and interactive relationships among the variables in the TP53 dataset. As suggested by the heatmap illustrating correlated interactions, strong positive correlations suggest coordinated regulation across most functional transactivation scores, while PC1 (the first primary component of the PCA) revealed that a majority of the dataset's variations can be attributed to one dominant factor of gene activity. However, the smaller contribution in PC2 (the second primary component of the PCA) indicates that there are more subtle differences between mutations, where their effects might vary depending on certain network-related differences in biological functions like apoptosis and DNA damage repair.

Three clustering analyses displayed direct functional groups among TP53 mutations, confirming the necessity of analyzing the TP53 dataset in clusters. The silhouette analysis indicated that two clusters capture the greatest distance between the various clusters in the data, while both K-means and hierarchical clustering identified three significantly distinct groups. The clusters represent the low, moderate, and high levels of activities and could be considered as loss-of-function, intermediate, and gain-of-function mutations.

The hierarchical clustering demonstrated substantial results, which indicate that these mutation groups form a structured spectrum. According to the dendrogram, mutations with intermediate and high activity appear to be more correlated than those with lower activity, displaying three “branches” and implying that functional impact gradually transitions rather than abruptly.

These findings may significantly improve understanding of TP53 mutation behavior and its clinical application. Biologically, identifying the number of mutation clusters reveals the interconnected yet variable effect TP53 mutations have on multiple networks, while from a clinical perspective, grouping mutations based on their biological functions may improve accurate treatment prediction and mutation classification. Better comprehending the nuances of a patient’s condition through these found clusters can greatly change the course of their care. Different TP53 mutations cause divergent treatment results; creating such a classification system may significantly improve the ability to reliably predict outcomes and explore more effective, personalized treatment options (Sabapathy and Lane).

Despite these strengths, the study has several limitations that should be considered when interpreting results. The analyses conducted throughout are based on a limited set of 8 functional transactivation scores and their transcriptional activity data, which captures only a small subset of their true activity. The clusters may not reflect the full biological range of TP53 mutations. Additionally, mean-scaled activity per mutation was used in some analyses, potentially hiding gene-specific effects and nuanced functional differences. Clustering methods may be sensitive to preprocessing choices like scaling and transformation, as they rely on cluster method hypotheses and algorithmic assumptions. This analysis was purely based on unsupervised computational research. Since the clinical outcomes are missing, the findings have yet to be biologically validated, despite being statistically meaningful.

For future research, incorporating additional clinical data could provide a more reliable view of the nature of the mutation. Furthermore, validating clusters would strengthen clinical datasets' relevance for practical applications.

This study used exploratory data analysis and supervised clustering techniques to examine the behavioral functions of TP53 mutations across various functional transactivation scores. Furthermore, the clusters identified for loss-of-function, neutral, and gain-of-function mutations reveal the existence of biological clusters.

By highlighting the variability of TP53 mutation along with its functional spectrum, the study found that mutations differently affect influential biological processes like apoptosis, DNA repair, and oxidative stress response. Understanding these differences helps provide a more nuanced perspective on how mutations affect cell behavior and cancer growth.

This study reinforces the value of data-driven analysis in understanding biological complexities and the importance of utilizing clustering methods to improve mutation classification. By advancing from traditional groupings, clustering offers more precise interpretation opportunities in genetic data and progress to personalized and targeted cancer treatments.

REFERENCES

Aubrey, BJ, et al. “Tumor-Suppressor Functions of the TP53 Pathway.” *Cold Spring Harbor Perspectives in Medicine*, vol. 6, no. 5, 2016, p. a026062. *PubMed Central*, https://pmc.ncbi.nlm.nih.gov/articles/PMC4852799/. Accessed 6 May 2026.

Cancer Biology Richardson, R. B. “p53 Mutations Associated with Aging-Related Rise in Cancer Incidence Rates.” *Cell Cycle*, vol. 12, no. 15, 2013, pp. 2468–2478. *PubMed Central*, https://pmc.ncbi.nlm.nih.gov/articles/PMC3841325/. Accessed 6 May 2026.

Hassin, Ori, et al. “Different Hotspot p53 Mutants Exert Distinct Phenotypes and Predict Outcome of Colorectal Cancer Patients.” *Nature Communications*, vol. 13, no. 1, 2022, p. 2800, Springer Nature, https://doi.org/10.1038/s41467-022-30481-7. Accessed 13 May 2026

Ikotun, Abiodun M., et al. “K-Means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data.” *Information Sciences*, vol. 622, 2023, pp. 178–210, doi:10.1016/j.ins.2022.11.139.

Jolliffe, Ian T., and Jorge Cadima. “Principal Component Analysis: A Review and Recent Developments.” *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, 2016, doi:10.1098/rsta.2015.0202.

Kato, Shuichi, et al. “Understanding the Function-Structure and Function-Mutation Relationships of p53 Tumor Suppressor Protein by High-Resolution Missense Mutation Analysis.” *Proceedings of the National Academy of Sciences*, vol. 100, no. 14, 2003, pp. 8424–8429. PubMed Entry.

Mantovani, Francesca, et al. "Mutant p53 as a Guardian of the Cancer Cell." *Cell Death & Differentiation*, vol. 26, no. 2, 2019, pp. 199–212. *PubMed Central*, https://pmc.ncbi.nlm.nih.gov/articles/PMC6356290/. Accessed 11 May 2026.

Miller, Jenna J., et al. "The Evolution of TP53 Mutations: From Loss-of-Function to Separation-of-Function Mutants." *Journal of Cancer Biology & Research*, vol. 4, no. 1, 2016, PMC, https://pmc.ncbi.nlm.nih.gov/articles/PMC5298884/. Accessed 6 May 2026.

Murtagh, Fionn, and Pedro Contreras. "Methods of Hierarchical Clustering." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, 2012, pp. 86–97, doi:10.1002/widm.53.

Rousseeuw, Peter J. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics*, vol. 20, 1987, pp. 53–65, doi:10.1016/0377-0427(87)90125-7.

Sabapathy, Kanaga, and David P. Lane. "Gain of Function Mutant p53 in Cancer." *Encyclopedia*, vol. 2, no. 4, 2022, pp. 1957–1970, https://encyclopedia.pub/entry/35066. Accessed 6 May 2026.

"TP53 Gene." *MedlinePlus Genetics*, U.S. National Library of Medicine, https://medlineplus.gov/genetics/gene/tp53/. Accessed 6 May 2026.

Vousden, Karen H., and Carol Prives. "Blinded by the Light: The Growing Complexity of p53." *Cell*, vol. 137, no. 3, 2009, pp. 413–31, https://www.cell.com/fulltext/S0092-8674(09)00944-3. Accessed 11 May 2026.