Identification and Characterization of Rare Pathogenic Gene*tic Variants Associated with Breast Cancer Susceptibility in 1000 Genome and gnomAD Population*

Kanishk Yadav[1*], Abhishek Prasad[2]

1. Bioinformatics Program, Department of Mathematical Sciences, University of Texas at El Paso, El Paso, Texas - 79968
2. Department of Anesthesiology and Peri-Operative Medicine, MD Anderson Cancer Center, Houston, Texas - 77030

*Corresponding Author: Kanishk Yadav *(email: kanishkyadav07@gmail.com)*

Keywords: Hereditary breast cancer; Germline variants; Breast cancer susceptibility genes; BRCA1; BRCA2; ClinVar; gnomAD; 1000 Genomes Project; Pathogenic variants; In silico pathogenicity prediction

Abstract:

Background: Inherited variants in cancer susceptibility genes play a critical role in breast cancer risk, yet comprehensive comparative analyses across clinically relevant genes remain limited. While *BRCA1* and *BRCA2* are well-established high-risk contributors, the broader landscape of germline pathogenicity across breast cancer associated genes warrants systematic evaluation.
Methods: We curated a panel of 15 breast cancer associated genes (*BRCA1*, *BRCA2*, *TP53*, *PTEN*, *CHEK2*, *ATM*, *CDH1*, *PALB2*, *LSP1*, *MAP3K1*, *NF1*, *RECQL*, *TOX3*, *FANCD2*, *RAD51C*) based on clinical and epidemiological relevance. Publicly available germline variant data from gnomAD v4.1 and the 1000 Genomes Project were integrated with ClinVar annotations. Variants were classified by clinical significance and filtered using allele frequency and phenotype annotation to stratify into high-risk and non-high-risk groups. We compared the distribution of variants and assessed penetrance patterns, and we evaluated the discriminatory power of computational pathogenicity predictors including CADD, REVEL, SIFT, and PolyPhen-2.
Results: *BRCA1* and *BRCA2* harbored the highest burden of pathogenic variants, consistent with their high penetrance status. Moderate-risk genes (*PALB2*, *RAD51C*, *CHEK2*) showed intermediate levels of pathogenicity, while low-risk genes such as *RECQL*, *TOX3*, and *FANCD2* were predominantly enriched for variants of uncertain significance. High-risk variants exhibited significantly elevated CADD (>30), REVEL (>0.75), and SIFT (<0.05) scores compared to non-high-risk variants ($p < 1\times10^{-2}$), whereas PolyPhen-2 showed limited discriminatory power. Odds ratio analysis further supported CADD, REVEL, and ClinVar pathogenicity labels as strong discriminators between variant categories.
Conclusions: Our analysis reaffirms the pathogenic burden concentrated in high penetrance genes and highlights the utility of combined in silico prediction tools and population frequency filters in variant prioritization. These findings support refined variant curation pipelines for hereditary breast cancer and provide a framework for interpreting gene-level pathogenicity across diverse susceptibility loci.

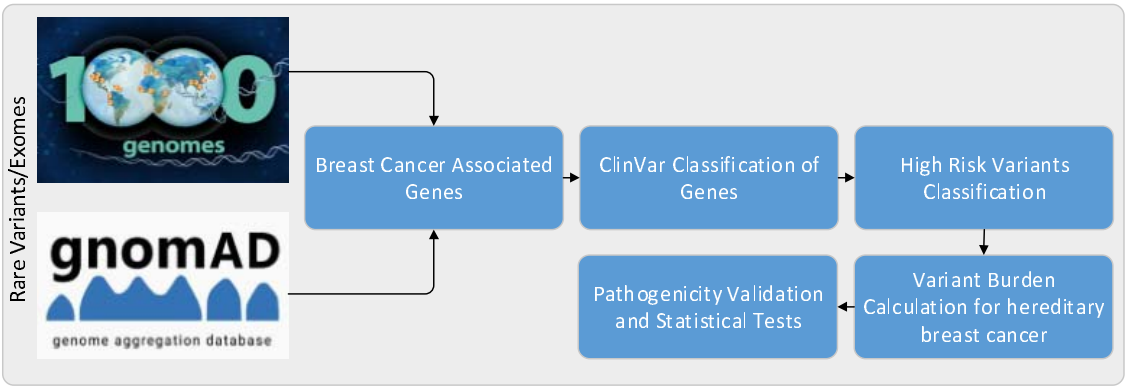
Introduction

Breast cancer is the most frequently diagnosed cancer in women worldwide and a leading cause of female cancer related mortality. In 2020, there were approximately 2.3 million new cases and over 670,000 deaths attributed to breast cancer globally, illustrating the significant burden of this disease [1]. Although the majority of breast cancers are sporadic, 5–10% are hereditary, and an even larger proportion may involve inherited susceptibility in familial clusters. In certain populations, up to 15–25% of breast cancer cases may be associated with germline mutations in cancer predisposition genes [2,27]. Historically, *BRCA1* and *BRCA2* have been the most studied genes, with pathogenic variants conferring lifetime breast cancer risks as high as 65% for *BRCA1* and 57% for *BRCA2* [6,7]. Other high penetrance genes, such as *TP53*, *PTEN*, and *CDH1*, are associated with hereditary cancer syndromes like Li-Fraumeni, Cowden, and hereditary diffuse gastric cancer [8–10].

Beyond these, moderate risk genes such as *PALB2*, *CHEK2*, *ATM*, and *RAD51C* also play clinically significant roles in genetic testing panels and contribute additional, though smaller, risk burdens [11–14]. The ClinVar database has become an essential open access resource for clinical variant interpretation, consolidating expert submissions, laboratory assertions, and supporting evidence [1,2]. However, interpretation challenges persist, variants of uncertain significance (VUS) still make up a large proportion of classified variants in cancer susceptibility genes [1,5,9]. Although ClinGen continues to refine gene and variant classification frameworks through its Variant and Gene Curation Expert Panels [23,41], no single universally accepted breast cancer gene panel exists for research standardization. Instead, researchers rely on curated sources such as the National Cancer Institute's PDQ Genetics summary and systematic reviews of genetic contributions to breast cancer to define working gene sets [27,28].

In this study, we leveraged population scale germline variant resources from gnomAD v4.1 and the 1000 Genomes Project [24,29] to extract and annotate variants across 15 breast cancer associated genes frequently referenced in clinical and research settings: *BRCA1*, *BRCA2*, *TP53*, *PTEN*, *CHEK2*, *ATM*, *CDH1*, *PALB2*, *LSP1*, *MAP3K1*, *NF1*, *RECQL*, *TOX3*, *FANCD2*, and *RAD51C* [2,27,28]. We then harmonized this data with ClinVar to evaluate the distribution of clinical significance classifications, stratified variants into high risk and non high risk categories based on frequency and annotation thresholds, and compared the effectiveness of in silico pathogenicity predictors such as CADD, REVEL, SIFT, and PolyPhen-2 in distinguishing variant classes [18–22].

Materials and Methods



Flowchart 1. *Overview of the analytical pipeline used to identify and evaluate high-risk germline variants associated with hereditary breast cancer*. Variant data from the 1000 Genomes Project and gnomAD v4 were used to extract rare single-nucleotide variants (SNVs) within a curated set of breast cancer susceptibility genes. These variants were annotated using ClinVar to classify their clinical significance, followed by identification of high-risk variants based on allele frequency, disease relevance, and predicted pathogenicity. Subsequent statistical validation and burden analyses were performed to quantify variant impact across genes.

Data sources and gene set

We aggregated germline single-nucleotide variants (SNVs) for breast-cancer associated genes from public resources: 1000 Genomes Project and gnomAD v4 (GRCh38) (accessed Aug 2025) [24,29]. Gene selection was guided by a recent peer-reviewed review and the NCI PDQ professional guideline, yielding the following set: *BRCA1*, *BRCA2*, *TP53*, *PTEN*, *CHEK2*, *ATM*, *CDH1*, *PALB2*, *LSP1*, *MAP3K1*, *NF1*, *RECQL*, *TOX3*, *FANCD2*, *RAD51C* [27,28]. Gene symbols followed HGNC standards to ensure consistent nomenclature across datasets [30]. To reduce potential errors, we verified gene lists against Ensembl BioMart and RefSeq identifiers, ensuring that gene boundaries and canonical transcripts were consistent with current reference annotations. This cross-check minimized misassignment of variants to paralogous loci, which can occur in large-scale databases.(Flowchart 1)

Variant harmonization and ClinVar Labeling

All variants were left-joined to ClinVar to obtain clinical significance (CLNSIG) labels (Pathogenic [P], Likely Pathogenic [LP], P/LP, VUS, Likely Benign [LB], Benign/Likely Benign [B/LB], Benign [B]) using dbSNP rsIDs (avsnp151) as the stable key [1,2,26]. Variants lacking an rsID (avsnp151 = ".") were excluded from Figure 1 to maintain uniqueness and cross-database comparability [26]. When the same rsID/gene appeared with discordant CLNSIG labels (e.g., both P and LP), we resolved to a single most-specific/most-severe label using the deterministic priority P > LP > P/LP > VUS > LB > B/LB > B (implemented by sorting on a numeric priority and retaining the first record). This rule

privileges definitive calls (P) over ambiguous or benign classifications and avoids double-counting. Additional harmonization included collapsing redundant ClinVar submissions under the same rsID, thereby reducing inflation from multiple submitter entries.

Counting unique variants per gene and CLNSIG class

After harmonization and de-duplication by (rsID, gene), we computed unique rsID counts per gene within each CLNSIG class. Counts reflect the number of distinct dbSNP identifiers in a given gene/CLNSIG bin, not alleles or submissions. This yields a gene-by-class matrix used to visualize the distribution of reported variation across 15 genes and 7 ClinVar categories. To ensure robustness, we also verified that the same variant did not map to multiple genes due to overlapping annotations, and any such ambiguous cases were excluded. The resulting dataset captures the breadth of known germline variation across clinically relevant breast cancer genes, emphasizing pathogenic and likely pathogenic variants while contextualizing them within the full spectrum of ClinVar labels.

Defining the high-risk variants with heavy variant burden

We defined the high-risk candidate set 'Cases' that simultaneously met three criteria: (i) rarity, operationalized as gnomAD exome allele frequency ≤1% [24,25]; (ii) disease relevance, requiring the ClinVar Diagnosis term (CLNDN) to contain "Hereditary_breast" (e.g., "Hereditary_breast_and_ovarian_cancer") [1]; and (iii) predicted deleteriousness, with CADD PHRED ≥30 to enrich for highly damaging changes [18,19]. (Thresholds for other predictors such as REVEL were not used to define Cases for Figure 2 and were evaluated separately in Figure 3.) To quantify burden, we first deduplicated records at the (rsID, gene) level, because a single variant can have multiple submissions/labels—by retaining one clinical label using a deterministic severity priority: P > P/LP > LP > VUS > LB > B/LB > B [1,2]. We then constructed, for each (gene, CLNSIG) bin, a numerator equal to the count of unique rsIDs among Cases ($\boldsymbol{N_{\{cases\}}}$) and a denominator equal to the count of all unique rsIDs after deduplication ($\boldsymbol{N_{\{total\}}}$). The "variant risk burden" was computed as 100 × (numerator/denominator); missing cells were imputed as 0. This measure reflects the proportion of catalogued variants within each gene/class that satisfy our rare, disease-linked, high-deleteriousness filter and is not a clinical penetrance estimate.

$$\boldsymbol{Variant\ Risk\ Burden} = \frac{\boldsymbol{N_{\{cases\}}}}{\boldsymbol{N_{\{total\}}}} \times \mathbf{100}$$

Gene ordering and visualization

Genes were assigned to high, moderate, or low penetrance classes using prior risk estimates [3–5]. For display, genes were ordered within each class by descending Pathogenic (P) burden, then VUS, then Benign (B) to emphasize clinically actionable content. We produced a stacked bar plot showing B, VUS, and P percentages per gene with dashed separators and labels for the three penetrance strata. Axes were labeled "Breast Cancer Associated Genes" (x) and "Variant Risk Burden (%)" (y) (Figure 2). To improve interpretability, we further highlighted the most frequently mutated genes within each penetrance category, providing a visual emphasis on loci of particular clinical and biological relevance. The choice of stacked percentages (rather than absolute counts) allows for fairer comparison across genes with

different total variant loads. Color schemes were selected to align with conventional clinical classification (e.g., red for pathogenic and moving down to blue for benign), thereby facilitating intuitive interpretation for both research and clinical audiences.

Rationale and QC

Using AF ≤ 1% focuses on rare alleles that are more likely pathogenic in Mendelian cancer genes [24,25]. A stringent CADD ≥ 30 enriches for variants with strong predicted functional impact [18,19]. Severity-based deduplication using dbSNP rsIDs ensures one vote per variant per gene and reduces bias from multiple submissions [1,2,26]. In addition, we excluded synonymous and intronic variants outside canonical splice sites to reduce noise from variants unlikely to alter protein function. QC steps also included verification of gene-variant mappings against multiple annotation sources (ANNOVAR, ClinVar, and gnomAD) to minimize misclassification errors. Where multiple isoforms existed, we prioritized canonical transcripts as defined by Ensembl for consistency across genes. Variant counts were cross-checked by independent scripts to ensure reproducibility of the burden profiles presented in Figure 2. All steps were performed in Python with pandas/NumPy for reproducibility [31–33]. Computational notebooks and code were version-controlled, with random seeds fixed where applicable, ensuring complete reproducibility and transparency of analyses.

Validation of the pathogenicity contributing factors in the for high-risk vs. non high-risk burden variants

To compare computational pathogenicity signal between high-risk and non-high-risk variant groups, we partitioned the severity-prioritized, de-duplicated rsID–gene set and, for each group, computed the proportion of unique variants exceeding established cut-points for widely used predictors—CADD PHRED >30, REVEL >0.75, SIFT <0.05 (damaging), PolyPhen-2 HDIV >0.85 (probably damaging)—and the proportion annotated Pathogenic in ClinVar. Proportions were calculated as 100 × {n=unique rsIDs meeting threshold}/{N=unique rsIDs in group} across all 15 genes. Between-group differences were assessed with two-sided Fisher's exact tests, and effect sizes were summarized as odds ratios (ORs) with 95% CIs using log(OR) (Woolf) approximations (Figure 4). Thresholds reflect original tool validations, systematic cross-tool evaluations, and contemporary ClinGen SVI guidance/calibration on the use and evidentiary weight of in-silico predictors (notably PP3/BP4), as well as dbNSFP aggregator releases that standardize predictor score retrieval and recommended cut-points [1,2,18–22,34–41,42,43]. Key benchmarking and guidance sources include Genome Biology's systematic review of in-silico algorithm use under ACMG/AMP [34], studies highlighting evaluation pitfalls and circularity [35], comparative and meta-predictor analyses [36,38], the dbNSFP v4 compendium [37], cancer-gene functional assay benchmarking with likelihood ratios for 44 tools [39], ClinGen SVI calibration of PP3/BP4 and recommendations for single pre-specified tools [40], and a Bayesian formalization of ACMG/AMP combining rules [41], with recent gene-by-gene calibration of REVEL/BayesDel further motivating fixed a priori thresholds [9].

Results

Distribution of ClinVar-Classified Variants across Breast Cancer Susceptibility Genes

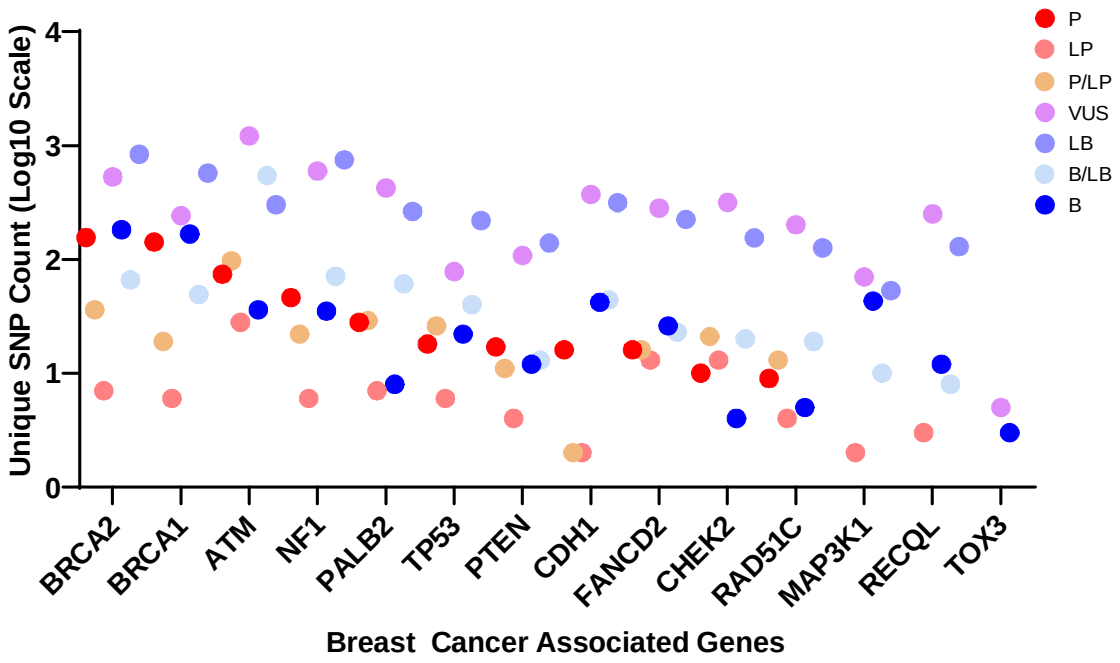


Figure 1. *Distribution of ClinVar-classified variants across breast cancer susceptibility genes.*
Log□□-scaled counts of unique SNPs per gene are shown, colored by clinical significance (P, LP, P/LP, VUS, LB, B/LB, B). *BRCA1* and *BRCA2* show the highest pathogenic variant counts, while low-penetrance genes like *MAP3K1*, *RECQL*, and *TOX3* have few. VUS remain prevalent across genes, reflecting challenges in interpretation.

We evaluated the unique ClinVar-reported SNPs across 15 breast cancer susceptibility genes, categorized by ClinVar clinical significance (Figure 1). Variant counts varied widely between genes, with *BRCA2* and *BRCA1* showing the highest numbers (>100 unique SNPs), followed by *ATM*, *NF1*, *PALB2*, *and TP53*, whereas *MAP3K1*, *RECQL*, *and TOX3* had very few unique SNPs. The proportion of entries classified as VUS remains substantial in ClinVar for cancer-associated genes, reflecting ongoing challenges in clinical interpretation [1,2]. Figure 1 demonstrates a declining gradient in total variant counts and in the proportion of pathogenic variants from high-penetrance genes (e.g., *BRCA2 to PALB2 to TP53*) to moderate (*CHEK2*, *RAD51C*) and low-penetrance loci (*MAP3K1*, *RECQL*, *TOX3*). These patterns are consistent with large studies showing that pathogenic variants are identified in a meaningful fraction of breast-cancer cases, with *BRCA1/2* contributing the largest share and additional risk arising

from other susceptibility genes [3–5].Figure 1 clearly demonstrates a declining gradient in both total variant counts and the proportion of pathogenic variants, from high-penetrance (e.g., *BRCA2 to PALB2 to TP53*) to moderate (*CHEK2*, *RAD51C*) and low-penetrance loci (*MAP3K1*, *RECQL*, *TOX3*).

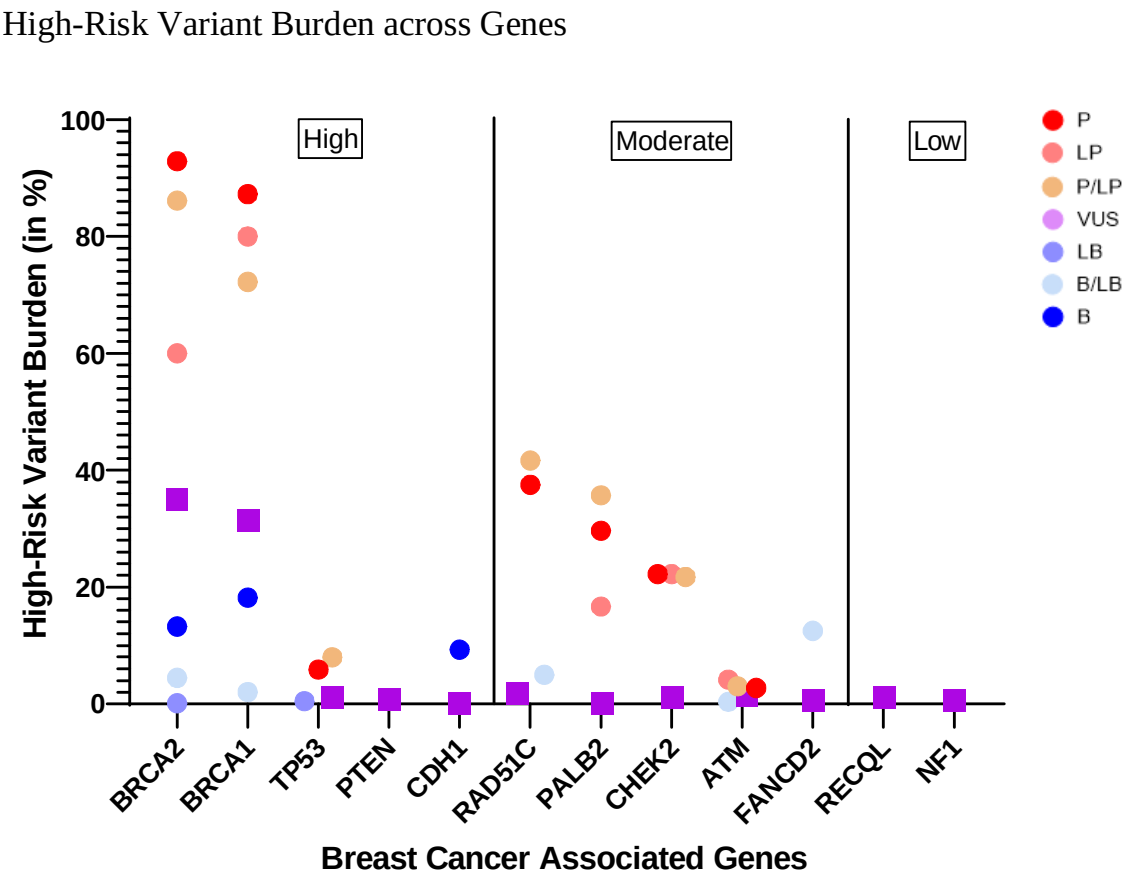


Figure 2. *High-risk variant burden across breast cancer susceptibility* genes. Percentage of high-risk variants (Pathogenic, Likely Pathogenic, and P/LP) per gene, stratified by penetrance class (high, moderate, low). Genes like *BRCA2* and *BRCA1* showed the highest burden (>85%), followed by *TP53*. Moderate-penetrance genes such as *PALB2* and *RAD51C* had intermediate burdens (~30–40%), while low-penetrance genes (*NF1*, *RECQL*, *FANCD2*) had minimal high-risk burden and were dominated by VUS.

We next quantified the burden of high-risk variants for hereditary breast cancer as defined in the Methods (Pathogenic [P], Likely Pathogenic [LP], and combined P/LP) across the 15 genes (Figure 2; Table 1). Genes were categorized into high, moderate, and low penetrance groups based on established risk estimates from prior studies [3–5]. High-penetrance genes showed the greatest burden of pathogenic variants. *BRCA2* exhibited the highest proportion of P variants (92.86%), followed by *BRCA1* (87.23%) and *TP53* (5.88%) [6–8]. Similarly, P/LP combined frequencies were highest in *BRCA2* (86.11%) and *BRCA1* (72.22%), with *TP53* markedly lower (8.00%) [6–8]. *PTEN* and *CDH1* also carried clinically relevant high-risk variants, albeit at lower frequencies than *BRCA1/2* [9,10]. In the moderate-penetrance

category, *PALB2* and *RAD51C* had the highest pathogenic variant proportions (29.63% and 37.00%, respectively), followed by *CHEK2* (22.22%) [11–13]. *ATM* carried a comparatively low fraction of pathogenic variants (2.74%) [14]. While *PALB2* and *RAD51C* showed substantial P/LP burdens (>35%), the remaining moderate-penetrance genes exhibited lower percentages [11–14]. For low-penetrance genes, the proportion of pathogenic variants was negligible, with most loci—such as *NF1*, *RECQL*, and Fanconi-pathway genes—showing ≤1–2% high-risk variant burden; most variants were VUS [15–17]. Overall, these analyses highlight a gradient in pathogenic burden that mirrors penetrance categories and agree with large panel and population studies reporting that *BRCA1/2* account for most actionable hereditary breast-cancer variants, with additional but smaller contributions from non-BRCA genes [3–5].In Silico Pathogenicity Predictions for High-Risk versus Non-High-Risk Variants

In-Silico Pathogenicity Predictions (High-Risk vs Non-High-Risk)

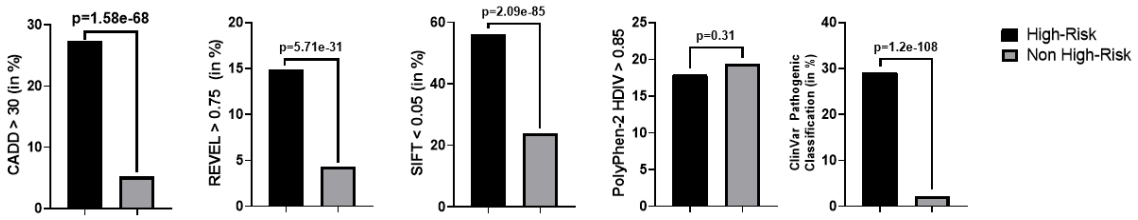


Figure 3. *Comparison of in silico pathogenicity scores between high-risk and non-high-risk variants.* High-risk variants showed significantly higher proportions exceeding established cutoffs for CADD (>30), REVEL (>0.75), and SIFT (<0.05) compared to non-high-risk variants (all $p < 10^{-3}$). No significant difference was observed for PolyPhen-2 HDIV (>0.85). ClinVar pathogenic classifications were also markedly enriched among high-risk variants (33.8% vs 1.9%, $p = 1.2\times10^{-108}$). These findings support the utility of select in silico scores and ClinVar assertions in distinguishing clinically relevant variants.

We compared multiple in silico scores between high-risk and non-high-risk variants (Figure 3). High-risk variants had a greater proportion with CADD > 30 (27.9% vs 6.5%, $p = 1.58\times10^{-68}$) [18,19] and REVEL > 0.75 (15.4% vs 4.1%, $p = 5.71\times10^{-31}$) [20]. A significantly larger fraction of high-risk variants had SIFT < 0.05 (57.3% vs 23.6%, $p = 2.09\times10^{-85}$) [21]. PolyPhen-2 HDIV > 0.85 was not different between groups (18.4% vs 19.7%, p = 0.31) [22]. As expected, the proportion of variants classified as pathogenic in ClinVar was far higher in the high-risk group (33.8% vs 1.9%, $p = 1.20\times10^{-108}$) [1,23]. Together, CADD, REVEL, and SIFT showed stronger discrimination than PolyPhen-2 HDIV, with ClinVar classifications aligning with our high-risk grouping.

Odds Ratios for Pathogenicity Predictors in High-Risk versus Non-High-Risk Variants

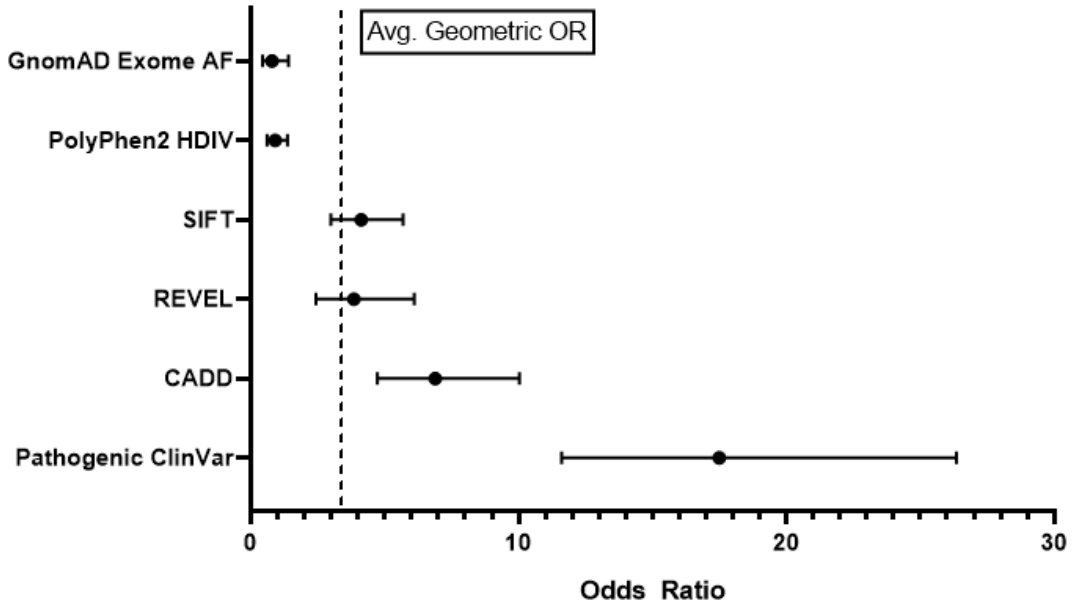


Figure 4. *Odds ratios (ORs) with 95% confidence intervals (CIs) for high-risk versus non-high-risk variants across key predictors.* The ClinVar pathogenic classification demonstrated the strongest discriminatory association (geometric mean OR ≈ 17.5), followed by CADD (≈ 6.9), SIFT (≈ 4.2), and REVEL (≈ 3.8). In contrast, PolyPhen-2 HDIV (≈ 0.9) and gnomAD exome allele frequency (≈ 0.8) were weak or non-informative in distinguishing high-risk variants. ORs were estimated using 2×2 Fisher's exact tests; point estimates represent geometric means across three replicates, and error bars denote observed minimum and maximum ORs.

To quantify discriminatory performance between high-risk and non-high-risk variant groups, we estimated odds ratios (ORs) with 95% confidence intervals (CIs) for each predictor using two-sided Fisher's exact tests (Figure 4). ClinVar pathogenic classification yielded the strongest association, with ORs ranging from 14.81 to 20.60 and a geometric mean of approximately 17.54, highlighting its value as a clinical gold standard for identifying pathogenic variants [1,23]. Among computational prediction tools, CADD showed the highest discriminatory power (geometric mean OR ≈ 6.92), reflecting its integration of diverse annotation, functional, and conservation-based features into a genome-wide pathogenicity score [18,19]. SIFT and REVEL followed with moderate associations (geometric means ≈ 4.23 and 3.84, respectively), indicating that protein conservation and ensemble-based models are informative in predicting variant impact [20,21]. In contrast, PolyPhen-2 HDIV exhibited limited discriminatory capacity (geometric mean OR ≈ 0.90),

consistent with prior evaluations reporting lower performance and calibration for clinical use [22,34–36].

gnomAD exome allele frequency yielded a geometric mean OR of 0.80, indicating that although high-risk variants are typically rare, rarity alone is insufficient to distinguish pathogenicity without supporting clinical or functional annotation [24,25]. Overall, these results emphasize the importance of integrating multiple lines of evidence for variant classification and suggest that while some prediction tools are informative, none alone can substitute for curated clinical databases or experimentally validated evidence.

Discussion

This study presents a comprehensive computational analysis of germline single-nucleotide variants (SNVs) in 15 breast cancer susceptibility genes using publicly available variant databases and ClinVar annotations. By integrating pathogenicity predictors (e.g., CADD, REVEL, SIFT, PolyPhen-2 HDIV) with allele frequency data from gnomAD and disease associations from ClinVar, we quantified the burden and distribution of high-risk variants across genes and assessed the discriminatory power of commonly used in silico tools.

Our results reaffirm the dominant contribution of *BRCA1* and *BRCA2* to hereditary breast cancer risk, with both genes showing the highest number of unique pathogenic or likely pathogenic (P/LP) variants and the greatest high-risk variant burden. This aligns with decades of literature identifying *BRCA1/2* mutations as the most penetrant drivers of familial breast cancer, with lifetime risk estimates ranging from 45% to 70% [6,7]. Genes such as *TP53*, *PTEN*, and *CDH1* also harbored clinically relevant high-risk variants but at lower frequencies, consistent with their established classification as high-penetrance genes involved in syndromic cancer risk (e.g., Li-Fraumeni and Cowden syndromes) [8–10]. Among moderate-penetrance genes, *PALB2*, *CHEK2*, and *RAD51C* showed measurable pathogenic variant frequencies, in agreement with population-based sequencing studies [11–13].

The observed gradient in pathogenic variant burden, from high to moderate to low-penetrance genes—is concordant with prior risk modeling studies [3–5]. For example, *ATM* and *FANCD2* showed low P/LP proportions, while *NF1*, *RECQL*, and *TOX3* had negligible high-risk variant content, suggesting that their overall contribution to inherited breast cancer may be limited or context-dependent [14–17]. Notably, *NF1* has previously been associated with moderate risk in specific subpopulations, though the majority of variants remain of uncertain significance [15]. When stratifying variants into high-risk versus non-high-risk groups based on allele frequency, disease label, and deleteriousness (CADD $\geq$ 30), we observed significant enrichment of functional impact signals in the high-risk group. CADD, REVEL, and SIFT scores were significantly elevated among high-risk variants, with odds ratios ranging from 3.8 to 6.9, while PolyPhen-2 HDIV showed no appreciable difference. These findings are in line with prior evaluations of pathogenicity prediction tools, which have consistently identified CADD and REVEL as among the best-performing algorithms for rare variant classification [34–38]. Our results also reinforce the limitations of relying on a single in silico tool and underscore the importance of integrating multiple evidence lines, as outlined in ACMG/AMP variant interpretation guidelines [23,40].

ClinVar pathogenic classification emerged as the most powerful discriminator of high-risk status, with an odds ratio exceeding 17 and a highly significant p-value ($p < 1e\text{-}100$). This reflects the clinical utility of curated variant databases and expert-reviewed annotations.

However, our findings also highlight persistent challenges: the high proportion of VUS across many genes reflects the uncertainty still surrounding rare missense variants and the need for more functional data, particularly in under-characterized loci. Despite the strengths of this analysis, including rigorous variant deduplication and prioritization pipelines, there are limitations. First, ClinVar submissions may be biased toward well-studied genes such as *BRCA1/2*, potentially inflating their apparent variant burden. Second, pathogenicity predictors are trained largely on known variants and may be less accurate for genes with sparse functional data. Third, the reliance on public databases limits clinical granularity (e.g., family history, co-occurring mutations, ancestry), which are critical for accurate risk modeling.

Future directions include incorporation of larger-scale case-control sequencing datasets, functional assay data, and population-specific variant annotations. As variant databases continue to grow, harmonized frameworks combining clinical data, evolutionary conservation, functional assays, and machine learning approaches will be key to resolving the large pool of VUS and refining gene-level penetrance estimates.

In conclusion, our study reinforces the centrality of *BRCA1/2* in hereditary breast cancer risk while providing comparative insight into other susceptibility genes. Computational pathogenicity scores such as CADD and REVEL effectively enrich for high-risk variants and may serve as supportive evidence in variant interpretation. However, their performance varies, and integrating them with curated clinical databases remains essential for robust classification. These findings support ongoing efforts to calibrate in silico tools and advocate for standardized, evidence-based variant assessment pipelines in clinical genomics.

Conclusion

This study presents a comprehensive variant-level analysis of 15 breast cancer susceptibility genes using large public databases and ClinVar annotations. Our findings reaffirm the central role of *BRCA1* and *BRCA2*, which exhibited the highest burden of high-risk variants towards hereditary breast cancer, consistent with their well-established high penetrance. Moderate-penetrance genes such as *PALB2*, *CHEK2*, and *RAD51C* showed appreciable pathogenic variant burdens, supporting their inclusion in clinical testing panels. In contrast, low-penetrance genes like *MAP3K1*, *RECQL*, and *TOX3* were predominantly enriched for variants of uncertain significance, highlighting gaps in current variant interpretation efforts.

In silico tools such as CADD, REVEL, and SIFT effectively discriminated high-risk from non-high-risk variants, with significantly higher proportions of high-scoring variants observed in the high-risk group. However, PolyPhen-2 HDIV showed limited utility, underscoring the need to rely on better-calibrated and validated predictors. While low allele frequency is an important filtering criterion, our analysis reinforces that functional and clinical annotations are essential for identifying pathogenic variants.

Together, these findings highlight the utility and limitations of current computational tools and databases in variant interpretation. They also provide a scalable framework for assessing variant burden across cancer susceptibility genes. Continued integration of population genomics, functional validation, and clinical data will be key to resolving uncertain variants and enhancing precision medicine in hereditary breast cancer.

Data Availability

Data Is publicly available through:-
1) The NYGC 30× GRCh38 1000 Genomes annotated coding and rare-coding flat files were obtained from the IGSR/NYGC 30× data collection (https://www.internationalgenome.org/data-portal/data-collection/30x-grch38).
2) Exome variant data and allele frequencies for breast cancer genes were obtained from gnomAD v4.1.0 (https://gnomad.broadinstitute.org/), which provides publicly accessible, de-identified population genetic data.

Author contributions

Kanishk Yadav, MS – Conceptualization, Coding, Writing the manuscript, Figures

Abhishek Prasad, MD – Editing the manuscript

Competing interests

The authors have no conflicts of interest to declare.

Declarations/Ethics

Ethics approval and consent to participate: Not applicable. All data analyzed in this study were obtained from publicly available, de-identified aggregate genomic databases (gnomAD, 1000 Genomes Project, ClinVar). No new data were collected and no identifiable human data, human tissue, or direct human participation were involved.

References

1. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062-7. doi:10.1093/nar/gkx1153.
2. Landrum MJ, Kattman BL. ClinVar at five years: delivering on the promise. Hum Mutat. 2018;39(11):1623-30. doi:10.1002/humu.23641.
3. Easton DF, Pharoah PDP, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, et al. Gene-panel sequencing and the prediction of breast-cancer risk. N Engl J Med. 2015;372(23):2243-57. doi:10.1056/NEJMsr1501341.
4. Couch FJ, Shimelis H, Hu C, Hart SN, Polley EC, Na J, et al. Associations between cancer predisposition testing panel genes and breast cancer. JAMA Oncol. 2017;3(9):1190-6. doi:10.1001/jamaoncol.2017.0424.
5. Hu C, Hart SN, Gnanaolivu R, Huang H, Lee KY, Na J, et al. A population-based study of genes previously implicated in breast cancer. N Engl J Med. 2021;384(5):440-51. doi:10.1056/NEJMoa2005936.

6. Antoniou AC, Pharoah PDP, Narod S, Risch HA, Eyfjord JE, Hopper JL, et al. Average risks of breast and ovarian cancer associated with *BRCA1* or *BRCA2* mutations. Am J Hum Genet. 2003;72(5):1117-30. doi:10.1086/375033.
7. Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips KA, Mooij TM, Roos-Blom MJ, et al. Risks of breast, ovarian, and contralateral breast cancer for *BRCA1* and *BRCA2* mutation carriers. JAMA. 2017;317(23):2402-16. doi:10.1001/jama.2017.7112.
8. Bougeard G, Renaux-Petel M, Flaman JM, Charbonnier C, Fermey P, Belotti M, et al. Revisiting Li–Fraumeni syndrome from *TP53* mutation carriers. J Clin Oncol. 2015;33(21):2345-52. doi:10.1200/JCO.2014.59.5728.
9. Tan MH, Mester J, Peterson C, Yang Y, Chen JL, Rybicki LA, et al. Lifetime cancer risks in individuals with germline *PTEN* mutations. Clin Cancer Res. 2012;18(2):400-7. doi:10.1158/1078-0432.CCR-11-2283.
10. Pharoah PD, Guilford P, Caldas C; International Gastric Cancer Linkage Consortium. Incidence of gastric cancer and breast cancer in *CDH1* mutation carriers. J Natl Cancer Inst. 2001;93(18):1449-56. doi:10.1093/jnci/93.18.1449.
11. Antoniou AC, Casadei S, Heikkinen T, Barrowdale D, Pylkäs K, Roberts J, et al. Breast-cancer risk in families with mutations in *PALB2*. N Engl J Med. 2014;371(6):497-506. doi:10.1056/NEJMoa1400382.
12. Pujol P, de Paillerets BB, Coupier I, Gauthier-Villars M, Stoppa-Lyonnet D, Frenay M, et al. RAD51C germline mutations in breast and ovarian cancer families. PLoS One. 2012;7(1):e29346. doi:10.1371/journal.pone.0029346.
13. Weischer M, Bojesen SE, Ellervik C, Tybjærg-Hansen A, Nordestgaard BG. *CHEK2**1100delC genotyping for clinical assessment of breast cancer risk. J Clin Oncol. 2008;26(4):542-8. doi:10.1200/JCO.2007.12.5922.
14. Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, et al. Rare, evolutionarily unlikely missense substitutions in *ATM* confer increased risk of breast cancer. Hum Mutat. 2009;30(7):1097-110. doi:10.1002/humu.21047.
15. Sharif S, Moran A, Huson SM, Iddenden R, Shenton A, Howard E, et al. Women with neurofibromatosis 1 are at a moderately increased risk of developing breast cancer. J Med Genet. 2007;44(8):481-4. doi:10.1136/jmg.2007.049346.
16. Sun J, Wang Y, Xia Y, Xu Y, Ouyang T, Li J, et al. Mutations in RECQL gene are associated with predisposition to breast cancer. PLoS Genet. 2015;11(5):e1005228. doi:10.1371/journal.pgen.1005228.
17. Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, et al. Truncating mutations in BRIP1 (FANCJ) are low-penetrance breast-cancer susceptibility alleles. Nat Genet. 2006;38(11):1239-41. doi:10.1038/ng1902.
18. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310-5. doi:10.1038/ng.2892.
19. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886-94. doi:10.1093/nar/gky1016.
20. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99(4):877-85. doi:10.1016/j.ajhg.2016.08.016.
21. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812-4. doi:10.1093/nar/gkg509.
22. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9. doi:10.1038/nmeth0410-248.

23. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants (ACMG/AMP). Genet Med. 2015;17(5):405-24. doi:10.1038/gim.2015.30.
24. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43. doi:10.1038/s41586-020-2308-7.
25. Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, et al. Using high-resolution variant frequency data from gnomAD to empower clinical genome interpretation. Genet Med. 2017;19(10):1151-8. doi:10.1038/gim.2017.26.
26. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-11. doi:10.1093/nar/29.1.308.
27. Pal M, et al. Genetics of breast cancer: an overview. World J Surg Oncol. 2024;22:271. doi:10.1186/s12957-024-03553-9.
28. National Cancer Institute. Genetics of Breast and Gynecologic Cancers (PDQ®) – Health Professional Version. Updated 12 Apr 2024. Available at: https://www.cancer.gov/types/breast/hp/breast-ovarian-genetics-pdq
29. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation (1000 Genomes Project). Nature. 2015;526(7571):68-74. doi:10.1038/nature15393.
30. Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Res. 2021;49(D1):D939-46. doi:10.1093/nar/gkaa980.
31. McKinney W. Data Structures for Statistical Computing in Python. Proc. 9th Python in Science Conf. 2010:51-56.
32. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature. 2020;585(7825):357-62. doi:10.1038/s41586-020-2649-2.
33. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;9(3):90-5. doi:10.1109/MCSE.2007.55.
34. Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol. 2017;18:225. doi:10.1186/s13059-017-1353-5.
35. Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat. 2015;36(5):513-23. doi:10.1002/humu.22768.
36. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125-37. doi:10.1093/hmg/ddu733.
37. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12:103. doi:10.1186/s13073-020-00803-9.
38. Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, et al. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res. 2018;46(15):7793-804. doi:10.1093/nar/gky678.
39. Çubuk C, Serra-Garcia L, Khrustaleva T, et al. Clinical likelihood ratios and balanced accuracy for 44 in silico tools against multiple large-scale functional assays of cancer susceptibility genes. Genet Med. 2021;23(2):209-18. doi:10.1038/s41436-021-01265-z.

40. Pejaver V, Babbi G, Casadio R, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am J Hum Genet. 2022;109(11):2163-77.
41. Tavtigian SV, Greenblatt MS, Harrison SM, Nussbaum RL, Prabhu SA, Boucher KM, Biesecker LG; ClinGen SVI Working Group. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet Med. 2018;20(9):1054-60. doi:10.1038/gim.2017.210.
42. Woolf B. On estimating the relation between blood group and disease. Ann Hum Genet. 1955;19(4):251-3. doi:10.1111/j.1469-1809.1955.tb01348.x.
43. Agresti A. An Introduction to Categorical Data Analysis. 3rd ed. Hoboken, NJ: Wiley; 2019.

Table 1. Variant Burden of genes by their ClinVar classification for determined High-Risk variants threshold for hereditary breast cancer:

Gene	**B**	**B/LB**	**LB**	**LP**	**P**	**P/LP**	**VUS**
ATM	0	0.362319	0	4.166667	2.739726	3.030303	1.486375
BRCA1	18.18182	2.040816	0	80	87.23404	72.22222	31.53527
BRCA2	13.26531	4.477612	0.1221	60	92.85714	86.11111	35.16068
CDH1	9.302326	0	0	0	0	0	0.27027
CHEK2	0	0	0	22.22222	22.22222	21.73913	1.269841
FANCD2	0	12.5	0	0	0	0	0.714286
MAP3K1	0	0	0	0	0	0	0
NF1	0	0	0	0	0	0	0.671141
PALB2	0	0	0	16.66667	29.62963	35.71429	0.236407
PTEN	0	0	0	0	0	0	0.934579
RAD51C	0	5	0	0	37.5	41.66667	2
RECQL	0	0	0	0	0	0	1.204819
TOX3	0	0	0	0	0	0	0
TP53	0	0	0.47619	0	5.882353	8	1.298701