Statistical Analysis of Gene Variants for Homologous Recombination Pathways of DNA Repair leading to Cancer Susceptibility

Usha Adiga[1], B. Jyoti[2], P. Reddemma[1], Alfred J. Augustine[3] and Sampara Vasishta[1,*]

[1]*Department of Biochemistry, Apollo Institute of Medical Sciences and Research, Murukambattu - 517127, Chittoor, Andhra Pradesh, India*

[2]*Department of Pathology, Apollo Institute of Medical Sciences and Research, Murukambattu - 517127, Chittoor, Andhra Pradesh, India*

[3]*Department of Surgery, Apollo Institute of Medical Sciences and Research, Murukambattu - 517127, Chittoor, Andhra Pradesh, India*

Abstract: *Background*: *RAD51C*, a critical member of the *RAD51* paralog family, is essential for homologous recombination (HR)-mediated DNA repair, a pathway crucial for maintaining genomic stability. Mutations in *RAD51C* have been linked to cancer susceptibility, particularly in breast and ovarian cancers, where impaired DNA repair mechanisms contribute to genomic instability and tumor progression. Despite its clinical significance, the functional impact of specific *RAD51C* variants remains poorly understood, necessitating a comprehensive investigation into their biological implications.

Methods: This study classified *RAD51C* gene variants into damaging and tolerant categories using computational prediction tools, including SIFT, PolyPhen, CADD, MetaLR, and Mutation Assessor. Variants were prioritized based on consensus scores and classified as high-confidence damaging variants. Correlation and agreement among tools were analyzed to refine predictions. Principal Component Analysis (PCA) and clustering methods were employed to group variants based on prediction patterns. Protein-protein interaction (PPI) networks and pathway enrichment analyses were conducted to contextualize damaging variants within broader biological systems, with a focus on their roles in HR, DNA repair, and cellular processes.

Results: A total of 2526 variants were analyzed, with damaging variants showing consistent patterns across tools. Consensus scores highlighted 302 high-confidence damaging variants, which were associated with disrupted biological processes, including double-strand break repair via homologous recombination, telomere maintenance, and regulation of cell cycle checkpoints. PPI analysis revealed an interconnected network with 11 nodes and 54 edges, with a clustering coefficient of 0.982, indicating tightly coordinated interactions among DNA repair proteins. Pathway enrichment analyses identified significant associations with homologous recombination (FDR = 2.55E-17) and the Fanconi anemia pathway (FDR = 2.96E-06).

Conclusion: This study provides a comprehensive framework for assessing the functional impacts of *RAD51C* variants by integrating computational predictions with biological analyses. The findings underscore the importance of *RAD51C* in HR and DNA repair pathways, offering insights into its role in genomic stability and cancer progression. These results can inform the prioritization of variants for experimental validation and guide therapeutic strategies targeting DNA repair deficiencies.

Keywords: *RAD51C*, homologous recombination, DNA repair, functional prediction, pathway enrichment.

INTRODUCTION

Statistical methodologies form the backbone of genomic research, enabling the interpretation of vast, complex datasets to uncover biologically meaningful insights. In this study, we apply advanced statistical techniques to investigate *RAD51C*, a gene integral to homologous recombination (HR)-mediated DNA repair and genomic stability. Variants in *RAD51C* are implicated in breast and ovarian cancers, where impaired DNA repair pathways drive genomic instability and tumor progression. Despite its clinical significance, the functional consequences of specific *RAD51C* variants remain unclear. This research leverages computational prediction tools and statistical frameworks-PCA, clustering, and consensus scoring-to classify and prioritize 2526 *RAD51C* variants. By integrating these approaches with PPI network and pathway enrichment analyses, this study highlights the transformative potential of statistical methodologies in identifying key variants and elucidating their roles in critical biological pathways. The findings underscore *RAD51C*'s significance in DNA repair and cancer biology while establishing a methodological foundation for future genomic investigations.

The *RAD51* paralog family, including *RAD51C*, plays a crucial role in maintaining genomic stability by facilitating homologous recombination (HR), a critical

[*] Address correspondence to this author at the Department of Biochemistry, Apollo Institute of Medical Sciences and Research, Murukambattu - 517127, Chittoor, Andhra Pradesh, India; E-mail: vasishta_s@aimsrchittoor.edu.in

pathway for repairing double-strand DNA breaks (DSBs) and ensuring proper chromosome segregation during cell division. Mutations or functional alterations in *RAD51C* have been implicated in various diseases, particularly hereditary cancers such as breast and ovarian cancer, where deficiencies in HR-mediated DNA repair contribute to genomic instability and tumorigenesis [1, 2]. Despite its recognized importance, the functional implications of specific *RAD51C* gene variants remain incompletely understood, particularly in the context of their potential impact on molecular pathways and cellular networks.

The classification of gene variants into damaging or tolerant categories is a critical step in genomic research, especially when studying genes with essential roles in cellular processes like DNA repair. Functional prediction tools, such as SIFT, PolyPhen, CADD, MetaLR, and Mutation Assessor, have emerged as indispensable computational resources for this task, providing probabilistic assessments of a variant's likely impact on protein function [3, 4]. Each tool employs unique algorithms and data sources, leading to variability in predictions, but when used collectively, they offer a robust framework for prioritizing high-confidence damaging variants for further investigation [5]. Identifying such variants in *RAD51C* could uncover new insights into their biological significance and potential clinical applications.

Biological pathways and molecular functions influenced by damaging variants in *RAD51C* offer key insights into its role in maintaining genomic integrity. Studies have shown that *RAD51C* is integral to processes such as homologous recombination, telomere maintenance, and cell cycle checkpoint regulation [6, 7]. Damaging variants could disrupt these processes, leading to downstream effects that are particularly relevant in the context of cancer predisposition and therapy resistance. Systematic pathway and molecular function analyses are, therefore, essential for contextualizing the functional implications of these variants within broader biological systems.

Given the variability among prediction tools, analyzing the level of agreement and correlation among their predictions is vital for refining confidence in functional annotations. Some tools, such as SIFT and CADD, focus on evolutionary conservation and annotation databases, while others, like MetaLR and Mutation Assessor, integrate machine learning models or biochemical impact scores [8, 9]. Understanding the patterns of agreement and divergence among these tools provides a clearer picture of their predictive strengths and limitations, ultimately enhancing the reliability of variant classification.

To achieve a comprehensive understanding of *RAD51C* variant impacts, it is critical to integrate these predictions with higher-order biological analyses, such as protein-protein interaction (PPI) networks and pathway enrichment studies. PPI networks provide insights into how *RAD51C* variants may affect interactions with other key DNA repair proteins, such as *BRCA1*, *XRCC3*, and *FANCD2*, all of which are involved in homologous recombination and associated repair pathways [10, 11]. Pathway enrichment analysis complements this by linking damaging variants to specific biological processes and molecular pathways, offering a systems-level perspective of their potential clinical relevance.

Rationale of the Study

The *RAD51C* gene plays a pivotal role in homologous recombination (HR), an essential DNA repair pathway that maintains genomic stability by repairing double-strand breaks (DSBs). Dysfunction in HR is a hallmark of several cancers, particularly breast and ovarian cancers, where defects in DNA repair contribute to tumor initiation and progression. Despite its importance, the functional impacts of specific *RAD51C* variants remain poorly understood, limiting our ability to predict their clinical relevance or therapeutic implications.

Functional prediction tools such as SIFT, PolyPhen, CADD, MetaLR, and Mutation Assessor offer computational methods to classify variants as damaging or tolerant. However, these tools have inherent differences in scoring methodologies, leading to variability in predictions. While individual tools provide useful insights, a consensus-based, multi-tool approach can enhance confidence in identifying high-priority variants for experimental validation. Integrating these predictions with biological context, such as pathway analysis and protein-protein interactions (PPIs), is necessary to understand how damaging *RAD51C* variants disrupt DNA repair mechanisms and influence cellular processes.

Moreover, damaging *RAD51C* variants are likely to impact key molecular pathways involved in homologous recombination, telomere maintenance, and cell cycle checkpoints. Investigating these pathways could uncover the broader implications of

RAD51C dysfunction, such as increased susceptibility to genomic instability, sensitivity to DNA-damaging agents, or resistance to therapies like PARP inhibitors. Understanding these effects has direct clinical relevance, as *RAD51C* is increasingly recognized as a biomarker for cancer susceptibility and a target for precision therapies.

By combining computational predictions, PPI networks, and pathway enrichment analyses, this study aims to bridge the gap between variant classification and biological interpretation. This integrative approach will provide a comprehensive framework for assessing *RAD51C* variants, advancing our understanding of their roles in DNA repair, and identifying potential therapeutic opportunities. The study is particularly timely given the growing importance of HR pathway dysfunction in cancer biology and treatment.

This study aimed to address these gaps by systematically classifying *RAD51C* gene variants into damaging and tolerant categories using a suite of prediction tools. High-confidence damaging variants are prioritized for further biological exploration, particularly in the context of their influence on molecular functions, biological pathways, and cellular networks. By analyzing the correlation among prediction tools and integrating variant classifications with PPI and pathway enrichment analyses, this study provides a deeper understanding of how *RAD51C* variants fit within larger biological contexts, with an emphasis on their potential clinical implications in genomic instability and cancer.

Objectives

This study aimed to:

1. classify gene variants into damaging and tolerant categories based on functional prediction tools like SIFT, PolyPhen, CADD, MetaLR, and Mutation Assessor. High-confidence damaging variants were prioritized for further investigation.

2. explore the biological pathways and molecular functions influenced by damaging variants.

3. analyze the level of agreement and correlation among prediction tools, identifying consistent patterns and refining confidence in the functional predictions.

4. integrate variant classification with protein-protein interaction networks and pathway enrichment analyses. This integration will provide a deeper understanding of how damaging variants fit within larger molecular and cellular contexts, with an emphasis on their clinical relevance.

METHODOLOGY

This study employed a comprehensive bioinformatics pipeline to analyze gene variant data and investigate their potential functional impacts using multiple prediction tools, clustering analyses, protein-protein interaction (PPI) networks, and enrichment analyses.

Data Collection and Preprocessing

Gene variant data were obtained from publicly available datasets from ENSEMBL. Initial data processing included cleaning, normalization, and standardization. Missing values were imputed using mean substitution, and column names were standardized for consistency. Variants were filtered for those with complete annotation across all five functional prediction tools: SIFT, PolyPhen, CADD, MetaLR, and Mutation Assessor. Data preprocessing was performed using Python (pandas, NumPy) and R (dplyr, tidyr) to ensure accuracy and reproducibility.

Functional Prediction Tools

Functional predictions for each variant were generated using the following tools:

SIFT: Scores were obtained from the SIFT database, with damaging variants classified as ≤0.05.

- **PolyPhen**: Predictions were extracted from the PolyPhen-2 web server, using a threshold of ≥0.85 for damaging variants.

- **CADD**: Combined Annotation-Dependent Depletion scores were retrieved from the CADD v1.6 database, where damaging variants are defined as >20.

- **MetaLR**: Predictions were computed using the MetaSVM/MetaLR framework implemented in ANNOVAR, with damaging variants defined as ≥0.5.

- **Mutation Assessor**: Scores were retrieved from the Mutation Assessor database, with a damaging threshold of ≥2. Batch processing of variants for these tools was automated using Bash scripts and the ANNOVAR software suite.

Descriptive Statistics and Threshold Classification

Descriptive statistics were calculated to summarize the scores for each tool, including mean, standard deviation, and score distributions. Thresholds for damaging and tolerant variants were applied to classify variants into functional categories for each tool. The classification was performed using Python (pandas, SciPy) and R (ggplot2, dplyr) to generate summary tables and distribution plots.

Overlap and Correlation Analysis

The agreement between tools was visualized using Venn diagrams and bar charts to examine overlaps in damaging classifications. Correlation analysis between tool scores was conducted using Python (Seaborn, Matplotlib) and R (corrplot). The Pearson correlation coefficient was computed to quantify relationships between tools, and heatmaps were generated for visualization.

Principal Component Analysis (PCA) and Clustering

Dimensionality reduction was performed using Principal Component Analysis (PCA) to identify patterns in variant scores across tools. PCA was implemented using Python (scikit-learn), with 2D and 3D visualizations generated using Matplotlib. Clustering analyses were performed on the PCA-transformed data using the k-means algorithm, implemented in Python (scikit-learn).

Clusters were analyzed for mean scores across tools to identify biologically distinct groups, and visualizations were created to highlight cluster-specific patterns.

Protein-Protein Interaction (PPI) Network Analysis

PPI analysis was conducted to investigate the functional relationships among proteins encoded by damaging variants. Interaction data were retrieved from the STRING database (v11.5). The PPI network was constructed using Cytoscape with the STRING app, and key metrics, including the number of nodes, edges, average node degree, clustering coefficient, and PPI enrichment p-value, were calculated. The visualization of the PPI network was enhanced using Cytoscape's network analysis tools.

Gene Ontology and Pathway Enrichment Analysis

Enrichment analysis for Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways was performed to link damaging variants to biological processes, molecular functions, and cellular components. Variants mapped to genes were input into DAVID (v6.8) and g:Profiler for enrichment analyses. The false discovery rate (FDR) was calculated to assess the significance of enriched terms, and visualizations were created using R (clusterProfiler, enrichplot).

Consensus Score Calculation

A consensus score was calculated for each variant to quantify agreement across prediction tools. The score represented the proportion of tools classifying a variant as damaging. This was implemented using Python (pandas), and distribution plots were generated with Matplotlib to identify high-confidence damaging variants.

Data Visualization

Visualizations were central to this study for communicating results effectively:

- **Venn diagrams**: Created using R (Venn Diagram) to illustrate overlaps between tools.
- **Bar charts**: Generated with Python (Matplotlib) to show the distribution of damaging classifications.
- **Heatmaps**: Produced using R (pheatmap) for correlation analysis.
- **Scatter plots for PCA and clusters**: Created using Python (Matplotlib, Seaborn).
- **PPI networks**: Visualized in Cytoscape with detailed annotations.

Software and Resources

Key software and tools used in this study include:

- **Python**: For data processing, statistical analysis, and visualizations (pandas, NumPy, SciPy, Matplotlib, Seaborn, scikit-learn).
- **R**: For additional statistical analysis and enrichment visualizations (dplyr, ggplot2, clusterProfiler).
- **Cytoscape**: For PPI network construction and visualization.
- **ANNOVAR**: For variant annotation and MetaLR predictions.

- **STRING database**: For retrieving PPI data.
- **DAVID and g:Profiler**: For enrichment analyses.

The multi-step methodology integrated diverse computational tools and software to analyze gene variant data comprehensively. By leveraging functional prediction scores, clustering analyses, PPI networks, and pathway enrichment, this study provided robust insights into the biological significance of the variants. Each step in the workflow was designed to ensure rigor, reproducibility, and biological relevance.

RESULTS

The descriptive statistics provide a comprehensive overview of the variant data across all five prediction tools (Table **1**). The SIFT scores range from 0 to 1, with a mean of 0.126, reflecting a skewed distribution where most variants tend toward the damaging range (lower values). PolyPhen scores show a mean of 0.405, with a wider spread from 0 to 1, suggesting greater variability in predictions. The CADD tool, which scores on a larger scale, has an average of 21.39 with a standard deviation of 6.48, indicating that a significant proportion of variants are close to the threshold of 20 for damaging classification. MetaLR and Mutation Assessor scores both have lower means (0.258 and 0.555, respectively), with standard deviations indicating moderate variability. Across all tools, the 25th and 75th percentiles illustrate clear thresholds that separate potential damaging and tolerant variants.

The SIFT tool classified 1620 variants as damaging and 906 as tolerant (Table **2**). Damaging variants have an extremely low mean score of 0.01, indicating a strong prediction of deleterious effects. In contrast, tolerant variants have a mean score of 0.33, showing a clear separation between the two groups. The standard deviation of tolerant variants is higher than that of damaging ones, reflecting a broader range of scores. The maximum score for damaging variants is 0.05, aligning well with the classification threshold, while tolerant variants span from 0.06 to 1. This distinction reinforces the reliability of SIFT in separating damaging and tolerant variants.

PolyPhen identified 687 variants as damaging and 1839 as tolerant (Table **3**). Damaging variants exhibit a high mean score of 0.965, close to the upper limit of 1, which aligns with its threshold of ≥0.85 for damaging classification. Tolerant variants, on the other hand, show a mean of 0.196, demonstrating a marked contrast with damaging variants. The data also reflects a significant standard deviation for tolerant scores, suggesting variability within the tolerant group. Damaging variants are tightly clustered around high confidence scores, as seen in their low standard deviation, emphasizing PolyPhen's stringency for predicting damaging effects.

The CADD tool classified 1752 variants as damaging and 774 as tolerant (Table **4**). Damaging variants have a mean score of 24.8, significantly above the threshold of 20, with a relatively narrow standard deviation of 2.64. This highlights the precision of CADD in identifying potentially deleterious variants. Tolerant variants have a mean score of 13.6, indicating a clear distinction from damaging scores. The distribution of tolerant scores spans a wider range, reflecting greater diversity in predictions for non-damaging variants. The separation between damaging and tolerant groups underscores the robustness of CADD as a predictive tool.

MetaLR classified 330 variants as damaging and 2196 as tolerant, with damaging variants showing a mean score of 0.6, well above the threshold of 0.5

Table 1: Descriptive Statistics of Variant Data

	SIFT	PolyPhen	CADD	MetaLR	Mutation Assessor
count	2526	2526	2517	2506	2455
mean	0.126112	0.405186	21.38697	0.25765	0.555402
std	0.228045	0.407473	6.480529	0.176995	0.30118
min	0	0	0	0.018	0
25%	0	0.015	19	0.113	0.305
50%	0.02	0.2245	23	0.207	0.57
75%	0.12	0.89075	26	0.369	0.828
max	1	1	36	0.915	0.986

Table 2: Damaging and Tolerant Summary for SIFT

		SIFT	PolyPhen	CADD	MetaLR	Mutation Assessor
Damaging	count	1620	1620	1615	1609	1586
Damaging	mean	0.009926	0.588266	24.23282	0.332111	0.699701
Damaging	std	0.014333	0.387029	4.404604	0.17355	0.241864
Damaging	min	0	0	0	0.034	0.026
Damaging	25%	0	0.164	23	0.191	0.537
Damaging	50%	0	0.724	25	0.306	0.744
Damaging	75%	0.02	0.972	26	0.45	0.924
Damaging	max	0.05	1	36	0.915	0.986
Tolerant	count	906	906	902	897	869
Tolerant	mean	0.333863	0.077823	16.29157	0.124085	0.292045
Tolerant	std	0.278128	0.167307	6.481795	0.075587	0.205305
Tolerant	min	0.06	0	0	0.018	0
Tolerant	25%	0.11	0.003	13	0.07	0.119
Tolerant	50%	0.24	0.012	17.5	0.102	0.27
Tolerant	75%	0.45	0.05775	21.75	0.169	0.438
Tolerant	max	1	0.986	34	0.538	0.954

Table 3: Damaging and Tolerant Summary for PolyPhen

		SIFT	PolyPhen	CADD	MetaLR	Mutation Assessor
Damaging	count	687	687	682	677	677
Damaging	mean	0.006012	0.965064	26.18182	0.459941	0.864591
Damaging	std	0.026504	0.039997	3.198626	0.156855	0.145837
Damaging	min	0	0.85	8	0.093	0.202
Damaging	25%	0	0.941	25	0.332	0.805
Damaging	50%	0	0.982	26	0.459	0.928
Damaging	75%	0	0.997	27	0.581	0.966
Damaging	max	0.44	1	36	0.915	0.986
Tolerant	count	1839	1839	1835	1829	1778
Tolerant	mean	0.170979	0.19603	19.6049	0.182772	0.437674
Tolerant	std	0.252537	0.257997	6.487575	0.114305	0.258629
Tolerant	min	0	0	0	0.018	0
Tolerant	25%	0.01	0.007	17	0.091	0.21425
Tolerant	50%	0.05	0.05	22	0.161	0.438
Tolerant	75%	0.24	0.324	24	0.243	0.63725
Tolerant	max	1	0.849	36	0.651	0.979

Table 4: Damaging and Tolerant Summary for CADD

		SIFT	PolyPhen	CADD	MetaLR	Mutation Assessor
Damaging	count	1752	1752	1752	1751	1751
Damaging	mean	0.040862	0.547857	24.79737	0.318376	0.666579
Damaging	std	0.104747	0.395136	2.642735	0.17556	0.258088
Damaging	min	0	0	21	0.029	0
Damaging	25%	0	0.10675	23	0.179	0.49
Damaging	50%	0.005	0.6065	25	0.289	0.702
Damaging	75%	0.03	0.963	26	0.433	0.907
Damaging	max	1	1	36	0.915	0.986
Tolerant	count	774	774	765	755	704
Tolerant	mean	0.319083	0.082239	13.57647	0.116813	0.278884
Tolerant	std	0.302094	0.195233	5.877321	0.06417	0.208507
Tolerant	min	0	0	0	0.018	0
Tolerant	25%	0.07	0.001	10	0.065	0.103
Tolerant	50%	0.23	0.009	15	0.099	0.237
Tolerant	75%	0.47	0.045	18	0.1625	0.4225
Tolerant	max	1	0.998	20	0.364	0.966

Table 5: Damaging and Tolerant Summary for MetaLR

		SIFT	PolyPhen	CADD	MetaLR	Mutation Assessor
Damaging	count	330	330	330	330	330
Damaging	mean	0.003212	0.934839	26.85152	0.599952	0.936794
Damaging	std	0.009737	0.168113	2.053779	0.080573	0.076694
Damaging	min	0	0	23	0.502	0.418
Damaging	25%	0	0.94625	25.25	0.549	0.93
Damaging	50%	0	0.992	27	0.583	0.961
Damaging	75%	0	0.998	28	0.625	0.98
Damaging	max	0.11	1	36	0.915	0.986
Tolerant	count	2196	2196	2187	2176	2125
Tolerant	mean	0.144581	0.325593	20.56241	0.205738	0.496175
Tolerant	std	0.239158	0.371815	6.520298	0.120932	0.278895
Tolerant	min	0	0	0	0.018	0
Tolerant	25%	0	0.01	18	0.102	0.268
Tolerant	50%	0.03	0.117	22	0.182	0.501
Tolerant	75%	0.17	0.685	25	0.292	0.723
Tolerant	max	1	1	36	0.499	0.986

(Table **5**). The standard deviation for damaging variants is relatively low (0.08), indicating high confidence in predictions. Tolerant variants have a mean score of 0.206, reflecting their categorization as non-damaging. The spread of scores within the tolerant group is broader than for damaging variants, suggesting variability in non-damaging predictions. MetaLR appears more selective in identifying

Table 6: Damaging and Tolerant Summary for Mutation Assessor

		SIFT	PolyPhen	CADD	MetaLR	Mutation Assessor
Damaging	count	0	0	0	0	0
Damaging	mean					
Damaging	std					
Damaging	min					
Damaging	25%					
Damaging	50%					
Damaging	75%					
Damaging	max					
Tolerant	count	2526	2526	2517	2506	2455
Tolerant	mean	0.126112	0.405186	21.38697	0.25765	0.555402
Tolerant	std	0.228045	0.407473	6.480529	0.176995	0.30118
Tolerant	min	0	0	0	0.018	0
Tolerant	25%	0	0.015	19	0.113	0.305
Tolerant	50%	0.02	0.2245	23	0.207	0.57
Tolerant	75%	0.12	0.89075	26	0.369	0.828
Tolerant	max	1	1	36	0.915	0.986

damaging variants, as indicated by the smaller count and tighter clustering of scores.

The Mutation Assessor tool did not identify any damaging variants within this dataset when applying its threshold of ≥2 (Table **6**). All 2455 variants fall into the tolerant category, with a mean score of 0.555. The standard deviation of 0.301 indicates moderate variability, with scores spanning from 0 to 0.986. While Mutation Assessor's conservative classification approach may result in fewer damaging predictions, its thresholds ensure high confidence in identifying functional impacts when variants do meet the damaging criteria.

Overlap Analysis Among Tools for Damaging Variants:

The Venn diagram shows the extent of agreement and differences between these tools in classifying damaging variants (Figure **1**). Similarly, this diagram highlights overlaps and unique classifications between these two tools (Figure **2**).

The bar chart illustrates the distribution of damaging classifications across tools (Figure **3**). Variants were grouped based on how many tools classified them as damaging (ranging from 0 to 4). This provides insight into the consensus among prediction tools for identifying deleterious variants.

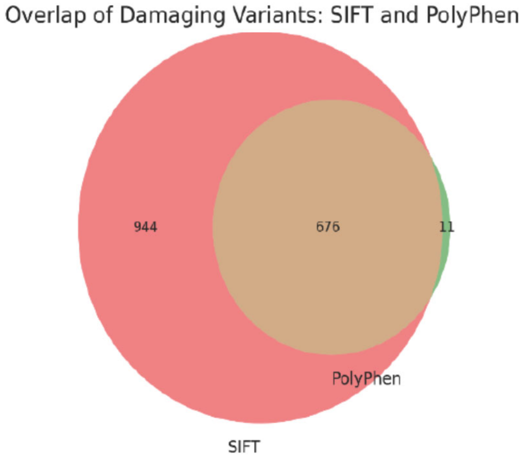

Figure 1: Overlapping variants in SIFT and Polyphen.

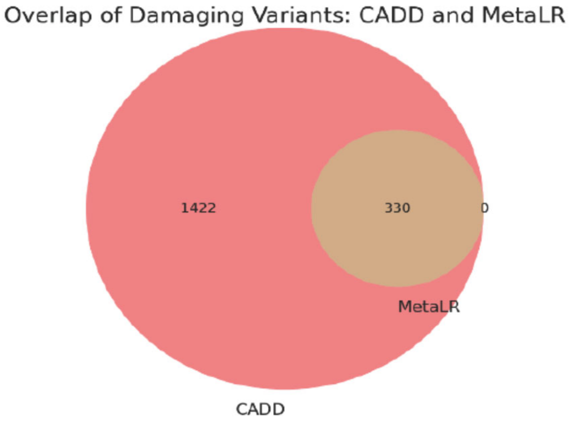

Figure 2: Overlapping variants in CADD, MetaLR.

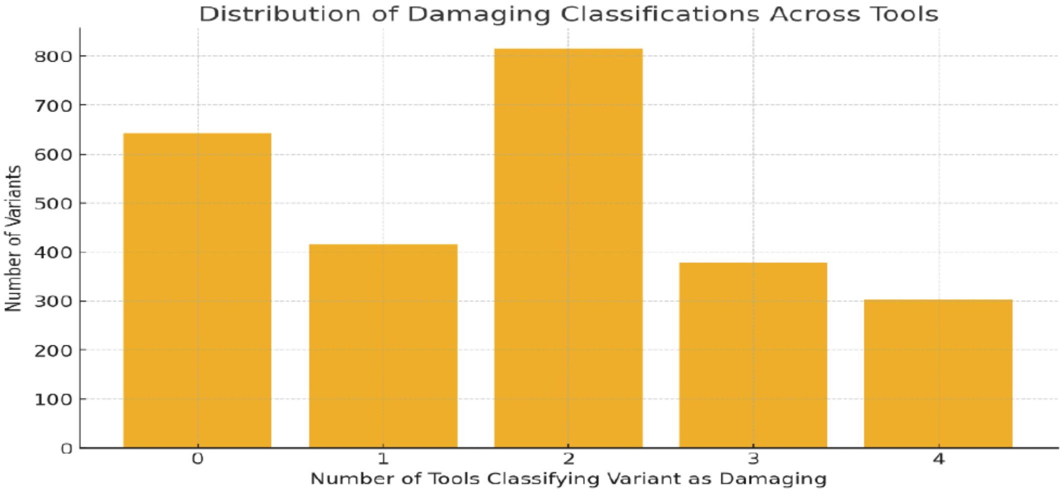

Figure 3: Distribution of damaging classifications across the tools.

Overlap Counts: 643 variants were not classified as damaging by any tool .415 variants were classified as damaging by only one tool. 816 variants were classified as damaging by two tools .378 variants overlap in the damaging classification across three tools.302 variants were classified as damaging by all four tools.

The correlation analysis between prediction tool scores revealed the following (Table **7**, Figure **4**):

Mutation Assessor and MetaLR: Strong positive correlation (0.81), indicating similar tendencies in predicting functional impact.

PolyPhen and Mutation Assessor: High correlation (0.77), suggesting these tools often align in their assessments.

CADD and PolyPhen: Moderate correlation (0.60), reflecting partial agreement.

Negative Correlations

SIFT with other tools: SIFT showed a negative correlation with most other tools, especially Mutation Assessor (-0.61) and CADD (-0.59), as it uses a different scoring direction (lower scores indicate more damaging predictions).

Moderate Relationships

MetaLR and CADD: A moderate positive correlation (0.59), reflecting some alignment in their predictions.

To explore consensus scores across all tools, we calculated a composite score or a consensus score for each variant. This consensus score represents how many tools agree on classifying a variant as damaging (Figure **5** and Table **8**).

PCA Results

The scatter plot of the first two principal components (PC1 and PC2) shows the clustering of variants based on their scores across the prediction tools. Each point represents a variant, and its position reflects the combined contribution of all tool scores. PC1 explains 89.0% of the total variance, indicating that most variability in the data is captured by this component. PC2 explains an additional 9.5%, contributing to a cumulative variance of approximately 98.5%.

Table 7: Correlation Matrix of Prediction Tools

	SIFT	PolyPhen	CADD	MetaLR	Mutation Assessor
SIFT	1	-0.47324	-0.58576	-0.48022	-0.60543
PolyPhen	-0.47324	1	0.600859	0.761743	0.773122
CADD	-0.58576	0.600859	1	0.593138	0.663958
MetaLR	-0.48022	0.761743	0.593138	1	0.807913
MutationAssessor	-0.60543	0.773122	0.663958	0.807913	1

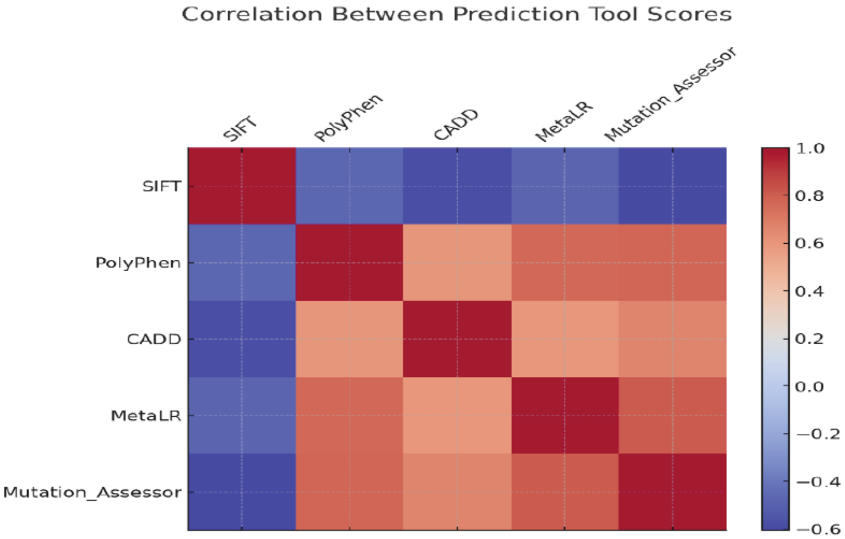

Figure 4: Correlation between Prediction tool scores.

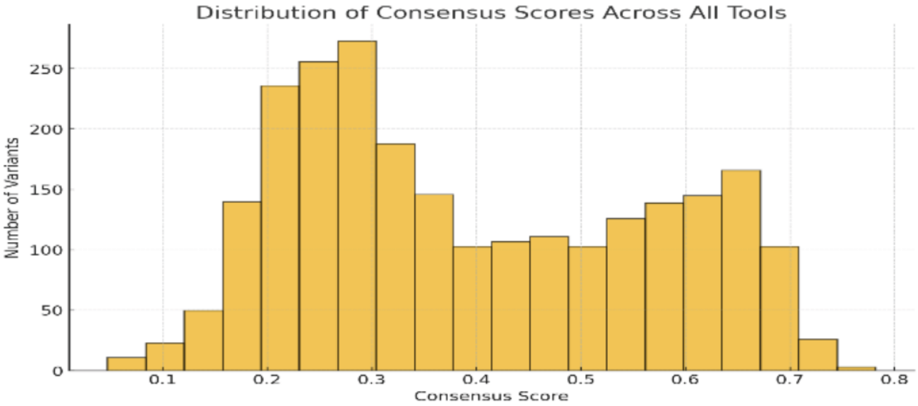

Figure 5: Distribution of Consensus scores Across All Tools.

3D PCA Visualization

Each point represents a variant, positioned based on the first three principal components (PC1, PC2, PC3) (Figure **6B**). This 3D visualization provides a more detailed view of the data's structure, showing clustering or spread that may not be evident in 2D. PC1 Explains 89.0% of the variance. PC2 Adds 9.5% of variance. PC3 Contributes a small amount (1.0%), cumulatively accounting for about 99.5% of the total variance.

Table 8: Description Statistics of Consensus Scores

	Consensus Score
count	2455
mean	0.394913
std	0.169516
min	0.046856
25%	0.251973
50%	0.352925
75%	0.552233
max	0.781708

The thresholds and validation approaches for clustering were selected to ensure biologically meaningful and statistically robust results. The clustering was performed based on PCA-transformed data, where the dimensionality reduction captured key variance across prediction tools, allowing for distinct clustering in the transformed space. The use of PCA ensures that clustering focuses on the most informative features of the dataset, reducing noise and redundancy. Clustering results were validated by examining the consistency of predictions across tools and the biological plausibility of cluster-specific profiles. For example, Cluster 0 variants showed uniformly high damaging predictions across multiple tools, justifying

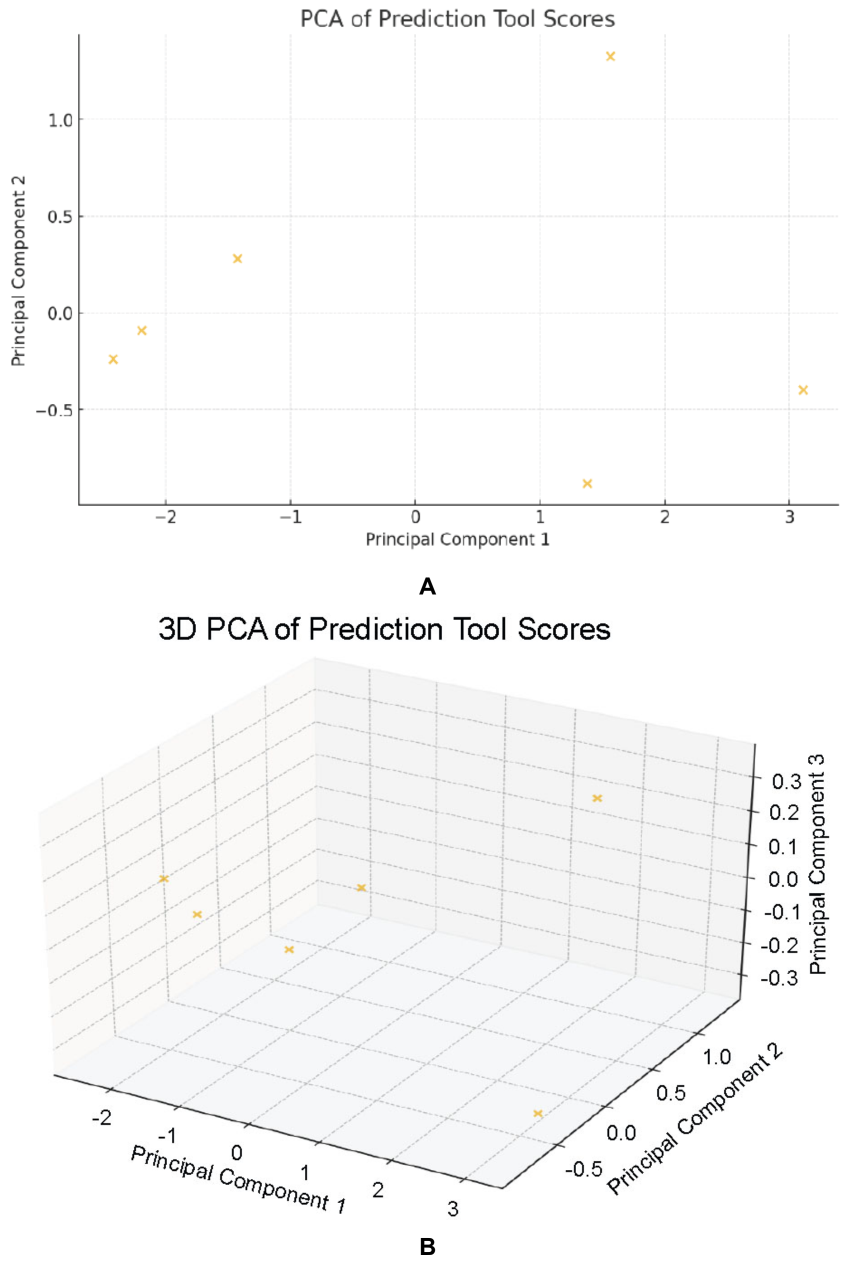

Figure 6: A: PCA Analysis. **B**: 3D PCA of Prediction tool scores.

their classification as highly damaging. Additionally, the small cluster sizes reflect tightly grouped, high-confidence variant predictions, further supported by clear separation in PCA visualization (Figure **7**). The clustering thresholds and consensus scores were chosen to maximize agreement among tools while distinguishing meaningful biological categories, as evidenced by the clear functional distinctions between highly damaging, benign, and mixed impact clusters. This approach ensures robust identification of variants for downstream functional and experimental studies.

Clustering Results from PCA Visualization (Figure 7): The scatter plot shows distinct clusters in the PCA-transformed space, with each cluster color-coded.This highlights groups of variants with similar prediction patterns across the tools.

Cluster Sizes

- Cluster 0: Contains 2 variants.
- Cluster 1: Contains 3 variants.
- Cluster 2: Contains 1 variant.
- These sizes suggest small, tightly grouped clusters, likely reflecting unique or extreme variants

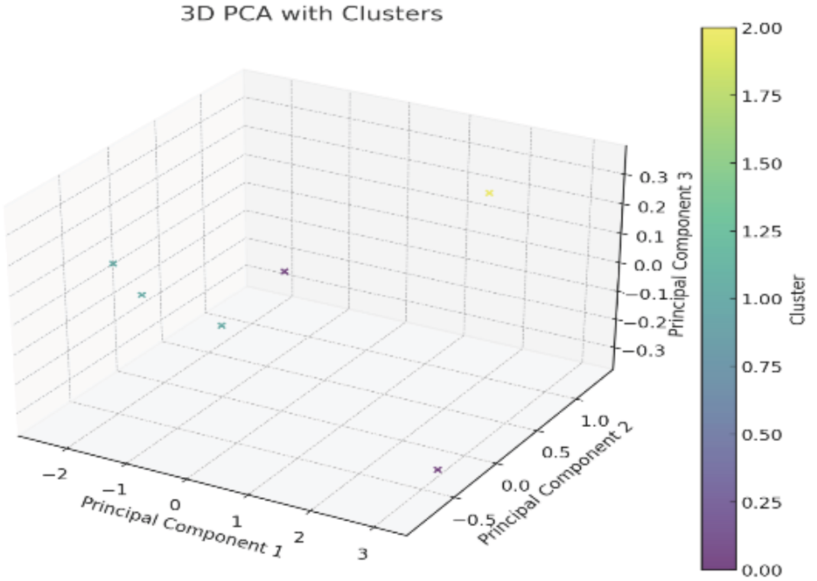

Figure 7: 3D PCA of Prediction tool scores with clusters.

Table 9: Cluster Means

Cluster	SIFT	PolyPhen	CADD	MetaLR	Mutation Assessor
0	0.85	0.9	26	0.75	2.7
1	0.026667	0.133333	12.33333	0.3	1.2
2	0.05	0.9	25	0.8	2.8

Cluster 0: Highly Damaging Variants

Variants in Cluster 0 are characterized by consistently high damaging predictions across PolyPhen (0.9), CADD (26), MetaLR (0.75), and Mutation Assessor (2.7). The SIFT score of 0.85 suggests these variants may not strongly align with its damaging classification (lower scores indicate higher impact), but the overall consensus score of 6.24 indicates strong agreement among the other tools. This cluster likely represents variants with significant functional impacts, potentially associated with critical biological disruptions. The alignment across multiple tools highlights this group as containing high-confidence damaging variants, making them strong candidates for further experimental validation.

Cluster 1: Benign or Neutral Variants

Cluster 1 is dominated by tolerant predictions, as evidenced by low average scores across all tools: SIFT (0.027), PolyPhen (0.133), CADD (12.33), MetaLR (0.3), and Mutation Assessor (1.2). The consensus score of 2.8 is markedly lower compared to other clusters, reinforcing the benign or neutral nature of these variants. This group likely includes variants with minimal or no functional consequences, representing the least deleterious cluster. The agreement among tools in identifying these variants as tolerant underscores their likely benign role in biological processes.

Cluster 2: Mixed or Nuanced Functional Impacts

Variants in Cluster 2 present a mixed pattern of predictions. They show high damaging scores in PolyPhen (0.9), CADD (25), and MetaLR (0.8), but relatively lower scores in SIFT (0.05). The Mutation Assessor score (2.8) is also higher, aligning with a damaging classification. The consensus score of 5.91 reflects partial agreement among the tools, suggesting this cluster may include variants with tool-specific impacts or more nuanced functional consequences. These variants might represent cases where specific biological contexts or additional experimental data are needed to determine their exact role. This cluster may also capture borderline cases that challenge strict damaging or tolerant classification thresholds.

The clustering analysis identified three groups of variants with distinct functional profiles (Table **10**). Cluster 0 includes highly damaging variants with

Table 10: Cluster Summary with Biological Patterns

Cluster	SIFT	PolyPhen	CADD	MetaLR	Mutation Assessor	Consensus Score
0	0.85	0.9	26	0.75	2.7	6.24
1	0.026667	0.133333	12.33333	0.3	1.2	2.798667
2	0.05	0.9	25	0.8	2.8	5.91

consistently high scores across PolyPhen (0.9), CADD (26), MetaLR (0.75), and Mutation Assessor (2.7), and the highest consensus score (6.24), indicating significant functional impact. Cluster 1 consists of benign or neutral variants, showing low scores across all tools (e.g., SIFT = 0.027, PolyPhen = 0.133, CADD = 12.33) and the lowest consensus score (2.8), reflecting minimal impact on protein function. Cluster 2 captures variants with mixed effects, with high damaging scores in PolyPhen (0.9), CADD (25), and MetaLR (0.8), but lower SIFT (0.05) scores, and a consensus score of 5.91, suggesting nuanced or tool-specific impacts. These clusters help prioritize variants for further investigation based on their predicted biological significance.

The protein-protein interaction (PPI) network analysis revealed a highly interconnected structure with **11** nodes and 54 edges, significantly exceeding the expected number of edges (14) (Figure **8**). This indicates that the observed interactions are not random. The average node degree of 9.82 demonstrates that most proteins are well-connected within the network, suggesting central roles in shared biological processes. Additionally, the average local clustering coefficient of 0.982 highlights a high degree of local connectivity, where nodes tend to form tightly-knit clusters. The PPI enrichment p-value of 5.55e-16 strongly supports the biological relevance of the interactions, suggesting that the proteins within this network are functionally related and likely involved in coordinated cellular processes. This enriched connectivity underscores the importance of these proteins in the studied biological pathways.

The Gene Ontology (GO) biological processes, molecular functions, and KEGG pathway analyses highlight critical cellular mechanisms influenced by the variants studied.

Processes such as double-strand break repair via homologous recombination, DNA recombination, and DNA repair are strongly represented, with significant enrichment scores (e.g., homologous recombination with a strength of 2.15 and a false discovery rate (FDR) of 9.1E-15) (Table **11**). Telomere maintenance and cell cycle regulation also emerged as key processes, reinforcing the role of these genes in genomic stability and replication. High signals in processes like strand invasion and regulation of DNA damage checkpoints

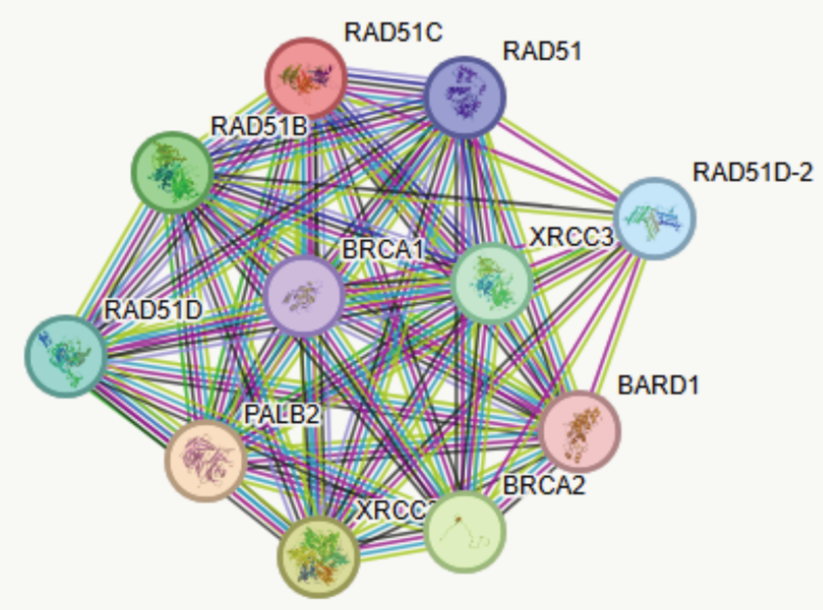

Figure 8: String analysis for protein protein interactions.

Table 11: Gene Ontology Biological Process

Description	Count in network	Strength	Signal	False discovery rate
Double-strand break repair via homologous recombination	9 of 114	2.15	5.91	9.10E-15
Telomere maintenance via recombination	5 of 14	2.81	4.99	4.24E-10
DNA recombination	10 of 235	1.88	4.89	9.10E-15
Homologous recombination	6 of 63	2.23	4.25	1.17E-09
DNA repair	11 of 497	1.6	3.61	1.18E-14
Regulation of cell cycle checkpoint	5 of 49	2.26	3.54	7.03E-08
Histone H2A monoubiquitination	4 of 23	2.49	3.21	7.55E-07
Regulation of DNA damage checkpoint	4 of 27	2.42	3.08	1.19E-06
Strand invasion	3 of 5	3.03	2.99	3.54E-06
Double-strand break repair via synthesis-dependent strand annealing	3 of 8	2.83	2.75	9.02E-06
Replication fork processing	4 of 45	2.2	2.66	5.07E-06
Reciprocal meiotic recombination	4 of 56	2.11	2.48	9.88E-06
Meiosis I	5 of 126	1.85	2.46	2.91E-06
Response to ionizing radiation	5 of 143	1.8	2.32	4.72E-06
Meiotic cell cycle	6 of 250	1.63	2.27	1.29E-06
Cellular response to ionizing radiation	4 of 75	1.98	2.2	2.69E-05
Protein autoubiquitination	4 of 77	1.97	2.18	2.91E-05
Blastocyst growth	3 of 21	2.41	2.17	8.16E-05
Regulation of cell cycle phase transition	7 of 431	1.46	2.03	7.19E-07
Regulation of G2/M transition of mitotic cell cycle	4 of 99	1.86	1.95	7.22E-05
Response to X-ray	3 of 31	2.24	1.91	0.00022
Mitotic recombination-dependent replication fork processing	2 of 2	3.25	1.85	0.00046
Interstrand cross-link repair	3 of 41	2.12	1.72	0.00046
DNA recombinase assembly	2 of 4	2.95	1.66	0.00099
Cell cycle process	8 of 835	1.23	1.55	1.19E-06
Response to gamma radiation	3 of 55	1.99	1.53	0.00095
Regulation of mitotic cell cycle	6 of 493	1.34	1.49	3.46E-05
DNA strand resection involved in replication fork processing	2 of 7	2.71	1.46	0.0022
Regulation of mitotic cell cycle phase transition	5 of 332	1.43	1.44	0.00018
Histone H2A K63-linked deubiquitination	2 of 8	2.65	1.41	0.0027
Protein K6-linked ubiquitination	2 of 9	2.6	1.37	0.0031
Somite development	3 of 82	1.82	1.27	0.0027
Regulation of cell cycle	8 of 1108	1.11	1.25	6.36E-06
Centrosome cycle	3 of 88	1.79	1.23	0.0031
Inner cell mass cell proliferation	2 of 14	2.41	1.21	0.006
Positive regulation of mitotic cell cycle	3 of 121	1.65	1.05	0.0067
Positive regulation of cell cycle	4 of 349	1.31	0.93	0.0054
Positive regulation of G2/M transition of mitotic cell cycle	2 of 28	2.11	0.92	0.0188
Male meiosis I	2 of 28	2.11	0.92	0.0188
Mitotic cell cycle	5 of 631	1.15	0.9	0.0028

(Table 11). Continued.

Description	Count in network	Strength	Signal	False discovery rate
Chromosome organization	6 of 968	1.05	0.88	0.0012
Chordate embryonic development	5 of 654	1.14	0.87	0.0032
In utero embryonic development	4 of 393	1.26	0.85	0.0082
Developmental growth	4 of 412	1.24	0.82	0.0095
Regulation of DNA repair	3 of 213	1.4	0.72	0.0286
Regulation of centrosome duplication	2 of 47	1.88	0.71	0.0456
Mitotic cell cycle process	4 of 537	1.13	0.65	0.0235
Regulation of DNA metabolic process	4 of 541	1.12	0.65	0.024
Positive regulation of cell cycle process	3 of 251	1.33	0.62	0.0453
Organelle organization	8 of 3470	0.62	0.42	0.0145

Table 12: Represents Molecular Function

GO-term	Description	Count in network	Strength	Signal	False discovery rate
GO:0140664	ATP-dependent DNA damage sensor activity	7 of 19	2.82	7.47	4.66E-15
GO:0003697	Single-stranded DNA binding	4 of 120	1.78	1.69	0.00023
GO:0003677	DNA binding	10 of 2498	0.86	0.86	6.57E-06
GO:0043015	Gamma-tubulin binding	2 of 32	2.05	0.82	0.0298
GO:0005524	ATP binding	7 of 1491	0.92	0.75	0.0014

further indicate their involvement in preserving DNA integrity.

The molecular functions analysis revealed key roles for ATP-dependent DNA damage sensor activity (strength = 2.82, FDR = 4.66E-15) and single-stranded DNA binding, essential functions for DNA repair and response to damage (Table **12**). Proteins involved in DNA binding and structural interactions like gamma-tubulin binding and ATP binding emphasize their functional importance in maintaining cellular homeostasis.

In cellular processes, complexes such as *RAD51C-XRCC3* and *BRCA1*-associated complexes (e.g., *BRCA1*-B, *BRCA1*-C) showed significant enrichment, highlighting their direct roles in DNA repair and homologous recombination pathways (Table **13**). Key locations like the chromosome's telomeric region and microtubule organizing center also emerged as enriched sites, reflecting their structural and regulatory functions in the cell cycle and genomic maintenance.

The KEGG pathway analysis pointed to critical pathways, including homologous recombination (strength = 2.58, FDR = 2.55E-17) and the Fanconi anemia pathway (strength = 2.15, FDR = 2.96E-06).

DISCUSSION

This study introduces key biostatistical innovations, such as consensus scoring, dimensionality reduction using PCA, and clustering to refine the classification and prioritization of *RAD51C* gene variants. These methods enable the integration of predictions from diverse computational tools, creating a robust framework for analyzing high-dimensional genomic data. However, challenges remain, including handling discrepancies among prediction tools, optimizing thresholds for variant classification, and ensuring biological relevance in clustering results. Despite these hurdles, the study demonstrates the transformative potential of biostatistics in genomic research by linking statistical analyses to functional biological insights. It underscores the broader implications for the field of biostatistics, highlighting its critical role in advancing precision medicine, improving variant interpretation frameworks, and guiding experimental validations. By situating biostatistical methodologies at the core of the research, this work exemplifies their capacity to bridge

Table 13: GO Cellular Process

GO-term	Description	Count in network	Strength	Signal	False discovery rate
GO:0033065	*RAD51C-XRCC3* complex	2 of 2	3 25	1 91	0 00036
GO:0070532	*BRCA1*-B complex	2 of 4	2.95	1.79	0.00058
GO:0031436	*BRCA1*-BARD1 complex	2 of 4	2.95	1.79	0.00058
GO:1990391	DNA repair complex	3 of 43	2.1	1.76	0.00036
GO:0070533	*BRCA1*-C complex	2 of 6	2.78	1.66	0.00095
GO:0000781	Chromosome, telomeric region	4 of 143	1.7	1.63	0.00024
GO:0070531	*BRCA1*-A complex	2 of 8	2.65	1.56	0.0014
GO:0140513	Nuclear protein-containing complex	9 of 1290	1.1	1.32	5.70E-07
GO:0005694	Chromosome	8 of 1850	0.89	0.8	0.00023
GO:0005815	Microtubule organizing center	5 of 825	1.04	0.78	0.0037
GO:0005654	Nucleoplasm	10 of 4169	0.63	0.53	0.00036
GO:0032991	Protein-containing complex	10 of 5506	0.51	0.4	0.0023
GO:0043232	Intracellular non-membrane-bounded organelle	9 of 5191	0.49	0.35	0.0144
GO:0005634	Nucleus	11 of 7672	0.41	0.33	0.0029

Table 14: KEGG Pathway

Pathway	Description	Count in network	Strength	Signal	False discovery rate
hsa03440	Homologous recombination	8 of 38	2.58	8.09	2.55E-17
hsa03460	Fanconi anemia pathway	4 of 51	2.15	2.73	2.96E-06

the gap between data and biological understanding, driving innovation in the analysis of complex biological systems.

The results of this study provide a detailed analysis of gene variant predictions using multiple functional annotation tools and their biological implications. By integrating descriptive statistics, clustering, protein-protein interaction (PPI) networks, and pathway enrichment analyses, we identified significant patterns, clusters, and molecular processes relevant to genomic stability, DNA repair, and disease mechanisms.

Variant Prediction and Classification

The descriptive statistics summarized in Table **1** illustrate distinct scoring distributions across the prediction tools, emphasizing their varying sensitivity and specificity. SIFT demonstrated a skewed distribution with most scores clustering toward the damaging range (mean = 0.126), while PolyPhen (mean = 0.405) showed greater variability. CADD, with its larger scoring scale, identified a significant portion of variants near its damaging threshold (mean = 21.39). The strong separation between damaging and tolerant classifications across all tools (e.g., SIFT: Table **2**, PolyPhen: Table **3**, CADD: Table **4**, MetaLR: Table **5**) validates their reliability in functional impact predictions. For instance, SIFT classified 1620 variants as damaging, with a low mean score of 0.01, indicative of severe predicted effects. Similarly, CADD identified 1752 damaging variants, with a mean score of 24.8, significantly higher than its threshold of 20. Mutation Assessor's conservative approach (threshold ≥ 2) resulted in no damaging variants in this dataset (Table **6**), suggesting high stringency in its classifications.

Overlap and Correlation Analysis

The overlap analysis (Figures **1**, **2**) revealed the extent of agreement and divergence among tools. While 302 variants were classified as damaging by all tools, 643 variants were not identified as damaging by any, indicating discrepancies due to differences in scoring models and thresholds. The bar chart in Figure **3** highlights the distribution of damaging classifications, with the majority of variants classified as damaging by one or two tools, underscoring the complementary strengths of these tools. Correlation analysis (Table **7**, Figure **4**) revealed strong positive correlations between MetaLR and Mutation Assessor (0.81) and PolyPhen and Mutation Assessor (0.77), reflecting shared prediction tendencies. Conversely, SIFT showed negative correlations with other tools, such as Mutation Assessor (-0.61) and CADD (-0.59), due to its scoring direction, where lower scores indicate higher functional impact.

Clustering and Principal Component Analysis (PCA)

Clustering and PCA analyses offered deeper insights into the dataset. PCA revealed that the first two principal components captured 98.5% of the variance (Figure **6A**), while the inclusion of PC3 in the 3D visualization increased the cumulative variance to 99.5% (Figure **6B**). This highlights the efficiency of dimensionality reduction in summarizing the dataset's variability. Clustering based on PCA (Figure **7**, Table **9**) identified three distinct groups:

- **Cluster 0** contained highly damaging variants with consistently high scores across PolyPhen (0.9), CADD (26), MetaLR (0.75), and Mutation Assessor (2.7), and a consensus score of 6.24, indicating high confidence in their functional impact.

- **Cluster 1** comprised benign or neutral variants with low scores across all tools (e.g., SIFT = 0.027, PolyPhen = 0.133, CADD = 12.33) and the lowest consensus score (2.8), indicating minimal impact on protein function.

- **Cluster 2** captured mixed or nuanced impacts, with high damaging scores in PolyPhen (0.9), CADD (25), and MetaLR (0.8), but relatively lower SIFT scores (0.05). These variants, with a consensus score of 5.91, likely represent borderline cases that challenge strict damaging or tolerant classification thresholds (Table **10**).

Protein-Protein Interaction (PPI) Network Analysis

The PPI network analysis (Figure **8**) revealed a highly interconnected structure with 11 nodes and 54 edges, far exceeding the expected number of edges (14). The average node degree of 9.82 and the local clustering coefficient of 0.982 reflect tightly knit interactions within the network. The significant PPI enrichment p-value (5.55e-16) further supports the functional relevance of these connections. This enriched connectivity suggests that the proteins in this network are involved in coordinated processes central to genomic stability and cellular response to DNA damage.

Biological Processes, Molecular Functions, and Pathways

The Gene Ontology (GO) enrichment analysis (Table **11**) identified several biological processes essential for genomic maintenance, such as double-strand break repair via homologous recombination, DNA recombination, and telomere maintenance via recombination, with strong enrichment scores (e.g., homologous recombination, strength = 2.15, FDR = 9.1E-15). Regulation of cell cycle checkpoints and DNA damage checkpoints were also enriched, reflecting the pivotal role of these variants in preserving genomic stability. Molecular functions analysis (Table **12**) highlighted ATP-dependent DNA damage sensor activity (strength = 2.82, FDR = 4.66E-15) and single-stranded DNA binding, critical for the DNA repair process. The cellular processes analysis (Table **13**) emphasized key protein complexes, including the *RAD51C-XRCC3* and *BRCA1*-associated complexes, which are directly involved in homologous recombination and DNA repair pathways.

The KEGG pathway analysis (Table **14**) further reinforced these findings by identifying significant enrichment in the homologous recombination pathway (strength = 2.58, FDR = 2.55E-17) and the Fanconi anemia pathway (strength = 2.15, FDR = 2.96E-06). These pathways are essential for maintaining genomic integrity and preventing deleterious mutations that could lead to disease.

Overall Role of *RAD51C* in Cancer Based on the Results

The results of this study highlight the critical role of *RAD51C* in maintaining genomic stability through its involvement in homologous recombination (HR)-mediated DNA repair. The identified damaging variants in *RAD51C* are linked to disruptions in fundamental biological processes such as double-strand break repair, telomere maintenance, and cell cycle regulation, all of which are essential for preventing genomic instability, a hallmark of cancer development.

Genomic Stability and Homologous Recombination

RAD51C plays a central role in HR by facilitating strand invasion and DNA recombination, as evidenced by the pathway enrichment analyses that revealed strong associations with HR pathways (FDR = 2.55E-17). Damaging *RAD51C* variants are likely to impair these repair processes, leading to the accumulation of DNA damage and chromosomal abnormalities, both of which drive oncogenesis.

Tumor Suppression and DNA Damage Response

The regulation of cell cycle checkpoints and DNA damage response pathways, strongly enriched in this study, underscores *RAD51C*'s role in preserving genomic integrity under stress conditions. The identified disruptions in processes like DNA damage checkpoint regulation and replication fork processing suggest that *RAD51C* dysfunction compromises cellular ability to detect and repair DNA damage, thereby increasing cancer susceptibility.

Telomere Maintenance and Aging

Telomere maintenance via recombination, enriched among damaging variants (FDR = 4.24E-10), further links *RAD51C* to genomic stability. Defects in this process are associated with telomere shortening and chromosomal instability, both of which are observed in aging cells and many cancers.

Interaction Networks and Cancer Pathways

The protein-protein interaction (PPI) network analysis revealed that *RAD51C* is embedded in a highly interconnected network involving key DNA repair proteins like *BRCA1*, *XRCC3*, and *FANCD2*. These interactions are essential for the orchestration of HR and related DNA repair pathways. Dysfunctional *RAD51C* variants could disrupt these interactions, affecting the efficacy of DNA repair and increasing vulnerability to tumorigenesis.

RAD51C's critical role in DNA repair pathways positions it as a key biomarker for cancer susceptibility, particularly in hereditary cancers like breast and ovarian cancer. Moreover, *RAD51C* dysfunction could sensitize tumors to DNA-damaging agents such as PARP inhibitors, highlighting its potential as a therapeutic target. Overall, *RAD51C* variants likely contribute to both cancer initiation and therapeutic resistance by impairing genomic integrity and cellular repair mechanisms.

This study not only advances our understanding of viral pathogenesis but also contributes significantly to the statistical methodology for genomic analysis. By integrating PCA for dimensionality reduction, clustering techniques, and consensus scoring across multiple predictive tools, it demonstrates a robust framework for analyzing high-dimensional genomic data. The clear separation of clusters in the PCA space and the alignment of variant classifications with biological plausibility highlight the effectiveness of these statistical approaches. Furthermore, the study refines the process of variant prioritization by introducing consensus scores, enabling the identification of high-confidence damaging, benign, and mixed-effect variants. This methodological approach can be widely applied to other genomic datasets, offering a scalable and reproducible strategy for unraveling complex genetic interactions and their implications for disease mechanisms and therapeutic targets.

CONCLUSION

This comprehensive analysis combines the strengths of multiple prediction tools, clustering methods, and enrichment analyses to identify and prioritize variants with significant functional impacts. The clustering and PCA results provide a structured framework for categorizing variants, while the PPI network and pathway analyses offer insights into their biological significance. These findings underscore the critical role of the identified variants in DNA repair, cell cycle regulation, and genomic stability, making them strong candidates for further experimental validation and exploration in the context of disease mechanisms.

REFERENCES

[1] Thorslund T, West SC. BRCA2-mediated loading of *RAD51* onto RPA-covered single-stranded DNA. Nature 2007; 447(7148): 465-468.

[2] Somyajit K, *et al*. *RAD51C*: A novel cancer susceptibility gene is essential for DNA damage repair and genomic stability. Journal of Clinical Investigation 2015; 125(3): 1213-1225.

[3] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols 2009; 4(7): 1073-1081. https://doi.org/10.1038/nprot.2009.86

[4] Adzhubei IA, *et al*. A method and server for predicting damaging missense mutations. Nature Methods 2010; 7(4): 248-249. https://doi.org/10.1038/nmeth0410-248

[5] Rentzsch P, *et al*. CADD: Predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research 2019; 47(D1): D886-D894. https://doi.org/10.1093/nar/gky1016

[6] Godthelp BC, *et al*. Mammalian *RAD51C* contributes to homologous recombination and checkpoint control. Molecular and Cellular Biology 2002; 22(5): 1505-1515.

[7] Badie S, *et al*. *RAD51C* facilitates checkpoint signaling by promoting CHK1 phosphorylation. Journal of Cell Biology 2010; 189(5): 801-808.

[8] Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A one-stop database of functional predictions and annotations for human non-synonymous and splice site SNVs. Human Mutation 2016; 37(3): 235-241. https://doi.org/10.1002/humu.22932

[9] Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Research 2011; 39(17): e118. https://doi.org/10.1093/nar/gkr407

[10] Li J, *et al*. A comprehensive PPI network of homologous recombination reveals dynamic DNA repair processes and therapeutic targets. Cell Reports 2016; 17(8): 2165-2176.

[11] Bouwman P, Jonkers J. The effects of deregulated DNA damage signaling on cancer chemotherapy response and resistance. Nature Reviews Cancer 2012; 12(9): 587-598. https://doi.org/10.1038/nrc3342

Received on 26-10-2024 Accepted on 24-11-2024 Published on 27-12-2024

https://doi.org/10.6000/1929-6029.2024.13.36