Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

EthniVar: a source-stratified catalogue of BRCA1/2 germline variants quantifies the classification gap for variants reported in Indian patients

Paras Verma
Plaksha University
Swagata Halder
Plaksha University
Monika Sharma
monika.sharma@plaksha.edu.in

Plaksha University

Research Article

Keywords: BRCA1, BRCA2, Variants of Unknown Significance (VUS), ancestry diversity, variant classification

Posted Date: September 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10983210/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Pathogenicity classification for BRCA1/2 germline variants has been derived largely from individuals of European ancestry, leaving South Asian populations markedly under-represented in ClinVar and gnomAD. The clinical consequence is that a variant reported predominantly in Indian patients may lack the population-frequency, functional and expert-curation evidence available for a variant of comparable biological consequence in a European-ancestry patient. The magnitude of this curation gap for BRCA1 and BRCA2, and the extent to which recent advances in computational prediction can bridge it, have not been quantified at the variant level.

Methods

We compiled a source-stratified catalogue of 3,315 unique BRCA1/2 germline variants (BRCA1: 1617; BRCA2: 1698) from literature and databases, grouped by the population descriptors recorded in their source metadata as European/Caucasian (n = 1301), Chinese (n = 1139) and Indian (n = 569); these reflect how variants entered the shared evidence base rather than genetically inferred ancestry. Variants were validated against their original genome assembly, lifted to hg38, and re-annotated with ANNOVAR and dbNSFP4.7a. We scored variants with six dbNSFP rank scores spanning supervised, evolutionary, protein-language and structure-informed approaches, and calculated an unweighted consensus for research prioritization.

Results

Indian-catalogue-specific variants showed a substantially larger curation gap, with only 37.4% carrying definitive ClinVar classifications, compared with 61.0% for European/Caucasian-catalogue-specific and 59.8% for Chinese-catalogue-specific variants. The consensus framework achieved AUC = 0.99 on 631 benchmark variants (sensitivity 99.4%, specificity 93.3% among 575 directional calls) and recovered the expected purifying selection signal, with a 3.4-fold ratio of median maximum allele frequency for benign-range versus pathogenic-range variants. Applied to the 356 uninformative Indian-catalogue-specific variants, the framework returned HIGH or MEDIUM confidence for 75 (21.1%), of which 33 had intermediate scores and 42 were directional (31 pathogenic-range, 11 benign-range). None of the 21 variants flagged by the conflicting allele-frequency screen met the dual-evidence criteria.

Conclusions.

EthniVar documents a larger unresolved classification fraction among variants reported only in Indian sources and provides a transparent research shortlist for further curation. Its consensus score, intended for prioritization narrows interpretation burden, but cannot close the gap: it resolved fewer than a quarter of uninformative Indian-catalogue-specific variants and none of those carrying conflicting classifications. The residual fraction requires review of ClinVar submissions, population-matched allele counts, segregation

evidence, and calibrated functional assays. The curated dataset and individual predictor scores are available as a live web database (https://brca-ethnivar.vercel.app).

INTRODUCTION

Natural selection acts most strongly on variants that disrupt essential gene function, and this constraint underlies both the pathogenicity and the rarity of damaging variants in high-penetrance cancer predisposition genes [1].

Inherited pathogenic variants in *BRCA1* and *BRCA2* confer one of the largest known elevations in lifetime breast and ovarian cancer risk. In a large prospective study, cumulative breast-cancer risk by age 80 was estimated at 72% for BRCA1 carriers and 69% for BRCA2 carriers, while the corresponding ovarian-cancer risks were 44% and 17%, respectively [2, 3]. Their identification informs clinical management through enhanced screening, risk-reducing surgery, and targeted therapies such as PARP inhibitors [4, 5]. The evidence base supporting these classifications, however, has been built largely on individuals of European ancestry [6–8]. For populations historically under-represented in genomic databases (including South Asian, Southeast Asian, and African populations), this leads to clinical inequity, where a patient of Indian ancestry and a patient of European ancestry carrying the same variant may receive different clinical reports [9–11]. Reducing this disparity is essential for advancing equitable genomics initiatives in healthcare [12–14].

The established clinical actionability of BRCA1/2 pathogenic variants has made these genes among the most widely sequenced cancer predisposition genes worldwide [15, 16]. For BRCA1/2 specifically, the ClinGen ENIGMA Variant Curation Expert Panel has developed gene-specific specifications that calibrate the ACMG/AMP framework for their interpretation [17]. Despite this, a majority of unique missense variants in *BRCA1* and *BRCA2* remain classified as variants of uncertain significance (VUS) in ClinVar [18], and newly observed variants often lack sufficient evidence for classification. Importantly, the uninformative variant fraction - VUS, variants with conflicting interpretations, and variants absent from ClinVar - is not distributed evenly across populations, because both the curation infrastructure and the upstream functional studies derive predominantly from European-ancestry data. This accordingly may contribute to under-represented populations accumulating uninformative classifications at substantially higher rates [6–10]. A direct precedent for clinical consequences of this skew comes from Manrai et al. [19], who showed that hypertrophic-cardiomyopathy variants once classified as pathogenic were reclassified once African-American population frequencies were incorporated, a case in which ancestry bias in variant databases translated into clinical misinterpretation. A related pattern has been documented for Indian clinical context, where Jadeja et al. [11] describe how a VUS in India is treated with substantially more clinical caution than the same call in the United States, where comparable results are often interpreted as near-negative.

Computational prediction offers one scalable aid. Supervised meta-predictors trained on clinical databases [20, 21] and unsupervised evolutionary or protein-language models that derive pathogenicity estimates from cross-species sequence variation without clinical labels [22, 23] have expanded the resources available for variant assessment. Their performance is increasingly benchmarked within the ACMG/AMP framework as computational pathogenic-supporting or benign-supporting evidence (PP3/BP4), provided they are calibrated gene by gene [17]. Population allele frequency (AF) data provide a complementary line of evidence, informing

purifying-selection-based assessment and constituting a key ACMG/AMP criterion in their own right [7, 24, 25]. Against a background of rising breast cancer incidence in South Asia [26, 27], the equity-relevant question is how far combining these tools limits inheritance of European-ancestry bias from their training data [28], and whether they resolve the specific BRCA1/2 variants disproportionately carried by under-represented groups.

In this study, we address the variant interpretation gap for BRCA1/2 germline variants reported in Indian breast cancer patients, compared against Chinese-source and European/Caucasian-source variant catalogues. Population descriptors in this field are applied inconsistently and the available terms are imperfect; we use these labels as they appear in the source records from which each catalogue was assembled, treating them as descriptors of how and where variants entered the shared evidence base rather than as genetically defined ancestry groups. Their derivation and limitations are set out in Methods. Our unit of analysis is the distinct normalised variant rather than the participant, and cross-group comparisons therefore describe catalogue composition rather than population allele architecture.

Because Indian genomic data remain sparse, we assembled an integrated Indian-source catalogue from the Bharat Cancer Genome Atlas (BCGA), the comparative dataset of Bhaskaran et al. [29], and targeted literature mining of breast cancer studies published subsequently (see Methods). Variants were drawn primarily from disease-focused sources rather than a population-frequency reference, distinguishing this resource from gnomAD-derived prior work and IndiGenomes [30].

We pursued three aims. First, we quantified the ClinVar curation gap among catalogue-exclusive variants and characterised how it differs across the three catalogues. Second, we developed and benchmarked a six-predictor consensus framework spanning supervised, evolutionary, protein-language and structure-informed paradigms [28], evaluated against ClinVar two-star variants and gnomAD v4.1 allele frequency data. This composition reduces dependence on any single method, but we emphasise that supervised predictors are themselves trained on classification sets in which European-ancestry variants predominate; the framework inherits rather than corrects this bias, and we treat that inheritance as a constraint on what computational evidence can deliver. Third, among variants showing population-enriched allele frequency patterns alongside conflicting ClinVar classifications, we examined whether consensus scoring and allele frequency evidence together could resolve the conflict.

The work is intended as a quantitative mapping of where computational evidence narrows the interpretation gap and an explicit account of where it does not. The curated dataset and individual predictor scores are released openly (https://brca-ethnivar.vercel.app) to support follow-up by functional and population-sequencing groups working toward more equitable BRCA1/2 interpretation in South Asia.

RESULTS

2.1 BRCA1/2 variant landscape across the three populations

After validation and hg38 re-annotation, the combined dataset comprised 3,315 unique BRCA1/2 germline variants, including 1,617 BRCA1 variants (48.8%) and 1,698 BRCA2 variants (51.2%), distributed across the

three population catalogues as shown in Fig. 1A.

Analysis of variant sharing across populations showed that only 9.2% of variants were shared by at least two populations, and approximately 1% were shared across all three populations (Fig. 1A). Thus, the vast majority of variants were population-specific: 1,301 variants were specific to European/Caucasian catalogue (~ 39%), 1,139 to Chinese catalogue (~ 34%), and 569 to Indian catalogue (~ 17%). The BRCA2/BRCA1 variant ratio varied modestly across catalogues (1.16 European/Caucasian, 1.10 Indian, 1.01 Chinese; Supplementary Figure S1A); the Indian catalogue showed a slightly higher relative proportion of BRCA2 variants than the Chinese catalogue and a slightly lower proportion than the European/Caucasian catalogue.

We next examined the genomic distribution of these variants. Approximately 10% occurred in non-coding regions, most of which were intronic, followed by splice-site variants defined as intronic changes located within ± 2 bp of exon boundaries (Fig. 1B). Among the remaining ~ 90% exonic variants, molecular consequence analysis showed that missense variants, resulting in amino acid substitutions, were the most frequent category, accounting for 34.5% of all variants (Fig. 1C). Loss-of-function variants collectively represented approximately 50% of all variants and included frameshift indels and substitutions, nonsense variants, and start-loss variants. At the gene level, *BRCA2* harbored a higher proportion of missense variants than *BRCA1* in our dataset, 40.4% versus 36.3%, respectively (Supplementary Figure S1B), consistent with the longer coding sequence of *BRCA2* compared with *BRCA1* (3,418 amino acids versus 1,863 amino acids). Minor variant classes included synonymous variants (3.2%) and in-frame indels and substitutions (1.8%). Having characterized the genomic and functional composition of *BRCA1/2* variants, we next evaluated their clinical interpretability.

ClinVar annotation of the dataset revealed substantial gaps in clinical interpretability (Fig. 1D). Nearly half of the variants were classified as pathogenic or likely pathogenic (~ 49%), whereas only ~ 10% were annotated as benign or likely benign, reflecting the disease-focused nature of the studies from which these variants were sourced. Importantly, ~ 13% of variants were absent from ClinVar (n = 430), while 0.7% (n = 23) had ‘not_provided’ clinical significance and an additional ~ 27.0% were assigned uncertain or conflicting significance, including VUS (5.9%, n = 197) and conflicting classifications (21.1%, n = 700). Collectively, approximately two-fifths of the dataset lacked definitive clinical resolution and represent key targets for further interpretation using computational prediction approaches, population frequency data, or an integrated combination of both.

2.2 Differences in ClinVar curation across source catalogues.

Given the abundance of pathogenic variants in our dataset, together with a substantial fraction of uninformative variants including those with conflicting classifications, VUS, or absence from ClinVar, we next analyzed variants by the source-population label. This was necessary because aggregate-level analysis can obscure important patterns. Although previous studies have reported poorer representation of Asian populations in genomic variant repositories [14], the extent of this gap has not been comprehensively characterized using a recent disease-derived germline variant catalogue annotated with updated ClinVar classifications.

2.2.1 ClinVar completeness

Comparing definitive ClinVar classifications (pathogenic/likely pathogenic or benign/likely benign), we observed that only 37.4% of Indian-catalogue-specific variants (213/569) had definitive classifications, compared to 61.0% for European/Caucasian-catalogue-specific variants (794/1301) and 59.8% for Chinese-catalogue-specific variants (681/1139) ($\chi^2 = 99.63$, df = 2, $p = 2.32 \times 10^{-22}$ (3-way), Fig. 2A). Chinese and European/Caucasian curation rates did not differ significantly (59.8% vs 61.0%, $\chi^2 = 0.34$, p = 0.56), although their qualitative composition differed with 88.5% of Chinese definitive variants classified pathogenic versus 78.2% European/Caucasian. The Indian catalogue showed significantly lower definitive classification rates than both comparators (versus European/Caucasian: $\chi^2 = 87.74$, Holm-adjusted $p = 2.24 \times 10^{-20}$; versus Chinese: $\chi^2 = 75.12$, Holm-adjusted $p = 8.84 \times 10^{-18}$). The pattern held gene-wise, with Indian BRCA2 variants most severely under-curated (43.1% definitive) relative to European/Caucasian (57.2%) and Chinese BRCA2 variants (66.0%) (Supplementary Figure S2A). Within the curated Indian set, BRCA1 variants showed a higher pathogenic fraction than BRCA2 variants (Supplementary Figure S2B).

Examining subtypes of the uninformative category (Supplementary Figure S2C), Indian- and Chinese-catalogue variants were absent from ClinVar in approximately one quarter of cases (25.3% and 24.4%, respectively), compared with 0.4% of European/Caucasian-catalogue variants. The Indian-catalogue variants had the highest VUS rate at 11.2%, followed by the European/Caucasian catalogue at 7.2% and the Chinese catalogue at 3.4%.

Among variants with conflicting classifications, the European/Caucasian catalogue showed the highest rate at 31.4%, exceeding the Indian catalogue at 24.1% and the Chinese catalogue at 11.3%. The high conflicting rate in the European/Caucasian catalogue occurred alongside near-complete ClinVar coverage (0.4% absent), whereas the elevated Indian rate occurred alongside 25.3% absence.

2.2.2 Pathogenic fraction among definitively classified variants

Restricting to variants with definitive ClinVar classifications, pathogenic ratios ranged from 78% to 89% across catalogues: 81.2% for Indian-catalogue-specific, 78.2% for European/Caucasian-catalogue-specific and 88.5% for Chinese-catalogue-specific variants. Fisher's exact test showed no significant difference between the Indian and European/Caucasian catalogues (OR = 0.83, Holm-adjusted p = 0.395), whereas the Chinese catalogue showed a significantly higher pathogenic ratio than both the Indian (OR = 0.559, Holm-adjusted p = 0.015) and European/Caucasian catalogues (OR = 0.464, Holm-adjusted $p = 3.34 \times 10^{-7}$) (Fig. 2B). These values describe unique reported variants rather than carrier prevalence, and differences between them reflect catalogue composition rather than population-level pathogenic-variant frequency.

2.3 Molecular class composition of the uninformative fraction

To assess whether uninformative ClinVar categories were distributed evenly across molecular classes, we examined completeness by exonic variant type. Loss-of-function (LOF) variants - frameshift, nonsense and

start-loss - showed near-complete definitive classification (1,368/1,664; 82.2%) (Figs. 3A–C). Missense variants showed the opposite pattern: only 261/1,144 (22.9%) carried definitive classifications, with the remaining 883 classified as conflicting, VUS, or absent/not provided (Fig. 3C). Uncertainty was thus approximately 4.3-fold enriched among missense variants relative to LOF variants (77.1% versus 17.8% uncertain).

Because missense variants dominated the uninformative fraction, we tested whether missense proportion differed across the three catalogues. A chi-square test of independence showed a significant association ($\chi^2 = 202.8$, df = 2, $p = 8.97 \times 10^{-45}$, Cramér's V = 0.392; Fig. 3B), with the Indian catalogue intermediate (69.1%) between the European/Caucasian (84.2%) and Chinese (40.8%) catalogues. The Indian catalogue therefore showed the lowest definitive classification rate despite an intermediate missense burden, indicating that missense composition alone does not account for its coverage gap.

Applying the pathogenic proportions observed in the European/Caucasian and Chinese catalogues to the 356 uninformative Indian-catalogue-specific variants gives a projected 99–278 additional pathogenic variants, a 57–161% increase over the 173 currently classified as pathogenic or likely pathogenic. The lower figure assumes the pathogenic proportion across all Indian-catalogue-specific variants matches the overall rate in the comparator catalogues (47.7% European/Caucasian, 52.9% Chinese); the upper figure assumes uninformative variants resolve at the rate observed among informative variants (78.2% and 88.5% respectively; Fig. 2B). The upper scenario is unlikely to hold, since the uninformative pool is enriched for missense variants, which carry lower prior pathogenic probability than the LOF variants dominating the informative set. The projection is therefore indicative of the scale of unresolved classification rather than an estimate of the number of pathogenic variants present.

2.4 Computational approaches to bridge the variant classification gap

To assess uninformative variants computationally, we implemented a consensus scoring framework integrating six pathogenicity predictors from dbNSFP v4.7a (see Methods for predictor rationale): BayesDel_noAF, a Bayesian meta-predictor [31] specified by ENIGMA for BRCA1/2 [17]; AlphaMissense, a structure-informed deep learning model [32]; EVE, an unsupervised evolutionary variational autoencoder [22]; ESM1b, a protein language model [23]; ClinPred, a clinical ensemble gradient boosting method [20]; and VEST4, an HGMD-trained random forest [21]. The selection spans supervised methods (BayesDel_noAF, ClinPred, VEST4), unsupervised methods (EVE, ESM1b) and a sequence- and structure-informed model (AlphaMissense), prioritising methodological coverage while recognising possible shared features and training-set overlap. The mean of the normalised scores is referred to as the consensus score [28].

We first evaluated whether the predictors provided sufficiently non-identical information to justify a consensus approach. Pairwise Spearman correlations ranged from 0.537 to 0.852 (mean 0.717; Fig. 4A), with only three of fifteen pairs exceeding 0.8 (EVE–ClinPred, EVE–VEST4, VEST4–BayesDel), indicating substantial non-redundancy across predictors.

Consensus scores required valid scores from at least two of the six predictors. On this criterion, 1,591 variants (48.0% of the dataset) received consensus scores. Coverage differed sharply by molecular class: scores were assigned to 100% of missense variants but only 26.7% of LOF-class variants (Fig. 3D). Stratified by uninformative ClinVar category, coverage was 91.6% for conflicting variants, 91.3% for variants with 'not provided' significance, 70.6% for VUS, and 26.3% for ClinVar-absent variants.

2.5 ClinVar label concordance of the consensus score

We assessed concordance against 631 definitively labelled ClinVar variants with review status of at least two stars (493 pathogenic, 138 benign). Fifty-six variants (22 pathogenic, 34 benign) scored within the intermediate range of 0.40–0.70 and were excluded from directional calling, giving a called set of 575 variants (471 pathogenic, 104 benign). The predominance of pathogenic variants reflects both the disease-focused study design and ClinVar's primary role in aggregating pathogenicity evidence. Across the called set, the consensus score achieved sensitivity 99.36%, specificity 93.27%, accuracy 98.26% and precision 98.53% (Fig. 4B); the continuous-score AUC was 0.99 across all 631 benchmark variants.

Stratified by confidence category (see Methods), 444 variants were high confidence (1 further variant excluded as intermediate), 30 medium confidence (17 excluded), and 101 low confidence (38 excluded) (Fig. 4D).

2.5.1 Performance across confidence tiers

High confidence (n = 444). All 443 pathogenic variants were correctly identified, and the single benign variant in this tier was misclassified as pathogenic (TP = 443, FP = 1, TN = 0, FN = 0; Fig. 4C, left panel). Specificity in this tier is therefore 0/1 and statistically unstable. The false positive was BRCA2 c.9976A > T p.(Lys3326Ter) (NM_000059.4), a stop-gain variant truncating the terminal 93 residues, which received a consensus score of 0.77 and has a documented history of classification change [33].

Medium confidence (n = 30). All 24 pathogenic variants were correctly identified. Of the six benign variants, two were correctly classified and four were misclassified as pathogenic (TN = 2, FP = 4; Fig. 4C, middle panel).

Low confidence (n = 101). This tier comprised 4 pathogenic and 97 benign variants. One of four pathogenic variants was correctly identified (TP = 1, FN = 3), and 95 of 97 benign variants were correctly classified (TN = 95, FP = 2; Fig. 4C, right panel).

2.6 Application to uninformative variants

Of 1,350 uninformative variants (including 883 missense), 914 (67.7%) received consensus scores (Fig. 4E): 115 high confidence, 146 medium confidence and 653 low confidence, corresponding to 28.6% high- or medium-confidence and 71.4% low-confidence predictions. Directional calls comprised 118 pathogenic-range variants (91 high, 24 medium, 3 low confidence; 89 missense) and 496 benign-range variants (1 high, 28 medium, 467 low confidence; 494 missense). The remaining 300 scored variants fell within the intermediate range of 0.40–0.70 (23 high, 94 medium, 183 low confidence).

Among the 356 uninformative Indian-catalogue-specific variants, 252 (70.8%) received a consensus score: 33 high confidence, 42 medium confidence and 177 low confidence. Of the 75 high- or medium-confidence variants, 33 fell in the intermediate range and 42 were directional (31 pathogenic-range, 11 benign-range). Three further pathogenic-range calls occurred at low confidence, giving 34 pathogenic-range calls across all tiers; 120 low-confidence variants were benign-range.

Gene-wise, Indian-catalogue-specific BRCA1 variants yielded 27 high- or medium-confidence predictions, of which 19 were pathogenic-range, whereas BRCA2 yielded 48, of which 12 were pathogenic-range. The pathogenic-range fraction among high- and medium-confidence predictions was therefore higher for BRCA1 than BRCA2 (70.4% versus 25.0%; OR = 7.12, $p = 2.03 \times 10^{-4}$).

2.7 Descriptive Allele frequency analyses

2.7.1 Comparison of consensus score and allele frequency

We next compared score direction against population allele frequency (AF) as an additional line of evidence, on the premise that high-penetrance pathogenic variants occur at lower population frequencies under purifying selection whereas benign variants may persist at higher frequencies [34] (Fig. 5A).

Of the 1,591 variants with consensus predictions, 273 had a directional score and a maximum point AF greater than zero in at least one gnomAD group, based on gnomAD v4.1 annotations in dbNSFP: 49 pathogenic-range and 224 benign-range. Because each variant could carry up to three reference-group AF values (gnomAD SAS, EAS and NFE), we used the maximum AF across groups. Reference-group AF distributions are shown in Fig. 5B, and the relationship between log-scaled maximum AF and consensus score in Fig. 5C.

Pathogenic-range variants showed significantly lower AFs than benign-range variants (median 0.0029% versus 0.0100%; 3.4-fold ratio of medians; one-sided Mann–Whitney U = 3,657.5, $p < 0.0001$; rank-biserial r = 0.334; Fig. 5A). Ultra-rare variants (AF < 0.01%) comprised 69.4% of pathogenic-range variants (n = 34) versus 48.2% of benign-range variants (n = 108). Conversely, 17.0% of benign-range variants (n = 38) exceeded AF 0.1%, compared with 4.1% of pathogenic-range variants (n = 2).

Two variants were discordant, carrying AF > 0.1% despite pathogenic-range scores (Fig. 5C): BRCA2 c.9976A > T (NM_000059.4; AF = 0.85%; consensus = 0.772; high confidence) and BRCA2 c.8702G > A (NM_000059.4; AF = 0.12%; consensus = 0.736; low confidence). Both carry benign ClinVar classifications. The first is the same variant identified as the sole false positive in Section 2.5.1, here flagged independently by frequency evidence.

For subsequent descriptive integration, maximum AF was binned as AF-low (< 0.01%), AF-intermediate (0.01–0.1% inclusive) or AF-high (> 0.1%). These are study-specific descriptive categories derived from nominal point AF rather than filtering allele frequency, and are not directly transferable to ENIGMA BA1/BS1 criteria ([17]; see Methods).

2.7.2 AF-ratio screen

We compared inter-catalogue allele frequencies, for variants with available AF data from at least two reference groups. Among the 120 variants with positive AF in at least two reference groups, 76 were AF-differentiated, defined by a greater than three-fold ratio of maximum to minimum reference-group AF. These were distributed across SAS-maximum (n = 33, 43.4%), EAS-maximum (n = 32, 42.1%) and NFE-maximum (n = 11, 14.5%) groups, with AF distributions shown in Fig. 5B. Of the 76, 53 (69.7%) carried benign or likely benign ClinVar classifications, 21 (27.6%) carried conflicting classifications, one was classified pathogenic and one VUS. No NFE-maximum variant carried a conflicting classification.

We then examined the 21 AF-differentiated variants with conflicting classifications against their consensus scores. Maximum AF was in SAS for 12 and EAS for 9. Consensus scores were available for 19 (10 SAS-maximum, 9 EAS-maximum); two SAS-maximum variants lacked sufficient predictor coverage. Most scored variants were benign-range (n = 14) or intermediate (n = 4), with one EAS-maximum variant in the pathogenic range.

Seventeen of the 19 scored variants carried low-confidence scores, comprising 14 benign-range predictions and 3 retained as uncertain (see Methods 4.3.2). Two variants carried high-confidence scores, both EAS-maximum: BRCA2 c.9104A > G (NM_000059.4), with an intermediate consensus score of 0.62, and BRCA2 c.10150C > T (NM_000059.4), scoring within the pathogenic range but with a maximum point AF of 0.02% placing it in the AF-intermediate bin. Neither met the score/AF-bin rule.

Across the 21 variants, allele frequencies were generally low, with a mean maximum AF of 0.059%, within the AF-intermediate bin and therefore uninformative under either directional rule. No variant met the dual criterion for convergent evidence, defined as AF > 0.1% in the enriched reference group with a consensus score < 0.40 for benign-direction support, or AF < 0.01% with a consensus score > 0.70 for pathogenic-direction support (see Methods). All 21 variants therefore remained uncertain under the integrated framework (Fig. 6). The 14 low-confidence benign-range predictions in this subset may nonetheless offer provisional support for re-evaluation when combined with functional or segregation data.

DISCUSSION

This study provides a systematic, quantitative assessment of the BRCA1/2 variant interpretation gap in Indian-source records relative to Chinese-source and European/Caucasian-source comparators. Integrating ClinVar curation status, a six-predictor consensus framework, and population allele frequency (AF) data, we report three findings: Indian-catalogue-specific variants remain substantially under-curated; computational consensus provides supportive evidence for a minority of uninformative variants, and only where biological signal is unambiguous; and variants near interpretive decision boundaries remain unresolved by computational and frequency-based evidence alone.

The curation gap

Only 37.4% of Indian-catalogue-specific variants carried definitive ClinVar classifications. This is broadly comparable with Rony et al. [35], who assessed ClinVar coverage of cohort-unique gnomAD variants across nine breast cancer risk genes and reported 43.4% of South Asian and 30.9% of East Asian variants unreported. Our corresponding absence estimates were lower and more similar to one another (25.3% Indian,

24.4% Chinese), likely reflecting both restriction to BRCA1/2 and disease-focused ascertainment that preferentially captures variants already flagged as clinically relevant. These results extend prior descriptions of genomic diversity disparities [12, 13] by providing a gene-specific, disease-derived quantification of the deficit.

The gap differed qualitatively, not only in magnitude. European/Caucasian-catalogue variants showed near-complete ClinVar coverage (0.4% absent) but the highest conflicting classification rate, consistent either with extensive evaluation exposing genuine interpretive complexity or simply with a larger number of independent submitters. Chinese-catalogue variants showed the highest pathogenic ratio among definitively classified variants, likely reflecting ascertainment from clinically ascertained high-risk hereditary breast and ovarian cancer cohorts, which preferentially report pathogenic and likely pathogenic variants [1, 36]. Indian-source variants remain sparse across both studies and public database submissions, producing lower representation and a broader curation gap [37].

Indian-catalogue-specific variants combined high ClinVar absence, elevated VUS rates and substantial conflicting classifications. This profile is consistent with population specific genetic novelty likely combined with limited ancestry-matched clinical and functional evidence for variant interpretation [38, 39]; the present data cannot distinguish these from differences in source ascertainment or variant composition. Missense variants dominated the uninformative fraction, and are inherently harder to classify than loss-of-function variants under existing frameworks, a difficulty repeatedly demonstrated in the Critical Assessment of Genome Interpretation challenges [28]. Notably, however, the Indian catalogue carried an intermediate missense burden (69.1%) between the European/Caucasian (84.2%) and Chinese (40.8%) catalogues while showing the lowest definitive classification rate of the three. Molecular class composition therefore does not account for the gap, which points to curation infrastructure rather than variant biology as its primary driver.

Consensus scoring as supporting evidence

The six-predictor framework combines supervised predictors (BayesDel_noAF, ClinPred, VEST4) with unsupervised ones (EVE, ESM1b) and a structure-informed model (AlphaMissense) to reduce dependence on any single modelling assumption. Pairwise correlations averaging 0.717, with only three of fifteen pairs above 0.8, support the non-redundancy of this selection. Supervised methods may partially inherit biases from ClinVar-derived training labels, whereas unsupervised tools trained without clinical labels or frequency data contribute an independent evolutionary signal, and their agreement with supervised predictors is compatible with the consensus capturing biological signal rather than replaying clinical annotation. Partial circularity nonetheless remains, since several supervised predictors overlap with ClinVar-based resources; agreement between paradigms provides internal support for score consistency but does not eliminate the underlying bias. AF comparison offered further external support, with pathogenic-range variants showing 3.4-fold lower median maximum AF than benign-range variants, as expected under purifying selection [34].

Confidence tiers should be interpreted asymmetrically. High-confidence pathogenic-range calls were highly sensitive, but their specificity rests on a single benign benchmark variant and is therefore unstable; the benchmark set was approximately 82% pathogenic, so precision estimates are optimistic. Medium-confidence pathogenic-range calls warrant greater caution, with four of six benign variants in that tier

misclassified. Low confidence frequently coincided with benign ClinVar labels, but class imbalance and predictor disagreement preclude treating low confidence as evidence of benignity in its own right.

The single high-confidence false positive illustrates a more interesting failure than a simple error. BRCA2 c.9976A > T p.(Lys3326Ter) was flagged independently by both the concordance analysis and the frequency comparison, the latter as one of only two variants carrying AF above 0.1% despite a pathogenic-range score. Functional studies show no significant effect on cell viability, protein localisation or homologous recombination repair, yet epidemiological evidence supports a modest risk-modifying effect (OR ≈ 1.3), and Thompson et al. classified it as a weak risk modifier [33]. A consensus score of 0.77 may therefore capture a real but modest pathogenicity signal at a genuine biological boundary. Directional consensus scores can thus arise for context-dependent alleles, and borderline scores require independent evidence rather than mechanical thresholding.

Validation remains limited to BRCA1/2, to variants with definitive ClinVar classifications, and to a dataset enriched for pathogenic variants. The consensus score is therefore best interpreted as supportive computational evidence and a research prioritisation tool, not a standalone classifier, and would require gene-specific calibration before any PP3/BP4 application [17].

Why conflicting, AF-differentiated variants remained unresolved

None of the 21 AF-differentiated variants with conflicting ClinVar classifications met the dual-evidence criterion. This describes the yield of the chosen score, dispersion and AF thresholds; it does not establish that computational or population-frequency evidence cannot contribute to their eventual classification.

The result nonetheless illustrates the nature of conflicting classifications. Their allele frequencies clustered in an intermediate range that is too common to strongly support high -penetrance pathogenicity under strict ACMG BA1/BS1 logic, yet too rare to confidently exclude clinical relevance. Both high-confidence variants in the subset, BRCA2 c.9104A > G and BRCA2 c.10150C > T, were EAS-maximum and showed the same intermediate pattern, the latter scoring within the pathogenic range but at an AF of 0.02%. Where evidence is intermediate or discordant, disagreement is unlikely to be resolved by prediction or frequency data alone. That no NFE-maximum variant carried a conflicting classification reinforces the point: the conflicting-classification burden falls disproportionately on variants enriched in under-represented groups, precisely where the supporting evidence is thinnest. The 14 low-confidence benign-range predictions in this subset may still serve as provisional evidence alongside AF context for research prioritisation, and these variants should be prioritised for functional characterisation, segregation studies and reassessment as data accumulate.

Clinical implications for South Asian populations

The framework returned high- or medium-confidence predictions for 75 uninformative Indian-catalogue-specific variants, including 31 pathogenic-range calls among the 42 directional ones. Projection from comparator-catalogue pathogenic proportions suggests the uninformative pool may contain 99–278 additional pathogenic variants, though the upper scenario is unlikely given its missense enrichment; the projection indicates the scale of unresolved classification rather than a count. The higher pathogenic-range

fraction among BRCA1 than BRCA2 predictions (70.4% versus 25.0%) may reflect BRCA1's greater evolutionary constraint and the concentration of essential functional domains including RING and BRCT, or predictor behaviour, and requires independent testing.

In practice, a definitive classification rate of 37.4% means that a patient of Indian ancestry undergoing BRCA1/2 testing is considerably more likely than a patient of European ancestry to receive a variant of uncertain significance. The consequences are concrete: unresolved decisions about risk-reducing surgery, ambiguity in PARP inhibitor eligibility where pathogenic status governs access, and an inability to offer informative cascade testing to relatives, which propagates the deficit across a family.

These findings matter given the early-onset breast cancer burden in India [40] and the limited interpretation infrastructure available for South Asian populations [11]. GenomeAsia 100K [41] and subsequent South Asian whole-genome work have demonstrated fine-scale population structure and founder effects, supporting the use of India- or community-matched reference frequencies rather than reliance on broad SAS estimates [42]; Khanna et al. report germline alterations in cancer-predisposition genes among Indian breast cancers [43]. EthniVar complements this evidence base by assembling reported germline BRCA1/2 variants and identifying those that remain priorities for curation; it does not estimate population prevalence. Priorities include standardised submission of clinical observations, richer source and phenotype metadata, population-matched allele frequencies, segregation evidence and calibrated functional assays.

Strengths and limitations

The principal strength of this study is the integration and normalisation of germline BRCA1/2 variants from multiple Indian disease-focused sources - BCGA, curated literature compilations and targeted literature mining - into a single traceable catalogue, validated with VariantValidator and re-annotated with ANNOVAR to reduce errors from inconsistent nomenclature or transcript mapping. Combining consensus scores with population AF data applies two complementary lines of evidence, an approach increasingly expected in clinical genomics but rarely implemented for under-represented populations. The finding that none of the 21 conflicting, AF-differentiated variants achieved convergent resolution provides a transparent benchmark for future tools, datasets and functional studies.

Two design features carry consequences for interpretation. Variants were assembled from disease-focused studies and curated resources rather than a population-frequency reference, and source population descriptors were taken from contributing literature rather than inferred within a single biobank. This enabled aggregation where no comprehensive Indian resource exists, but introduces heterogeneous ascertainment and reporting practices. Catalogue-specific counts and pathogenic fractions therefore describe the assembled evidence base, not population prevalence, and because some compilations integrate database records rather than primary case reports, we do not claim every record was independently observed in a breast cancer patient.

Further limitations apply. Population descriptors are source-assigned rather than genetically inferred and collapse substantial internal diversity: “Indian” spans a subcontinent of populations distinguished by geography, language group and endogamy, and the European/Caucasian label covers heterogeneous populations including those with documented founder effects; our findings characterise no specific

population. The three catalogues were also constituted by different operations — the Indian by study provenance, the European/Caucasian by keyword matching against free-text source fields, the Chinese as published — so differences between them reflect reporting and curation practice as much as variant spectrum; records with unspecified source population, over half the European/Caucasian source file, were excluded, maximising assignment specificity at the cost of completeness. Validation used ClinVar classifications with two-star or higher review status, the best available standard but not infallible, particularly for variants from the under-represented populations this study concerns. The consensus score is an unweighted average treating all six predictors as equally informative; weighted or Bayesian integration might improve performance but would increase dependence on training-set composition. gnomAD v4.1 frequencies derive from cohorts that may not adequately represent South Asian communities, constraining ACMG frequency criteria for precisely these variants. Finally, some apparent ClinVar absences may reflect inconsistent variant naming in source publications, although systematic under-representation remains the dominant contributor.

CONCLUSION

Variants reported only in Indian sources had substantially lower definitive ClinVar coverage than those in the European/Caucasian and Chinese catalogues (37.4% versus 61.0% and 59.8%), with missense variants comprising most of the uninformative fraction. A six-predictor consensus framework returned high or medium confidence for 75 of 356 uninformative Indian-catalogue-specific variants, of which 42 were directional (31 pathogenic-range, 11 benign-range), and resolved none of the 21 AF-differentiated variants carrying conflicting classifications.

Computational consensus therefore narrows the interpretation burden but does not close the gap. The residual fraction is not a modelling problem but an evidence problem, requiring population-matched allele frequencies, segregation data, ClinVar submission by laboratories serving South Asian patients, and calibrated functional assays. Until that evidence exists, patients of Indian ancestry will continue to receive uncertain results at a materially higher rate than patients of European ancestry undergoing the same test. EthniVar provides a curated, versioned baseline and a prioritised shortlist to direct those efforts (https://brca-ethnivar.vercel.app).

MATERIALS AND METHODS

Data collection

4.1 Population descriptors and terminology

Throughout this study, the labels “Indian,” “Caucasian,” and “Chinese” denote source-assigned population descriptors carried in the originating databases and publications, not genetically inferred ancestry. No genotype-based ancestry inference (for example, principal component projection against a reference panel) was performed, as the compiled records are variant-level and lack the individual-level genotype data such inference requires. We therefore refer to the three groups as the Indian-source, European/Caucasian-source,

and Chinese-source catalogues, and to variants unique to one group as catalogue-specific rather than population-specific.

We retain the source descriptors verbatim to preserve traceability to the primary records. We note their limitations explicitly. “Caucasian” is used here only because it is the descriptor employed in the underlying curated dataset; it functions as an umbrella for a set of heterogeneous European and European-descended populations (see keyword list below) that are genetically distinguishable from one another, including populations with well-characterised founder effects such as Ashkenazi Jewish. “Indian” is a national rather than ancestral descriptor and collapses substantial documented substructure arising from geography, language group and endogamy across the Indian subcontinent. “Chinese” is similarly a national descriptor applied here as inherited from the source dataset.

The three catalogues were assembled by three different operations, and are comparable in the sense that each reflects how variants from that population have entered the shared evidence base - which is precisely the object of study here - but they are not matched cohorts. The Indian catalogue was defined by study provenance (variants reported in studies of patients in India); the European/Caucasian catalogue by keyword matching against free-text source-population fields; and the Chinese catalogue was adopted as published. Our central comparison is therefore a comparison of curation status across three differently constituted evidence streams, not a comparison of allele architecture between genetically defined populations. We report this asymmetry as a limitation in the Discussion.

Our framing follows the recommendations of the National Academies of Sciences, Engineering, and Medicine (2023) report on population descriptors in genetics and genomics [44], in that we define each descriptor operationally, state the process by which it was assigned, and avoid treating the descriptors as proxies for biological categories.

4.1.1 Indian source compilation

To represent the germline *BRCA* variants reported in Indian population, we compiled variants from two pre-existing datasets and supplemented them with updated literature mining to capture recently reported variants.

1. First, germline data were obtained from the Bharat Cancer Genome Atlas (BCGA) [45] which includes data from Bharat Cancer Germline (IITM 2025) and triple-negative breast cancer germline data from Mizoram University [46]. Using cBioPortal (https://bcga.iitm.ac.in/cbioportal/) *BRCA1* and *BRCA2* were queried, yielding 1,761 sample-wise *BRCA1* records and 1149 sample-wise *BRCA2* records (Supplementary File: P_Mutants.xlsx; sheets: “Raw BRCA1_BCGA” and “Raw BRCA2_BCGA”). After filtering for unique variant call format (VCF) records, 52 distinct *BRCA1* mutations and 94 distinct *BRCA2* mutations were retained (*Supplementary File: P_Mutants.xlsx; sheet: “BCGA Final”*).
2. Second, we incorporated variants from Bhaskaran et al.[29], which compiled and curated published literature on Indian *BRCA1/2* variants and reported t 260 *BRCA1* and 220 *BRCA2* variants.
3. Third, to capture additional variants, particularly those reported after the publication period covered by Bhaskaran et al. [29], we performed a PubMed search “BRCA variants India”. Abstracts were screened to

identify studies reporting germline BRCA1/2 variants in breast cancer patients. The selected studies corresponded to PMIDs: 30555256, 29470806, 32963034, 33552952, 35020120, 39506289, 38538877, 38315150, 39733403, 38439815 [29, 47–55] and are listed in *Supplementary File: P_Mutants.xlsx, Sheet: "Verma et al ConsideredStudies*". From these publications, variant-level information was extracted, including sequence accessions and genome assembly versions where available (*Supplementary File: P_Mutants.xlsx; Sheet: "Verma et al CompiledData"*). These variants were subsequently subjected to validation and normalization as described in the section "Variant validation and normalization".

4.1.2 Caucasian source compilation

To compile Caucasian *BRCA1/2* variants, we used the dataset curated by Bhaskaran et al [36], which was originally assembled for comparative analysis with Chinese and other Asian variants. Variant information was extracted from the supporting data file, IJC-145-962-s006, which integrates records from multiple sources, including BIC [56], BMD, ClinVar [57] LOVD [58], and ENIGMA [59]. Conservative filtering criteria were then applied to identify Caucasian-specific variants.

Variants were classified as Caucasian only when all associated source-population descriptors matched predefined Caucasian keywords and did not contain any predefined non-Caucasian keywords. Records with source-population values such as "none-specified," "unknown," "not specified," "-", or empty fields were excluded; these accounted for more than 50% of entries. However, variants were retained as Caucasian if "-" co-occurred with a Caucasian keyword in the source-population annotation. This filtering process yielded 1,747 BRCA1/2 variants, comprising 846 BRCA1 and 901 BRCA2 variants. (*Supplementary File: P_Mutants.xlsx, Sheet: "caucasianData")*

Caucasian keywords: *austrian, ashkenazi, ashkenazi jewish, british, bulgarian, caucasian, central europe, central/eastern europe, central/eastern european, dutch, jewish, eastern europe,eastern european, european, england, english, french, finland, german, dutch/german, galician, hungarian, irish, italian, italy, mediterranean, north european, nw european, norway, norwegian, polish, serbian, scottish, scot, greek, polish, west swedish, western swedish, portuguese, swedish, white-nonhispanic, white nonhispanic, welsh, white north american, western europe, western european and westerneuropean*.

Non-caucasian keywords: *arab, australia, afric, africa, african, asian, latin, carib, caribbean, native, near east, middle east, hawai, pacific, pacific islands, filip, hispanic/latino, china, japan, korea, mexican, hispanic, puerto, spanish, colombian, pakistan, india, malay, samoan, vietnam*.

Records with unspecified source-population values accounted for more than 50% of entries in the source file and were excluded. This conservative filter maximises specificity of the European/Caucasian assignment at the cost of sensitivity, and means the resulting catalogue is not an exhaustive inventory of variants reported in European-ancestry individuals.

4.1.3 Chinese source compilation

For the Chinese cohort, we sourced supplementary data file 3 of Bhaskaran et al's[29], which lists 748 *BRCA1* and 755 *BRCA2* variants. (*Supplementary File: P_Mutants.xlsx, Spreadsheet: "chinaData")*

Variant validation and normalization

https://rest.variantvalidator.org/VariantValidator/variantvalidator/[Assembly]/[mRNA/chromsomeAccession]:[respective_HGVS]/[mRNA_accession]?content-type=application/json

To enable cross-source-population comparison, all variants were reannotated and normalized to the hg38 genome assembly. Variant records were first validated using the VariantValidator REST API [60, 61], with following endpoint:

During validation, variants with inconsistencies in the mRNA transcript, reference allele sequence, cDNA HGVS description, or genome assembly were flagged as validation failures. Error reports generated by VariantValidator were reviewed manually, and minor typographical errors in mRNA transcript identifiers or cDNA Human Genome Variation Society (HGVS) nomenclature were corrected for the Indian cohort where the intended variant could be unambiguously inferred. (see *Supplementary File: P_Mutants.xlsx, Sheet: "Verma et al modifications"*). Details of corrections are also listed along with unresolvable cases in *Supplementary Text S1*.

table_annovar.pl vcfFile.vcf data/annovar/humandb/ -buildver hg38 -out output/ -protocol refGeneWithVer, clinvar_20250721, gnomad41_genome, avsnp151, dbnsfp47a -operation g,f,f,f,f -arg '-hgvs',,,, -nastring. -vcfinput runlog 2>&1

The resulting JSON outputs were deduplicated using hg38-specific VCF attributes, including chromosome, genomic position, reference allele, and alternate allele. Non-redundant variants were then reannotated using ANNOVAR [62, 63] with the command described below÷:

After filtering and annotation, the overall variant counts are summarized in Table 1.

Table 1

Summary of BRCA germline variant counts from multiple sources. This table presents the total number of variants in each dataset, along with counts of the unique and validated and distinct and curated subsets. "unique and validated" variants were determined from the total using gene name, mRNA sequence, cDNA HGVS and genome assembly and VariantValidator validation. The "distinct and curated" column reflects the final count after removing duplicates informed by VCF attributes and reannotation using ANNOVAR.

Source population	Dataset	*BRCA1* (total, (unique and validated), [distinct and curated])	*BRCA2* (total, (unique and validated), [distinct and curated])
Indian	BCGA	1761, (52), [52]	1149, (94), [94]
	Bhaskaran et al 2021 [29] (India)	260, (253), [246]	220, (210), [204]
	Self-compilation (*)	484 (313) [235]	271, (217) [198]
Caucasian	Bhaskaran et al 2019 [36] (§)	846 (840) [713]	901 (895) [785]
Chinese	Bhaskaran et al 2021 [29] (China) (¥)	748 (723) [719]	755 (722) [680]

(*) In self-compilation dataset (*Verma et al compiledData)*, after correcting formatting and typographical errors (recorded in *Verma et al modifications, with contextual explanation in Supplementary Text S1)* 319 *BRCA1* and 217 *BRCA2* variants remained. These variants were then validated (5 variants failed) and filtered on VCF attributes (Chromosome, position, ref sequence, alt sequence) to yield the unique counts in the respective table entries.

(**§**) Bhaskaran et al 2019 [36]: we were able to get 840 BRCA1 (713 unique) and 895 BRCA2 (784 unique) entries. The difference in counts arises from duplicate entries, for instance, c.3919_3919delG and c.3919delG indicate the same in BRCA2. (Failure details can be found in *Supplementary Text S3*).

(**¥**) Bhaskaran et al 2021 (China) [29]: We were able to validate 723 entries for BRCA1 (708 unique), and 722 for *BRCA2* (661 unique), Failure details can be found in *Supplementary Text S4*).

Variant Annotation and Allele Frequency Extraction

1.1.1 Consensus predictor selection and scoring framework

ANNOVAR annotation using dbNSFP 4.7a protocol provides scores from multiple variant-effect predictors, including individual predictors and meta-predictors such as REVEL, which itself incorporates several component predictors. Recent ClinGen and ACMG/AMP-related guidance has emphasized that computational evidence for PP3/BP4 should be based on appropriately calibrated tools and may incorporate newer predictors with state-of-the-art model architectures [64–66]. Accordingly, a consensus scoring framework was implemented using six Tier-1 predictors from dbNSFP v4.7a. These predictors were selected to capture methodological diversity across supervised and unsupervised paradigms and to reduce potential shared biases arising from training datasets dominated by well-represented populations, particularly well-represented ancestry groups, as discussed by Rastogi et al [28].

The six Tier-1 predictors were as follows:

- *BayesDel_noAF (meta-predictor)*: BayesDel_noAF is a Bayesian meta-predictor that integrates scores from multiple component tools, including SIFT, PolyPhen, and conservation-based features [31]. We used the no-allele-frequency version to reduce circularity and reduce rare-variant training bias that may arise from allele frequency differences in well-represented populations. BayesDel was included because it is the ENIGMA-specified predictor for *BRCA1/2* missense classification [17]. Previous benchmarking by Tian et al [67] also showed that BayesDel outperformed CADD, MetaSVM, and Eigen, and performed comparably to REVEL (86.4% consensus), with modest advantages in discordant cases.
- *AlphaMissense*: AlphaMissense predicts the impact of missense variants using protein structure and sequence context informed by AlphaFold-based modelling [32]. Its inclusion was motivated by its relevance for variants affecting structured regions of DNA damage repair proteins. Huang et al. reported strong performance of AlphaMissense in interpreting saturation genome editing data from the BRCA2 DNA-binding domain, where it performed slightly better than BayesDel when compared with clinical and functional data ([68].
- *EVE (Evolutionary model of Variant Effect)*. EVE is an unsupervised evolutionary model of variant effect based on a variational autoencoder trained on protein-specific multiple sequence alignments (gene

wise) from protein sequences of over 140,000 species. Because it is trained without clinical labels, EVE provides a predictor that is less susceptible to ClinVar circularity or population-label bias, while also providing probabilistic uncertainty estimates [22].

- *ESM1b (Evolutionary Scale Modeling-1b)*: ESM1b is a 650-million parameter transformer-based protein language model trained on large-scale protein sequence data (250 million protein sequences) and predicts variant effects without labelled pathogenic or benign training examples [23]. Together with EVE, ESM1b provides an unsupervised evolutionary signal that complements supervised predictors and helps reduce dependence on clinically curated datasets.
- *ClinPred*: ClinPred is an ensemble machine-learning predictor trained using ClinVar classifications and multiple engineered features [20]. Although incorporation of population-frequency information may introduce bias against underrepresented populations, ClinPred was included because it can capture clinically curated patterns and may provide information for variants in non-conserved regions that may be missed by purely evolutionary models.
- *VEST4 (Variant Effect Scoring Tool)*: VEST4 is a random forest-based predictor trained using HGMD variants and incorporates physical and evolutionary protein features to identify rare damaging variants [21]. We included VEST4 to capture functional deleteriousness with high sensitivity, despite the possibility of HGMD training-set bias. Its inclusion is also consistent with recent ACMG calibration studies that classify VEST4 among predictors capable of contributing strong computational evidence under appropriate calibration [65].

For each Tier-1 predictor, we used the corresponding '*_rankscore' field from dbNSFP v4.7a. These rank scores represent percentile-normalized values on a 0–1 scale, where values closer to 1 indicate greater predicted pathogenicity and values closer to 0 indicate a more benign prediction. For each variant, a consensus score was calculated as the unweighted mean of available Tier-1 rank scores, provided that valid scores were available from at least two of the six predictors.

To quantify agreement among contributing predictors, we calculated the coefficient of variation (CV; standard deviation divided by the mean) across available Tier-1 scores for each variant. Prediction confidence was assigned as follows: high confidence, CV < 0.15; medium confidence, 0.15 ≤ CV < 0.30; and low confidence, CV ≥ 0.30. Variants were then classified based on their mean consensus score as pathogenic-range if the score was > 0.70, intermediate if the score was ≥ 0.40 and ≤ 0.70, and benign-range if the score was ≤ 0.40.

In addition to the Tier-1 predictors used for consensus scoring, we evaluated variants using a set of previously well-established Tier-2 predictors to provide broader interpretive context. These include REVEL, MetaRNN, VARITY_R_LOO, SIFT, Polyphen2_HDIV, CADD_phred, PrimateAI, GERP++_RS, phyloP100way_vertebrate and phastCons100way_vertebrate. For all Tier-2 predictors except CADD_phred, normalized rank scores were used in the same manner as for Tier-1 predictors. Because CADD_phred is reported as a scaled logarithmic ranking, values were normalized by dividing by 35 and values greater than 1 were capped at 1. Tier-2 predictors were not included in the consensus score calculation but were retained in the accompanying data resource to support visual inspection and interpretation of individual variants.

4.3.2 Allele frequency data and validation

Population allele frequency (AF) data were extracted from gnomAD v4.1 annotations. AF denotes the reported gnomAD estimate and was not further processed using allele count, allele numbers, or confidence intervals. Non-Finnish European (NFE), South Asian (SAS), and East Asian (EAS) frequencies were used as ancestry proxies for Caucasian, Indian, and Chinese populations, respectively.

Consensus prediction performance was evaluated against definitive ClinVar classifications, defined as pathogenic/likely pathogenic or benign/likely benign variants with a review status of ≥ 2 stars. This validation set included 631 variants, comprising 493 pathogenic or likely pathogenic variants and 138 benign or likely benign variants. Variants with consensus scores in the intermediate range of 0.40–0.70 were excluded from binary performance evaluation. Accuracy, sensitivity, specificity, and precision were then calculated overall and within each confidence stratum. Called set metrics used 575 directional calls; receiver operating characteristic (ROC) curves and area under the curve (AUC) values used all 631 continuous scores and were calculated for the consensus score and for each individual predictor using scikit-learn v1.8.0.

For external comparison, consensus scores were evaluated against gnomAD v4.1 population AFs. For each variant the maximum point AF across NFE, SAS, and EAS was used. Maximum point AF was described as AF low (< 0.01%), AF-intermediate (0.01%-0.1%, inclusive) or AF-high (> 0.1%). These study specific bins were used to describe whether AF and score direction were concordant, or insufficient for directional interpretation. In particular, rarity alone was not considered evidence of pathogenicity.

For integrated assessment of the 21 conflicting, population-enriched variants, consensus predictions and AF evidence were combined using the following rule-based framework.:

- Convergent_benign: AF high AND benign-range prediction to indicate benign-direction evidence.
- Convergent_pathogenic: AF low AND pathogenic-range prediction to indicate concordant pathogenic-direction evidence.
- Discordant_high_AF_prediction_pathogenic: AF high AND pathogenic-range prediction, needs further evaluation.
- Low_AF_prediction_benign: AF low AND benign-range prediction, needs further evaluation. This combination is not intrinsically discordant, because benign variants can also be rare.
- Uncertain: Variants with low confidence (high dispersion) consensus predictions or intermediate consensus scores, defined as 0.40–0.70.

4.4 Database design and web interface

We have made the curated variants in the EthniVar dataset openly accessible. We implemented a lightweight relational database and public web interface (https://brca-ethnivar.vercel.app) in JavaScript and SQLite3 via the synchronous better-sqlite3 binding (v9.6). The resource is designed for updates as additional centralized sources become available. This application was developed and deployed with the help of Anthropic-Claude opus 4.7 and verified by the authors for data accuracy and its representation.

4.5 Data and code availability.

The interactive web resource is at https://brca-ethnivar.vercel.app. A static snapshot of the curated variant tables is provided as Supplementary File: P_Mutants.xlsx.

Declarations

Competing interests

The authors declare no competing interests.

Funding Declaration

P.V. acknowledge postdoctoral fellowship support and S.H. and M.S. acknowledge research support from Plaksha University with intramural grants - OOF/CEPH/2023-24/11 and OOR/BBIIF/2026-27/2. The funders had no role in designing the study, collecting the data, or the decision to submit the work for publication.

Author Contribution

M.S. and S.H. conceived and supervised the study. P.V. developed the framework, analysis, contributed to dataset curation. P.V. and M.S. contributed to drafting the main manuscript text. All authors approved the final version.

Acknowledgement

Authors thank Plaksha University for funding of research staff and infrastructural support.

Data Availability

The interactive web resource is at https://brca-ethnivar.vercel.app. A static snapshot of the curated variant tables is provided as Supplementary File: P_Mutants.xlsx.

References

1. Ciesielski TH, Sirugo G, Iyengar SK, Williams SM. Characterizing the pathogenicity of genetic variants: the consequences of context. NPJ Genomic Med. 2024;9:3. https://doi.org/10.1038/s41525-023-00386-5.
2. Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips K-A, Mooij TM, Roos-Blom M-J, et al. Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. JAMA. 2017;317:2402–16. https://doi.org/10.1001/jama.2017.7112.
3. Rebbeck TR, Friebel TM, Friedman E, Hamann U, Huo D, Kwong A, et al. Mutational spectrum in a worldwide study of 29,700 families with BRCA1 or BRCA2 mutations. Hum Mutat. 2018;39:593–620.

https://doi.org/10.1002/humu.23406.

4. Robson M, Im S-A, Senkus E, Xu B, Domchek SM, Masuda N, et al. Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation. N Engl J Med. 2017;377:523–33. https://doi.org/10.1056/NEJMoa1706450.
5. Pensabene M, Calabrese A, von Arx C, Caputo R, De Laurentiis M. Cancer genetic counselling for hereditary breast cancer in the era of precision oncology. Cancer Treat Rev. 2024;125:102702. https://doi.org/10.1016/j.ctrv.2024.102702.
6. Naslavsky MS, Scliar MO, Nunes K, Wang JYT, Yamamoto GL, Guio H, et al. Biased pathogenic assertions of loss of function variants challenge molecular diagnosis of admixed individuals. Am J Med Genet C Semin Med Genet. 2021;187:357–63. https://doi.org/10.1002/ajmg.c.31931.
7. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: Lessons from gnomAD. Hum Mutat. 2022;43:1012–30. https://doi.org/10.1002/humu.24309.
8. Wang SM. A global perspective on the ethnic-specific BRCA variation and its implication in clinical application. J Natl Cancer Cent. 2023;3:14–20. https://doi.org/10.1016/j.jncc.2022.12.001.
9. Appelbaum PS, Burke W, Parens E, Zeevi DA, Arbour L, Garrison NA, et al. Is there a way to reduce the inequity in variant interpretation on the basis of ancestry? Am J Hum Genet. 2022;109:981–8. https://doi.org/10.1016/j.ajhg.2022.04.012.
10. Agaoglu NB, Unal B, Hayes CP, Walker M, Ng OH, Doganay L, et al. Genomic disparity impacts variant classification of cancer susceptibility genes in Turkish breast cancer patients. Cancer Med. 2024;13:e6852. https://doi.org/10.1002/cam4.6852.
11. Jadeja N, Rajakumar N, Reddy N, Ali N, Lichten L. Reflections on my international genetic counseling rotations: Contrasts in practice between India and the United States. Genet Med Open. 2024;2:101871. https://doi.org/10.1016/j.gimo.2024.101871.
12. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538:161–4. https://doi.org/10.1038/538161a.
13. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019;177:1080. https://doi.org/10.1016/j.cell.2019.04.032.
14. Fatumo S, Chikowore T, Choudhury A, Ayub M, Martin AR, Kuchenbaecker K. A roadmap to increase diversity in genomic studies. Nat Med. 2022;28:243–50. https://doi.org/10.1038/s41591-021-01672-4.
15. Tavtigian SV, Byrnes GB, Goldgar DE, Thomas A. Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications. Hum Mutat. 2008;29:1342–54. https://doi.org/10.1002/humu.20896.
16. Balabanski L, Antov G, Dimova I, Ivanov S, Nacheva M, Gavrilov I, et al. Next-generation sequencing of BRCA1 and BRCA2 in breast cancer patients and control subjects. Mol Clin Oncol. 2014;2:435–9. https://doi.org/10.3892/mco.2014.251.
17. Parsons MT, de la Hoya M, Richardson ME, Tudini E, Anderson M, Berkofsky-Fessler W, et al. Evidence-based recommendations for gene-specific ACMG/AMP variant classification from the ClinGen ENIGMA

BRCA1 and BRCA2 Variant Curation Expert Panel. Am J Hum Genet. 2024;111:2044–58. https://doi.org/10.1016/j.ajhg.2024.07.013.

18. Dines JN, Shirts BH, Slavin TP, Walsh T, King M-C, Fowler DM, et al. Systematic misclassification of missense variants in BRCA1 and BRCA2 coldspots. Genet Med Off J Am Coll Med Genet. 2020;22:825–30. https://doi.org/10.1038/s41436-019-0740-6.
19. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic Misdiagnoses and the Potential for Health Disparities. N Engl J Med. 2016;375:655–65. https://doi.org/10.1056/NEJMsa1507092.
20. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. Am J Hum Genet. 2018;103:474–83. https://doi.org/10.1016/j.ajhg.2018.08.005.
21. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14(Suppl 3):S3. https://doi.org/10.1186/1471-2164-14-S3-S3.
22. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599:91–5. https://doi.org/10.1038/s41586-021-04043-8.
23. Brandes N, Goldman G, Wang CH, Ye CJ, Ntranos V. Genome-wide prediction of disease variant effects with a deep protein language model. Nat Genet. 2023;55:1512–22. https://doi.org/10.1038/s41588-023-01465-0.
24. Wang X, Zhang T, Qin Y, Zou Y, Luo H, Li H, et al. Assessing Allele Frequency Information: A Study of Variant Curation Expert Panel Guidelines. Hum Mutat. 2026;2026:1611484. https://doi.org/10.1155/humu/1611484.
25. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med Off J Am Coll Med Genet. 2015;17:405–24. https://doi.org/10.1038/gim.2015.30.
26. Ji P, Gong Y, Jin M-L, Hu X, Di G-H, Shao Z-M. The Burden and Trends of Breast Cancer From 1990 to 2017 at the Global, Regional, and National Levels: Results From the Global Burden of Disease Study 2017. Front Oncol. 2020;10:650. https://doi.org/10.3389/fonc.2020.00650.
27. Chaturvedi M, Vaitheeswaran K, Satishkumar K, Das P, Stephen S, Nandakumar A. Time Trends in Breast Cancer Among Indian Women Population: An Analysis of Population Based Cancer Registry Data. Indian J Surg Oncol. 2015;6:427–34. https://doi.org/10.1007/s13193-015-0467-z.
28. Rastogi R, Chung R, Li S, Li C, Lee K, Woo J, et al. Critical assessment of missense variant effect predictors on disease-relevant variant data. Hum Genet. 2025;144:281–93. https://doi.org/10.1007/s00439-025-02732-2.
29. Bhaskaran SP, Huang T, Rajendran BK, Guo M, Luo J, Qin Z, et al. Ethnic-specific BRCA1/2 variation within Asia population: evidence from over 78 000 cancer and 40 000 non-cancer cases of Indian, Chinese, Korean and Japanese populations. J Med Genet. 2021;58:752–9. https://doi.org/10.1136/jmedgenet-2020-107299.

30. Jain A, Bhoyar RC, Pandhare K, Mishra A, Sharma D, Imran M, et al. IndiGenomes: a comprehensive resource of genetic variants from over 1000 Indian genomes. Nucleic Acids Res. 2021;49:D1225–32. https://doi.org/10.1093/nar/gkaa923.
31. Feng B-JPERCH. A Unified Framework for Disease Gene Prioritization. Hum Mutat. 2017;38:243–51. https://doi.org/10.1002/humu.23158.
32. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Volume 381. New York, N: Science; 2023. p. eadg7492. https://doi.org/10.1126/science.adg7492.
33. Thompson ER, Gorringe KL, Rowley SM, Li N, McInerny S, Wong-Brown MW, et al. Reevaluation of the BRCA2 truncating allele c.9976A > T (p.Lys3326Ter) in a familial breast cancer context. Sci Rep. 2015;5:14800. https://doi.org/10.1038/srep14800.
34. Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, et al. Using high-resolution variant frequencies to empower clinical genome interpretation. Genet Med Off J Am Coll Med Genet. 2017;19:1151–8. https://doi.org/10.1038/gim.2017.26.
35. Rony R, Deng S, Yang S, Doig K, Goode DL. Putative breast cancer risk variants from populations of South Asian ancestry are under-represented in public variant classification databases. Breast Cancer Res BCR. 2026;28:5. https://doi.org/10.1186/s13058-025-02166-8.
36. Bhaskaran SP, Chandratre K, Gupta H, Zhang L, Wang X, Cui J, et al. Germline variation in BRCA1/2 is highly ethnic-specific: Evidence from over 30,000 Chinese hereditary breast and ovarian cancer patients. Int J Cancer. 2019;145:962–73. https://doi.org/10.1002/ijc.32176.
37. Narang A, Uppilli B, Vivekanand A, Naushin S, Yadav A, Singhal K, et al. Frequency spectrum of rare and clinically relevant markers in multiethnic Indian populations (ClinIndb): A resource for genomic medicine in India. Hum Mutat. 2020;41:1833–47. https://doi.org/10.1002/humu.24102.
38. Qin Z, Li J, Tam B, Sinha S, Zhao B, Bhaskaran SP, et al. Ethnic-specificity, evolution origin and deleteriousness of Asian BRCA variation revealed by over 7500 BRCA variants derived from Asian population. Int J Cancer. 2023;152:1159–73. https://doi.org/10.1002/ijc.34359.
39. Sharma-Oates A, Shaaban AM, Tomlinson I, Wynne L, Cazier J-B, Sundar S. Heterogeneity of germline variants in high risk breast and ovarian cancer susceptibility genes in India. Precis Clin Med. 2018;1:75–87. https://doi.org/10.1093/pcmedi/pby010.
40. Malvia S, Bagadi SA, Dubey US, Saxena S. Epidemiology of breast cancer in Indian women. Asia Pac J Clin Oncol. 2017;13:289–95. https://doi.org/10.1111/ajco.12661.
41. GenomeAsia100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature. 2019;576:106–11. https://doi.org/10.1038/s41586-019-1793-z.
42. Wall JD, Sathirapongsasuti JF, Gupta R, Rasheed A, Venkatesan R, Belsare S, et al. South Asian medical cohorts reveal strong founder effects and high rates of homozygosity. Nat Commun. 2023;14:3377. https://doi.org/10.1038/s41467-023-38766-1.
43. Khanna D, Ghosh A, Bhadwal P, Chirravuri-Venkata R, Dhingra L, Das S et al. Whole-Genome Landscape of Breast Cancers from India shows Distinct Clinically Actionable Subtypes [Internet]. Genetic and Genomic Medicine; 2026 [cited 2026 Sept 8]. https://doi.org/10.64898/2026.03.27.26349506

44. National Academies of Sciences, Engineering, and Medicine, Division of Behavioral and Social Sciences and Education; Health and Medicine Division; Committee on Population; Board on Health Sciences Policy. ; Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research. Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field [Internet]. Washington (DC): National Academies Press (US); 2023 [cited 2026 Sept 9]. http://www.ncbi.nlm.nih.gov/books/NBK589855/. Accessed 9 Sept 2026.
45. Mahalingam S, Scaria V, Sivasubbu S. The Bharat Cancer Genome Atlas: Charting India's Unique Cancer Landscape for Precision Oncology. Technol Cancer Res Treat. 2025;24:15330338251381404. https://doi.org/10.1177/15330338251381404.
46. Pachuau L, Lalremmawia H, Ralte L, Vanlalpeka J, Pautu JL, Chenkual S, et al. Uncovering novel pathogenic variants and pathway mutations in triple-negative breast cancer among the endogamous mizo tribe. Breast Cancer Res Treat. 2025;209:375–87. https://doi.org/10.1007/s10549-024-07501-9.
47. Mehta A, Vasudevan S, Sharma SK, Kumar D, Panigrahi M, Suryavanshi M, et al. Germline BRCA1 and BRCA2 deleterious mutations and variants of unknown clinical significance associated with breast/ovarian cancer: a report from North India. Cancer Manag Res. 2018;10:6505–16. https://doi.org/10.2147/CMAR.S186563.
48. Singh J, Thota N, Singh S, Padhi S, Mohan P, Deshwal S, et al. Screening of over 1000 Indian patients with breast and/or ovarian cancer with a multi-gene panel: prevalence of BRCA1/2 and non-BRCA mutations. Breast Cancer Res Treat. 2018;170:189–96. https://doi.org/10.1007/s10549-018-4726-x.
49. Kadri MSN, Patel KM, Bhargava PA, Shah FD, Badgujar NV, Tarapara BV, et al. Mutational Landscape for Indian Hereditary Breast and Ovarian Cancer Cohort Suggests Need for Identifying Population Specific Genes and Biomarkers for Screening. Front Oncol. 2020;10:568786. https://doi.org/10.3389/fonc.2020.568786.
50. Rajagopal T, Seshachalam A, Jothi A, Rathnam KK, Talluri S, Venkatabalasubranian S, et al. Analysis of pathogenic variants in BRCA1 and BRCA2 genes using next-generation sequencing in women with triple negative breast cancer from South India. Mol Biol Rep. 2022;49:3025–32. https://doi.org/10.1007/s11033-022-07129-2.
51. Gupta P, Thakur T, Chadda A, Irinike S, Khare S, Bal A. A Pilot Study on BRCA1/2 and PI3K Mutations Across Subtypes of Triple Negative Breast Cancer in North Indian Population. Appl Immunohistochem Mol Morphol AIMM. 2024;32:462–8. https://doi.org/10.1097/PAI.0000000000001231.
52. John AO, Singh A, Yadav P, Joel A, Thumaty DB, Fibi Ninan K, et al. The BRCA mutation spectrum among breast and ovarian cancers in India: highlighting the need to screen BRCA1 185delAG among South Indians. Eur J Hum Genet EJHG. 2024;32:1319–26. https://doi.org/10.1038/s41431-024-01596-w.
53. Gardi N, Chaubal R, Parab P, Pachakar S, Kulkarni S, Shet T, et al. Natural History of Germline BRCA1 Mutated and BRCA Wild-type Triple-negative Breast Cancer. Cancer Res Commun. 2024;4:404–17. https://doi.org/10.1158/2767-9764.CRC-23-0277.
54. Lila K, Bhanushali H, Chanekar Mi, Jatale R, Banerjee M, Dixit RB, et al. Mutation Spectrum Analysis of BRCA1/2 Genes for Hereditary Breast and Ovarian Cancer in the Indian Population. Asian Pac J Cancer Prev APJCP. 2024;25:4145–51. https://doi.org/10.31557/APJCP.2024.25.12.4145.

55. Pramanik R, Chitikela S, Deo SVS, Gogia A, Batra A, Kumar A, et al. Comprehensive germline profiling of patients with breast cancer: initial experience from a Familial Cancer Clinic. Ecancermedicalscience. 2024;18:1670. https://doi.org/10.3332/ecancer.2024.1670.
56. Szabo C, Masiello A, Ryan JF, Brody LC. The breast cancer information core: database design, structure, and scope. Hum Mutat. 2000;16:123–31. https://doi.org/10.1002/1098-1004(200008)16:2%3C123.::AID-HUMU4%3E3.0.CO;2-Y.
57. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42:D980–985. https://doi.org/10.1093/nar/gkt1113.
58. Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat. 2011;32:557–63. https://doi.org/10.1002/humu.21438.
59. Spurdle AB, Healey S, Devereau A, Hogervorst FBL, Monteiro ANA, Nathanson KL, et al. ENIGMA–evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. Hum Mutat. 2012;33:2–7. https://doi.org/10.1002/humu.21628.
60. Freeman PJ, Hart RK, Gretton LJ, Brookes AJ, Dalgleish R, VariantValidator. Accurate validation, mapping, and formatting of sequence variation descriptions. Hum Mutat. 2018;39:61–8. https://doi.org/10.1002/humu.23348.
61. Freeman PJ, Wagstaff JF, Fokkema IFAC, Cutting GR, Rehm HL, Davies AC, et al. Standardizing variant naming in literature with VariantValidator to increase diagnostic rates. Nat Genet. 2024;56:2284–6. https://doi.org/10.1038/s41588-024-01938-w.
62. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164. https://doi.org/10.1093/nar/gkq603.
63. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc. 2015;10:1556–66. https://doi.org/10.1038/nprot.2015.105.
64. Bergquist T, Stenton SL, Nadeau EAW, Byrne AB, Greenblatt MS, Harrison SM, et al. Calibration of additional computational tools expands ClinGen recommendation options for variant classification with PP3/BP4 criteria. Genet Med Off J Am Coll Med Genet. 2025;27:101402. https://doi.org/10.1016/j.gim.2025.101402.
65. Pejaver V, Byrne AB, Feng B-J, Pagel KA, Mooney SD, Karchin R, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am J Hum Genet. 2022;109:2163–77. https://doi.org/10.1016/j.ajhg.2022.10.013.
66. Stenton SL, Pejaver V, Bergquist T, Biesecker LG, Byrne AB, Nadeau EAW, et al. Assessment of the evidence yield for the calibrated PP3/BP4 computational recommendations. Genet Med Off J Am Coll Med Genet. 2024;26:101213. https://doi.org/10.1016/j.gim.2024.101213.
67. Tian Y, Pesaran T, Chamberlin A, Fenwick RB, Li S, Gau C-L, et al. REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. Sci Rep. 2019;9:12752. https://doi.org/10.1038/s41598-019-49224-8.

68. Huang H, Hu C, Na J, Hart SN, Gnanaolivu RD, Abozaid M, et al. Functional evaluation and clinical classification of BRCA2 variants. Nature. 2025;638:528–37. https://doi.org/10.1038/s41586-024-08388-8.

Figures

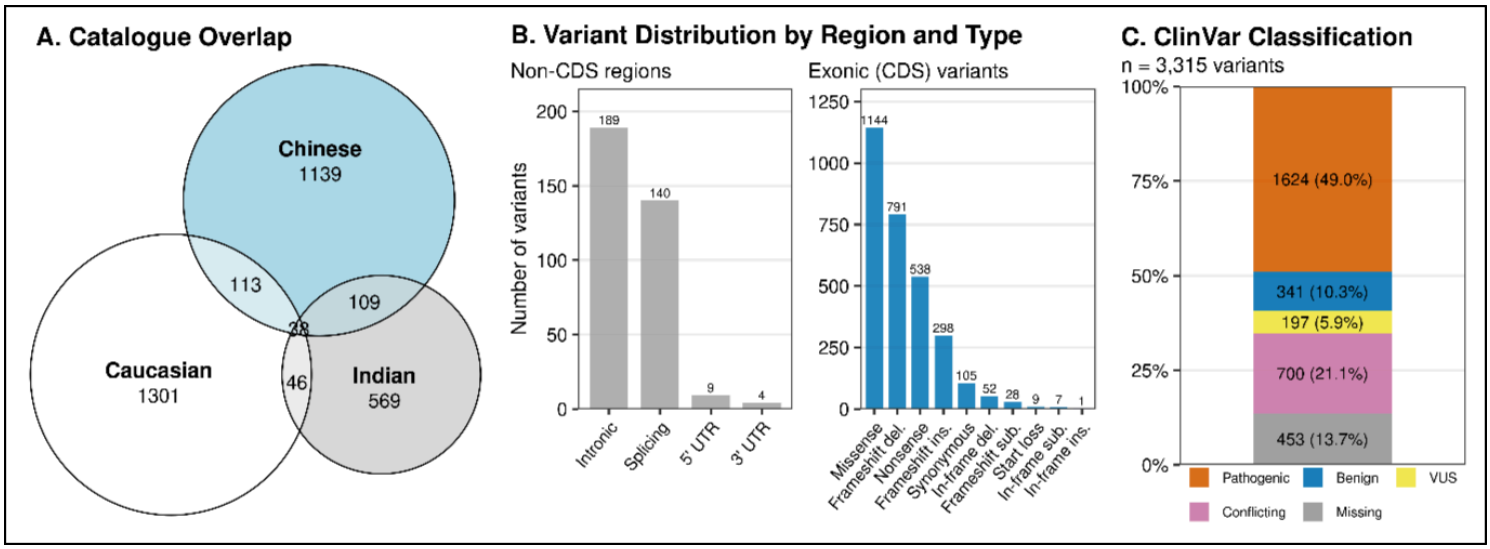


Figure 1

EthniVar dataset overview of BRCA1/2 germline variants across three populations. A) Venn diagram showing distribution of 3,315 *BRCA1/2* germline variants across European/Caucasian, Indian, and Chinese catalogues. **B)** Distribution of variants by genomic region, highlighting non-exonic variants, which comprised approximately 10% of the total dataset; and exonic variants (n=2973), stratified by molecular consequence. **D)** Stacked bar plot showing the distribution of ClinVar annotation status across all 3,315 variants. Pathogenic includes likely pathogenic variants and accounted for 48.9% of the dataset (n = 1,624); benign similarly includes likely benign variants.

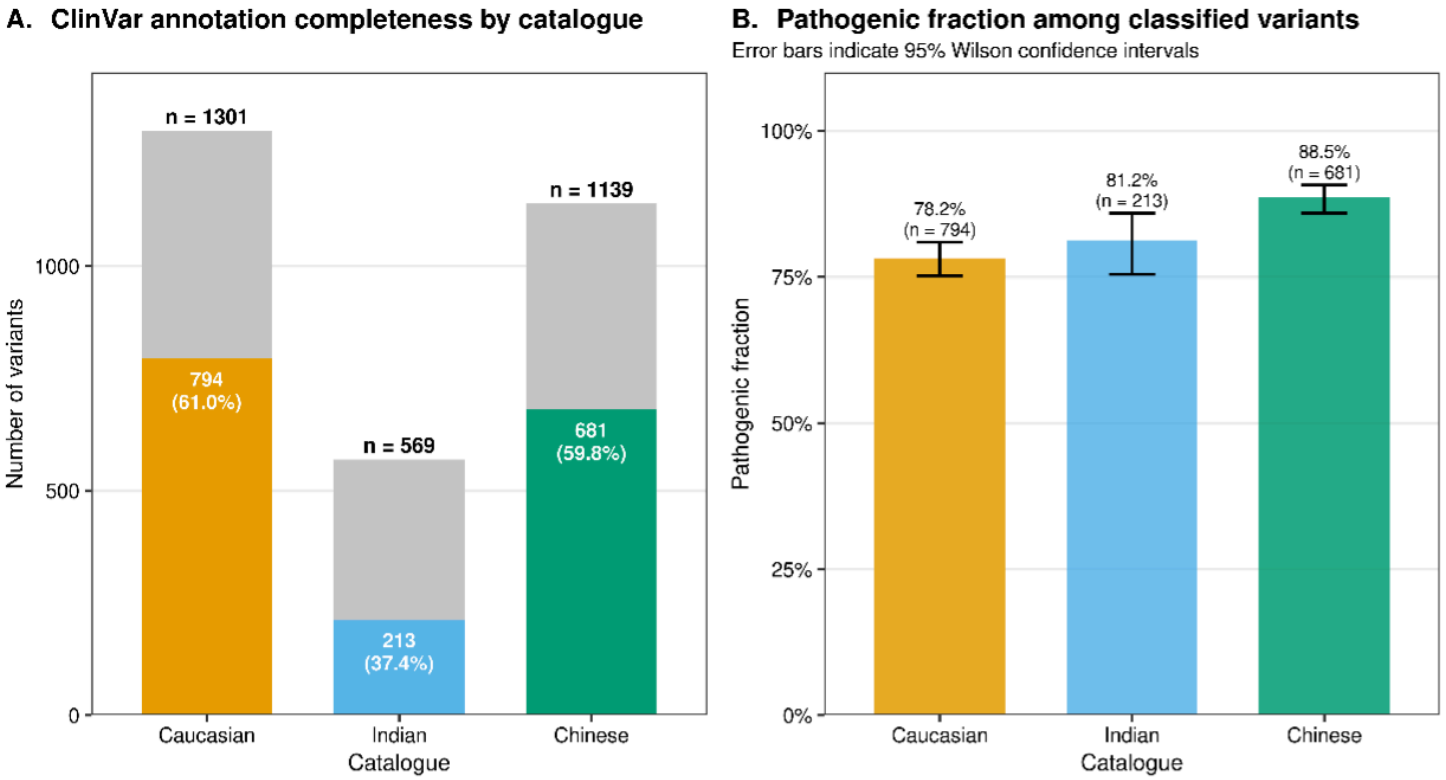


Figure 2

Population-specific variant patterns. **(A)** Shows ClinVar annotation completeness by population origin, where solid color fill in stacked bars show proportion of variants with definitive classifications (pathogenic/benign) and translucent gray fill above shows uncertain classifications. (B) Pathogenic ratios among definitively classified variants for populations. Error bars show 95% Wilson score confidence intervals (Cis; Indian-specific: 173/213, 95% CI: 75.4-85.9%; Caucasian-specific 621/794, 95% CI: 75.2-80.9%, and Chinese-specific: 603/681, 95% CI: 85.9-90.7%.

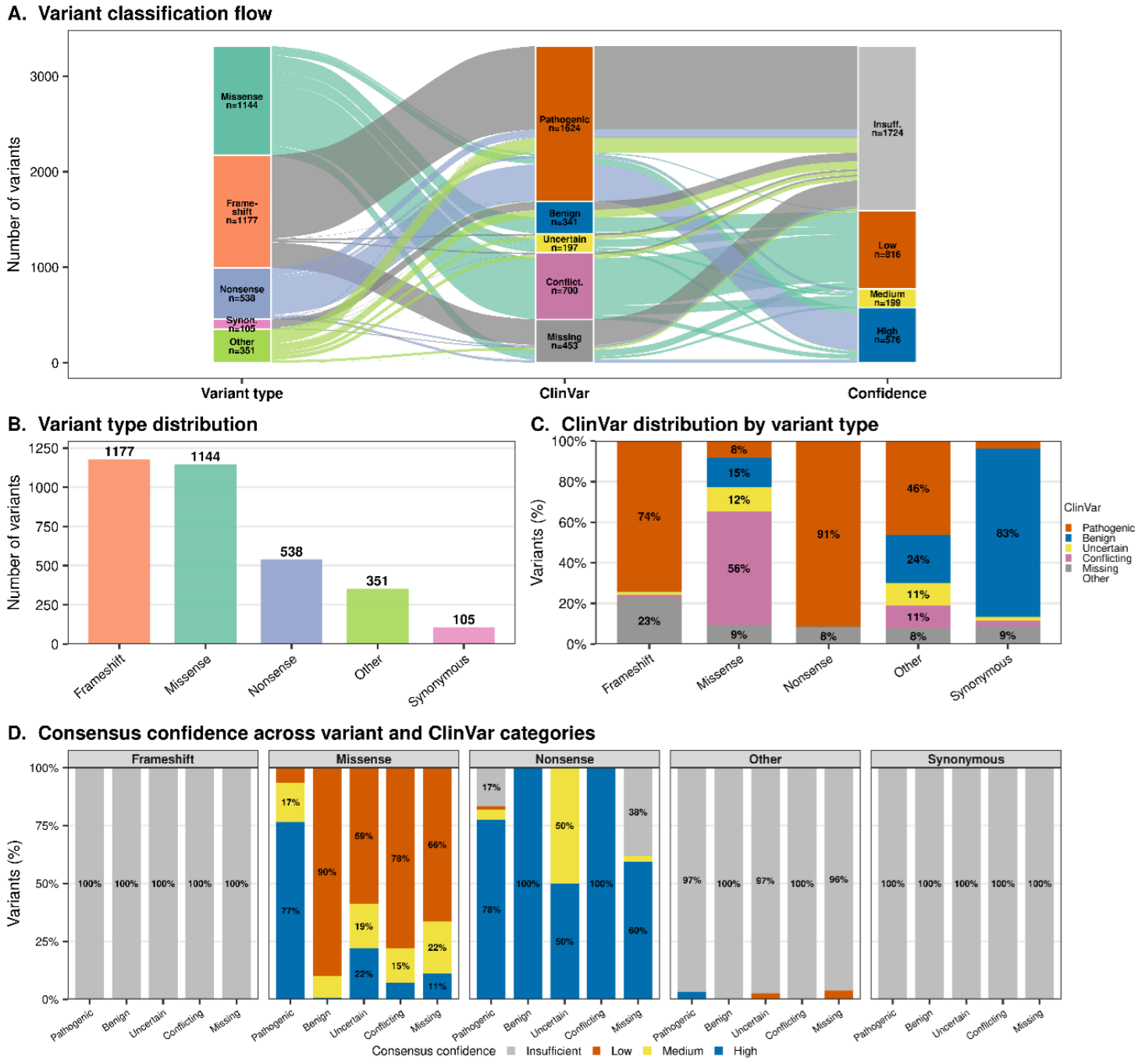


Figure 3

Variant classification flow among variant types, ClinVar annotation and consensus-score confidence across variant classes. A) Sankey diagram showing the three-level classification flow for variants in the combined dataset. The left axis represents variant type based on functional impact, including missense variants, frameshift variants, nonsense variants, synonymous variants, and other variants, comprising non-frameshift indels, substitutions, intronic variants, splice-region variants, and UTR variants. The middle axis represents ClinVar annotation categories: pathogenic, including likely pathogenic; benign, including likely benign; VUS; conflicting; and missing from ClinVar. The right axis represents confidence derived from agreement among normalized rank scores from the six Tier-1 predictors: BayesDel_noAF, AlphaMissense, ESM1b, VEST4, EVE, and ClinPred. High, medium and low confidence follow the coefficient of variation (CV) cutoffs defined in the

Methods. Variants were classified as insufficient when consensus scoring was unavailable because valid scores from at least two Tier-1 predictors could not be obtained. Ribbons connecting the axes indicate classification flow, with ribbon width proportional to the number of variants following each path. **B-D)** Quantitative summaries of the hierarchical classification structure B) Counts of variants by functional variant type C) Stacked bar plot showing the distribution of ClinVar annotation categories within each variant type. Each bar represents 100% of a given variant type, with colored segments indicating the proportion assigned to each ClinVar category; percentages ≥ 5% are labeled. D) Distribution of consensus score confidence across variant type and ClinVar annotation combinations shown in panels B and C.

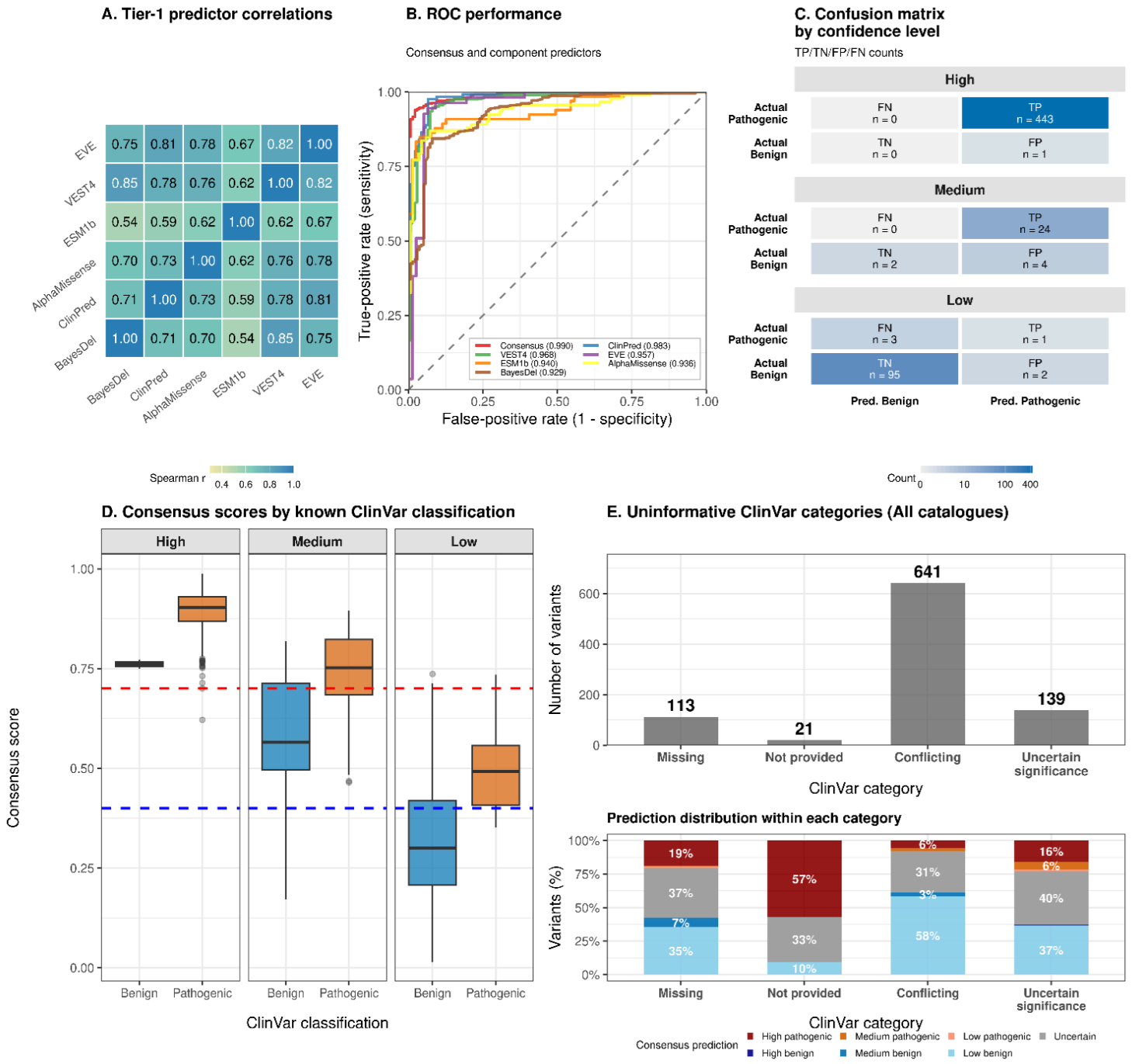


Figure 4

ClinVar label concordance of the consensus predictor and classification of uninformative variants A) Heatmap showing pairwise Spearman correlation coefficients among the six Tier-1 predictors: EVE, VEST4, ESM1b, AlphaMissense, ClinPred, and BayesDel_noAF. Correlation values range from 0.3 (low agreement, red) through 0.7 (moderate agreement, yellow) to 1.0 (perfect agreement, blue). **B)** Receiver Operating Characteristic (ROC) curves comparing the consensus meta-predictor with each individual Tier-1 predictor. Each curve is labeled with its corresponding area under curve (AUC). **C)** Confusion matrices for the consensus predictor stratified by confidence category. Tile color intensity represents the number of variants, with darker blue indicating higher counts. **D)** Consensus score distribution for ClinVar variants included in the validation set, stratified by confidence level: high, medium, and low. The red line marks the pathogenic prediction threshold of 0.70, and the blue line marks the benign prediction threshold of 0.40. Variants with consensus scores between 0.40 and 0.70 were classified as intermediate and excluded from the binary performance metrics shown in panel C. **E)** Prediction coverage and confidence-category stratification for uninformative variants. The missing category represents variants absent from ClinVar, “not provided” indicates variants present in ClinVar without submitted clinical significance, and VUS denotes variants of uncertain significance. The upper panel shows variant counts for each prediction category, while the lower panel shows prediction coverage within each uninformative class and its stratification by confidence category.

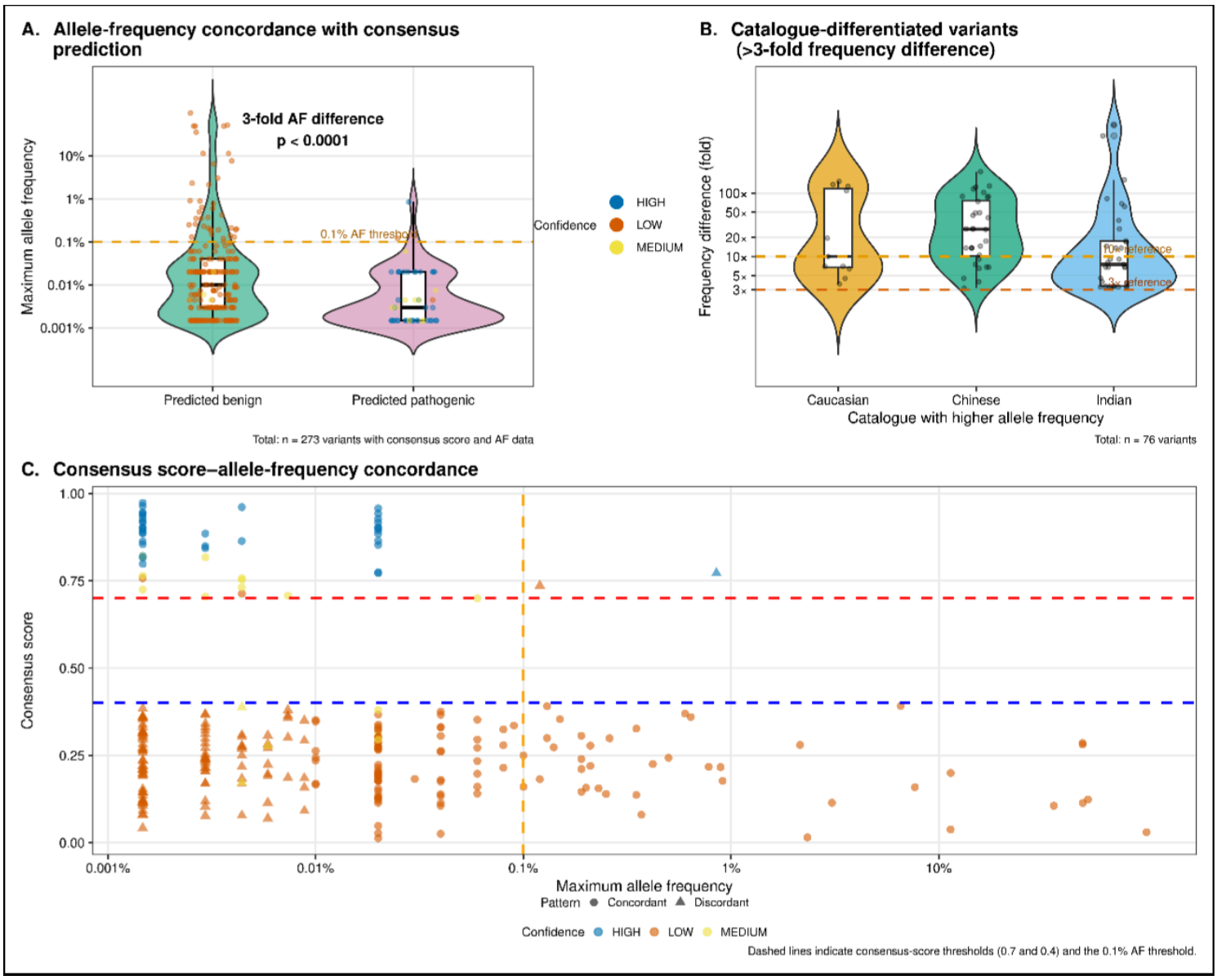


Figure 5

AF-ratio-screen and descriptive allele frequency analysis: A) Maximum NFE/SAS/EAS point AF for pathogenic-range and benign-range scores. B) Distribution of maximum AF (NFE/SAS/EAS) among 76variants passing the >3-fold AF ratio criterion . C), Scatter plot between consensus score and log10-transformed AF. Points are colored by consensus confidence level: high, medium, or low, and are shaped according to AF score relationship, with circles indicating concordant variants and triangles indicating discordant variants. Dashed reference lines indicate the pathogenic-range call threshold of 0.70, the benign-range call threshold of 0.40, and the AF threshold of 0.1%. Variants in the upper-left region, characterized by high consensus score and low AF, follow the expected pattern for rare pathogenic variants, whereas variants in the lower-right region, characterized by low consensus score and high AF, follow the expected pattern for common benign variants. Discordant triangle points represent potentially informative cases: variants in the upper-right region have pathogenic-range call despite relatively high AF whereas variants in the lower-left region have benign-range call despite rarity and may warrant re-evaluation using additional evidence.

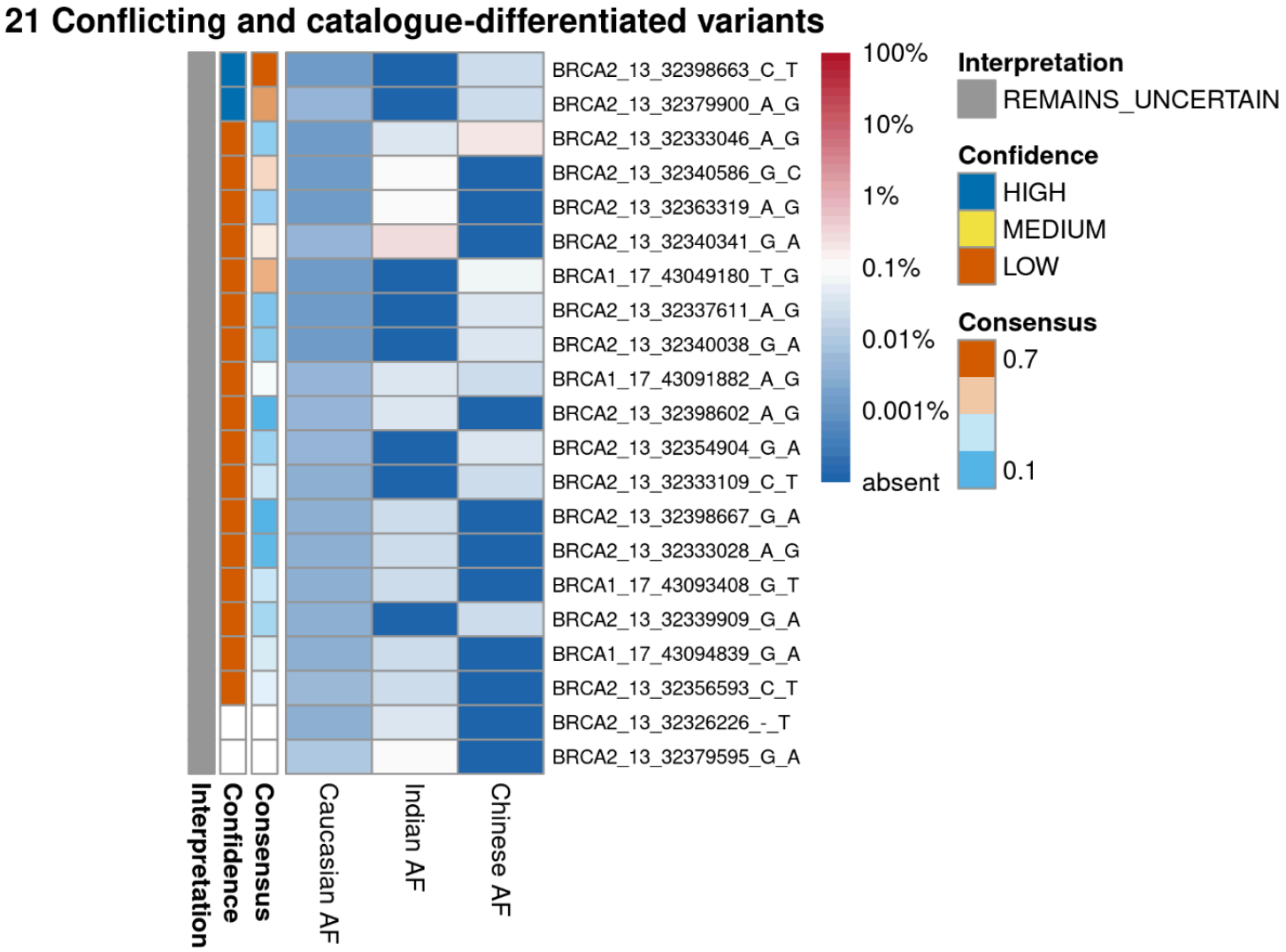


Figure 6

Integrated assessment of AF-differentiated variants with conflicting ClinVar classifications. Heatmap showing the 21 variants that were both AF-differentiated and classified as conflicting in ClinVar. Allele frequencies are displayed on a $\log_{10}$ scale, with blue indicating absent or rare variants and red indicating higher-frequency variants. Consensus scores are shown on a 0-1 scale, with blue indicating scores <0.40, consistent with likely benign prediction; white indicating intermediate scores between 0.40 and 0.70; and red indicating scores >0.70, consistent with likely pathogenic prediction. The final column (leftmost) indicates whether each variant met the prespecified score, dispersion and AF criteria; none of the 21 variants met the criteria for convergent evidence

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.pdf
- PMutants.xlsx