Calibration and refinement of ACMG/AMP criteria for variant classification with BayesQuantify

Sihan Liu[1, 2,*], Xiaoshu Feng[1,2], Fengxiao Bu[1, 2,*]

[1] Department of Otolaryngology-Head and Neck Surgery, West China Hospital, Sichuan University, Chengdu 610041, China
[2] Institute of Rare Diseases, West China Hospital, Sichuan University, Chengdu 610041, China

[*] Correspondence: bufengxiao@wchscu.cn; liusihan@wchscu.cn

Abstract

Background: Improving the precision and accuracy of variant classification in clinical genetic testing requires further specification and stratification of the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) criteria. While the Clinical Genome Resource (ClinGen) Bayesian framework enables quantitative evidence calibration for selected criteria, standardized tools to optimize evidence thresholds and systematically refine ACMG/AMP criteria remain underdeveloped.

Methods: To address this gap, we developed *BayesQuantify*, an R package that provides a unified resource for quantifying evidence strength for ACMG/AMP criteria based on the Bayesian framework. *BayesQuantify* accepts a variant classification file as input and automatically calculates the odds of pathogenicity for each evidence strength, incorporating user-provided prior probabilities of pathogenicity. Through bootstrapping, *BayesQuantify* generates thresholds by aligning the 95% lower boundary of positive likelihood ratio/local positive likelihood ratio values with the odds of pathogenicity for different evidence levels. Three independent datasets (the

ClinVar 2019 dataset, the ClinGen curated dataset, and the *PTEN* gene dataset) derived from ClinVar, HGMD, and gnomAD were utilized to evaluate the utility of *BayesQuantify*.

Results: Validation across three independent datasets demonstrates that *BayesQuantify* delivers objective, consistent refinements for both categorical and continuous ACMG/AMP evidence. Specifically, we replicated the PP3/BP4 thresholds for four computational tools (BayesDel, VEST4, REVEL, and MutPred2) recommended by the ClinGen Sequence Variant Interpretation Working Group using the ClinVar 2019 dataset. Our analysis also indicated that the PM2 criterion should be downgraded from moderate to supporting evidence, aligning with ClinGen recommendations. Importantly, we have established thresholds for supporting, moderate, and strong evidence for in-silico tools using this tool, thereby expanding the application of PP3/BP4 criteria for missense variants in the *PTEN* gene.

Conclusions: *BayesQuantify* is an accessible and user-friendly resource that enhances the rigor and reproducibility of ACMG/AMP criteria application. By facilitating evidence-based stratification and threshold optimization, the tool strengthens variant classification workflows, offering immediate value to clinical genetic testing laboratories and research communities. The package is freely available at https://github.com/liusihan/BayesQuantify.

Keywords: ACMG/AMP guidelines, Bayesian framework, Evidence refinement, R package, Likelihood ratio

Background

Interpreting sequence variants in human genes is a pivotal aspect of diagnosing and managing genetic diseases[1, 2]. However, this task presents considerable challenges due to the complexity and diversity of the genomic data and the lack of standardization and consistency across laboratories[3]. To address these issues, the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) have developed a set of guidelines for variant interpretation, comprising 28 criteria which support or refute the pathogenicity or benignity of a variant[4]. These criteria are grouped into five strength levels of evidence (stand-alone, very strong, strong, moderate, and supporting) and are combined according to specific rules to assign a variant into one of five categories: pathogenic (P), likely pathogenic (LP), uncertain significance (US), likely benign (LB), and benign (B). Since their introduction, the ACMG/AMP guidelines have facilitated the genetic diagnosis of suspected inherited disorders and have become widely adopted by clinical laboratories and researchers as the standard for variant interpretation and reporting[5-8].

In 2018, the ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI) proposed a Bayesian framework that transformed the ACMG/AMP guidelines into a probabilistic system[9]. This framework assumes four levels of evidence and exponentially scaled odds of pathogenicity (OP). It assigns a posterior probability of pathogenicity (Post_P) to a variant by multiplying the OP for each level of evidence (O_{PVSt}, O_{PSt}, O_{PM}, O_{PSu}, O_{BVSt}, O_{BSt}, O_{BM}, O_{BSu}) with a prior probability of pathogenicity (Prior_P). The resulting posterior probability is then compared with probability thresholds to categorize variants: (1) the pathogenic combinations would have Post_P >

0.99, (2) the Post_P of the likely pathogenic combinations would be 0.90 < Post_P ≤ 0.99, (3) the Post_P of the likely benign combinations would be 0.001 ≤ Post_P < 0.10, and (4) the one benign category would have Post_P < 0.001. By modelling the ACMG/AMP guidelines with a Prior_P of 0.10 and O_{PVSt} set to 350, 15 out of 17 existing ACMG/AMP evidence combinations (except BA1) were found to be mathematically consistent with the overall framework, validating the approach adopted by the ACMG/AMP.

The Bayesian framework presents opportunities for refining evidence and combining rules, as well as supporting efforts to automate components of variant pathogenicity assessments. Researchers have quantified and extended the strengths of evidence criteria such as PS1/PM5, PS3/BS3, and PP3/BP4 according to the Bayesian framework[10-16]. Importantly, thresholds for PS4, PS4_Moderate, and PS4_Supporting have been quantified by aligning positive likelihood ratio (LR^+) and local positive likelihood ratio (lr^+) values with theoretical OP for each evidence strength in our prior study[17]. Our study shows that the refinement of PS4 improves the identification of pathogenic variants, supporting that the specification and stratification of evidence is essential to establish an objective, unbiased and consistent variant classification[7, 8, 18]. However, there is a lack of existing software and tools tailored to quantify the strength of evidence and establish appropriate thresholds to refine the ACMG/AMP criteria.

In this study, we develop *BayesQuantify*, an R package that undergoes the ClinGen Bayesian framework for quantifying the strength of evidence for ACMG/AMP criteria. *BayesQuantify* provides functions to estimate OP for each level of evidence strength and the posterior probability of pathogenicity for the ACMG/AMP combining rules. Furthermore, *BayesQuantify* facilitates the

calculation of LR^+ and lr^+ for various tested cutoffs, aligning the 95% lower boundary of LR^+/lr^+ with OP to establish thresholds. Additionally, users can employ *BayesQuantify* to classify variants in accordance with ACMG/AMP combining rules and Bayesian framework, as well as visualize both the data and statistic results[4, 9]. The *BayesQuantify* is available from GitHub at https://github.com/liusihan/BayesQuantify.

Methods

BayesQuantify workflow overview

The functions available through *BayesQuantify* were designed to quantify the evidence strength and thresholds following the algorithm proposed by ClinGen SVI. The package requires one essential input: a *data.frame* containing variants classification results, including the variants, ACMG/AMP evidence, variant classification, and tested features (e.g. Table S1 and Table S2). The step-by-step usage of *BayesQuantify* consists of five main steps as follows (Figure 1):

1. **Step 1: Calculation of OP and Post_P for each evidence strength**: The SVI Working Group has provided an Excel spreadsheet to calculate Post_P from Prior_P and OP from pathogenic and benign evidence categories[9]. Based on this spreadsheet, users can manually adjust the values of Prior_P and O_{PVSt} to obtain the OP values corresponding to each level of evidence strength. However, this process was computationally inefficient and prone to user bias. To overcome these limitations, we developed *auto_select_postp*(), a core algorithm in *BayesQuantify*, which systematically evaluates O_{PVSt} values over a plausible range (1-10,000) in combination with user-defined Prior_P values. For each iteration, it calculates Post_P using Bayes' rule and validates compliance with ACMG/AMP rules for combining evidence and

predefined pathogenicity probability thresholds. The output identifies the minimum O_{PVSt} value that satisfies these constraints, ensuring robust, reproducible OP and Post_P calculation.

$$OP = O_{PVSt}^{\left(\frac{N_{PVSt}}{1}+\frac{N_{PSt}}{2}+\frac{N_{PM}}{4}+\frac{N_{PSu}}{8}-\frac{N_{BSu}}{8}-\frac{N_{BM}}{4}-\frac{N_{BSt}}{2}-\frac{N_{BVSt}}{1}\right)}$$

(Equation 1)

$$\text{Post_P} = \frac{OP * \text{Prior_P}}{\left((OP-1) * \text{Prior_P} + 1\right)}$$

(Equation 2)

2. **Step 2: Variants of US (VUS) subclassification:** The majority of human genomic variants are rare, resulting in a high prevalence of VUS in clinical classifications. To optimize statistical efficiency and reduce classification bias, we have implemented a systematic approach to categorize and filter VUS. The *VUS_classify*() function is developed to stratify variants into six confidence tiers (hot, warm, tepid, cool, cold, ice cold) as defined by the Association for Clinical Genomic Science (ACGS) Best Practice Guidelines (https://www.acgs.uk.com/quality/best-practice-guidelines/#VariantGuidelines). Variants classified as cool, cold, or ice cold are designated benign-leaning VUS (BL-VUS), whereas hot, warm, or tepid variants are classified as pathogenic-leaning VUS (PL-VUS). To establish a robust truth set for refining the evidence criteria, we recommended excluding PL-VUS to avoid confounding the truth set with ambiguous or potentially pathogenic classifications. This stratification ensures a high-confidence dataset for calibrating evidence thresholds while maintaining biological and clinical relevance.

3. **Step 3: Calculating LR^+/lr^+:** Within the ClinGen Bayesian framework, the OP quantifies the probability of observing specific evidence if a variant is pathogenic. To quantify evidence thresholds, *BayesQuantify* provided functions to calculate the LR^+/lr^+ using the following methodologies:

For binary variables (e.g. variant marked with PM2 or not): The *LR*() function computes various performance metrics by evaluating the presence or absence of specific evidence among disease-causing and non-pathogenic variants. These metrics include true positive (TP), false positive (FP), true negative (TN), and false negative (FN), overall accuracy, true positive rate (sensitivity), true negative rate (specificity), positive predictive value (PPV), negative predictive value (NPV), F1 score, and LR^+ (equation 3). The 95% confidence interval (CI) for LR^+ is estimated using bootstrapping in the R package bootLR [19]. The lower bound of LR^+ (LR^+_LB) is then utilized to determine evidence strength by comparing it with the OP thresholds generated in Step 1.

$$LR^+ = \frac{TP/(TP+FN)}{1-TN/(TN+FP)}$$

(Equation 3)

For quantitative variables (e.g. scores from in-silico tools used for tagging PP3/BP4): the *local_bootstrapped_lr()* function first sorts all unique tested cutoffs and positions each cutoff at the center of a user-defined sliding window (default: 0.01). The local posterior probability of pathogenicity and benignity is then calculated for each tested cutoff within the interval, considering a minimum quota of disease-causing and non-pathogenic variants (equation 4).

Additionally, the one-sided 95% confidence bound for each estimated lr^+ is determined through bootstrapping iterations. The *get_lr_threshold()* function can determine the thresholds for each evidence level by mapping the lower bound of local posterior probability with the thresholds generated in Step 1.

$$Local\ posterior\ probability = \frac{\#\mathrm{P/LP}\ variants\ in\ interval}{\#\mathrm{P/LP}\ \ in\ interval + (Weight * \#\mathrm{non-pathogenic}\ variants\ \ in\ interval)}$$

(Equation 4)

4. **Step 4: Variant classification:** *BayesQuantify* offers two functions for variant classification: The *ACMG_Classification*() function categorizes variants into five tiers (P, LP, B, LB, and VUS) using predefined evidence combinations from the ACMG/AMP guidelines; In parallel, the *BCF*() function applies to a Bayesian classification framework to classify variants by calculating posterior probabilities of pathogenicity with user-provided Prior_P and OP for each evidence level.
5. **Step 5: Visualization:** To improve user-friendliness, *BayesQuantify* includes several visualization functions: 1) the *multi_plot()* function generates composite plots to display variant characteristics (e.g., allele frequency, conservation scores) in the input data.frame; 2) the *heatmap_LR()* function creates heatmaps to visualize the distribution of positive likelihood ratios across tested cutoffs, highlighting thresholds where LR values meet or exceed OP bounds for pathogenicity or benignity; and 3) the *plot_lr()* function plots local posterior probability curves for quantitative variables, showing the distribution of lr^+.

All these steps are implemented within the R programming environment to consolidate all intermediate data generated throughout the workflow, ensuring transparency and user-friendly operation. For a detailed explanation of each step with code and examples, see the website of the package (https://github.com/liusihan/BayesQuantify).

Data sets

The ClinVar 2019 dataset

To evaluate the performance of *BayesQuantify*, the ClinVar 2019 dataset, compiled by Pejaver et al. was obtained from https://zenodo.org/records/8347415[10]. This dataset includes missense variants with an allele frequency (AF) below 0.01 in the Genome Aggregation Database (gnomAD v.2.1) and from genes with at least one pathogenic variant of any type in ClinVar[20]. In total, 11,834 variants (5,705 P/LP and 6,129 B/LB variants) from 1,914 genes are included in this dataset (Table S1).

The ClinGen Curated Variants dataset

The ClinGen Curated Variants dataset was retrieved from the ClinGen Evidence Repository as of April 7th, 2024 (https://erepo.clinicalgenome.org/evrepo/). This dataset contains expert-curated assertions on variant pathogenicity, along with supporting evidence summaries[18]. It includes classification summaries for 6,768 variants across 74 diseases, with 1,850 P, 1,463 LP, 679 LB, 775 B, and 2,001 US variants (Figure 2 and Table S2). Variants tagged with BA1 were excluded, resulting in the removal of 664 variants from downstream analysis.

188 **The *PTEN* dataset**

189 To perform the calibration for the *PTEN* gene (MIM# 601728), we constructed a dataset, referred
190 to as the *PTEN* dataset, which included variants classified as disease-causing and non-pathogenic.
191 Variants were sourced from three databases: ClinVar (v20241120)[21], HGMD (2024q1)[22], and
192 gnomAD (v4.1)[20]. These variants were filtered out based on the following criteria: (1) Only
193 missense variants located in the *PTEN* (NM_000314.8) gene with an AF below 0.01 in gnomAD
194 were retained; (2) For variants observed exclusively in gnomAD, only those variants annotated as
195 "PASS" in the "FILTER" column were included; (3) Variants with conflicting pathogenicity
196 annotations between ClinVar (P/LP) and HGMD (DM/DM?) were excluded; and (4) Variants
197 classified as uncertain significance or with conflicting classifications in ClinVar were also removed.
198 Disease-causing variants were defined as P/LP variants, either with a review status of
199 "criteria_provided,_multiple_submitters,_no_conflicts,", or "reviewed_by_expert_panel," or
200 those with consistent classifications between ClinVar and HGMD. Non-pathogenic variants were
201 included variants classified as B/LB variants in ClinVar or rare variants only observed in gnomAD
202 (Table S3).

203 **Estimation of the prevalence of pathogenic variants for the ClinGen Curated Variants**
204 **dataset**

205 To estimate the prevalence of pathogenic variants in the ClinGen Curated Variants dataset, we
206 performed the following steps to generate a reference dataset from gnomAD: 1) 332,531 variants
207 were extracted from gnomAD (v4.1), focusing on coding regions of genes with at least one P/LP
208 variant in the ClinGen Curated Variants dataset; 2) variants with an AF < 0.01 and those annotated

209 as "PASS" in the "FILTER" column were retained; and 3) P/LP variants present in ClinVar
210 (v20241120) or HGMD (2024q1) were removed. The final reference dataset consisted of 133,270
211 potentially non-pathogenic variants. Based on this, we estimated the prevalence of pathogenic
212 variants in the ClinGen Curated Variants dataset to be 2.48% (3,313/133,270).

213 **Estimation of the prevalence of pathogenic variants for the *PTEN* dataset**

214 For estimating the prevalence of pathogenic variants in the *PTEN* dataset, we utilized a combined
215 dataset from ClinVar (v20241120), HGMD (2024q1), and gnomAD (v4.1). Variants in the *PTEN*
216 gene with an AF > 0.01 in gnomAD, as well as those with conflicting pathogenicity annotations
217 between ClinVar (P/LP) and HGMD (DM/DM?), were excluded. The prevalence of pathogenic
218 variants in this dataset was found to be 4.68% (1,319/28,171).

219 **Calibration of computational tools in the *PTEN* dataset**

220 To systematically evaluate and calibrate computational tools, we annotated variants using
221 precomputed prediction scores from the Database of Human Nonsynonymous SNPs (dbNSFP 4.7a)
222 via the Variant Effect Predictor (VEP) [23, 24]. A total of 38 prediction scores from the dbNSFP
223 database, covering the entire human genome, were compared. All genomic positions for these
224 precomputed scores were based on the GRCh38/hg38 reference genome. To standardize the
225 assessment, we employed a rank score methodology, which scaled the raw prediction scores to a
226 0-1 range, reflecting the proportion of variants deemed less damaging. The performance of these
227 38 tools was then evaluated in the *PTEN* dataset using receiver-operating characteristic (ROC)
228 curves, generated with the R package pROC[25].

Results

Generating OP values and high-confidence datasets using *BayesQuantify*

Comprehensive OP values for ACMG/AMP criteria were established across three distinct datasets through the *auto_select_postp*() function in *BayesQuantify*. For the ClinVar 2019 dataset (Prior_P=0.0441, estimated by Pejaver et al.), pathogenic evidence required OP values ≥2.41 (Supporting), ≥5.80 (Moderate), ≥33.65 (Strong), and ≥1132 (Very Strong), while benign evidence thresholds were ≤0.415, ≤0.172, ≤0.03, and ≤0.001 respectively. The ClinGen Curated Variants dataset (Prior_P=0.0248) exhibited elevated pathogenic thresholds (O_{PSu}≥2.65, O_{PM}≥7.06, O_{PSt}≥49.9, O_{PVSt}≥2490) with corresponding benign thresholds at O_{BSu}≤0.376, O_{BM}≤0.142, O_{BSt}≤0.02, and O_{BVSt}≤0.0004. For the *PTEN* dataset, the OP values were: $O_{PSu} \geq 0.384$, $O_{PM} \geq 5.682$, $O_{PSt} \geq 32.28$, $O_{PVSt} \geq 1042$, $O_{BSu} \leq 0.42$, $O_{BM} \leq 0.176$, $O_{BSt} \leq 0.031$, and $O_{BVSt} \leq 0.001$, with a Prior_P of 0.0468.

To accurately refine the ACMG/AMP criteria, the *VUS_classify()* function was used to categorize variants of uncertain significance into six levels, generating high-confidence datasets by excluding PL-VUS variants. In the ClinGen Curated Variants dataset, this approach resulted in 5,137 variants, comprising 1,847 P, 1,463 LP, 123 B, 673 LB, and 1,031 BL-VUS variants. Correspondingly, the ClinVar 2019 dataset encompassed 5,705 disease-causing variants and 6,129 non-pathogenic variants, while the *PTEN* dataset consisted of 326 disease-causing variants and 69 non-pathogenic variants.

Validating the utility of *BayesQuantify* for evidence refinement and variant classification

To assess the accuracy of *BayesQuantify* in refining evidence for variant classification, we

250 replicated the calibration of four recommended computational tools (BayesDel, MutPred2,
251 REVEL, and VEST4) using the ClinVar 2019 dataset, as proposed by Pejaver et al. for determining
252 PP3 and BP4 evidence levels. Using 10,000 bootstrap iterations, we calculated local posterior
253 probabilities for each unique score generated by these tools. Our analysis confirmed that all four
254 tools achieved thresholds for strong evidence of pathogenicity (BayesDel ≥ 0.53, MutPred2 ≥
255 0.925, REVEL ≥ 0.921, VEST4 ≥ 0.96). For benignity, BayesDel (≤ -0.36) and VEST4 (≤ 0.32)
256 reached moderate evidence thresholds, while MutPred2 (≤ 0.02) and REVEL (≤ 0.04) met strong
257 evidence thresholds (Figure 3, Table S4). These results align with Pejaver et al.'s findings,
258 validating the reliability of *BayesQuantify* in evidence calibration.
259 We further evaluated *BayesQuantify*'s ability to classify variants into five categories (P, LP, B, LB,
260 and VUS) according to the ACMG/AMP guidelines and the Bayesian classification framework.
261 Using the ClinVar 2019 dataset, *BayesQuantify* demonstrated 99.02% concordance (n = 6,702
262 variants) with existing classifications (Table S5). Of the 66 discordant cases, discrepancies arose
263 from clinical judgement beyond guideline-based evidence. A representative example is the *CDH23*
264 variant NM_022124.6:c.380A>G (p.Asp127Gly), which is classified as a VUS according to the
265 ACMG/AMP criteria. This variant is absent in gnomAD v2.1.1 (PM2_Supporting) and
266 computationally predicted to be deleterious (REVEL = 0.796, PP3). Although detected in an
267 individual with autosomal recessive Usher syndrome (PP4) and confirmed to be in compound
268 heterozygosity with a pathogenic frameshift variant (NM_022124.6:c.1949dup, p.Leu651fs; PM3),
269 its Bayesian posterior probability (0.812) fell below the threshold for likely pathogenic
270 classification (≥ 0.9). Despite this, the ClinGen Hearing Loss Variant Curation Expert Panel

271 (VCEP) re-evaluated the variant on 18 January 2023 and retained its likely pathogenic designation
272 based on clinical expertise.

273 **Evaluating the clinical weight of PM2 criteria in variant pathogenicity classification**

274 The 2015 ACMG/AMP guidelines originally defined PM2 as a moderate pathogenic criterion,
275 applicable when a variant is absent or extremely rare (e.g., recessive conditions) in population
276 databases such as gnomAD. However, concerns arose about overestimating its predictive value, as
277 large-scale studies (e.g., ExAC) revealed that 54% of rare variants occurred only once, suggesting
278 that rarity alone is insufficient to infer pathogenicity. To address this concern, we evaluated the
279 clinical weight of PM2 criteria using *BayesQuantify*. In the ClinGen Curated Variants dataset, PM2
280 was designated as moderate evidence and supporting evidence in accordance with gene-/disease-
281 specific guidelines and emerged as the most frequently applied pathogenic criterion, being utilized
282 in 3,072 P/LP variants, 580 BL-VUS, and 114 B/LB variants, respectively (Figure S1). We
283 evaluated whether or not PM2 and PM2_Supporting met the relative odds of pathogenicity for a
284 Moderate (7.064:1) and Supporting pathogenic (2.658:1) evidence type. Consequently, a total of
285 3,766 variants were tagged with PM2_Supporting, including 3,072 P/LP variants and 694 BL-
286 VUS/B/LB variants. The accuracy was 0.819, PPV was 0.816, sensitivity was 0.928, specificity
287 was 0.620, and LR^+ was 1.497 (95% CI, 1.443–1.555), which was lower than the O_{PSu} value for
288 the supporting evidence level (Table 1 and Table S6). Furthermore, 1,102 variants were identified
289 with a moderate evidence level of PM2, comprising 1,002 P/LP variants and 100 BL-VUS/B/LB
290 variants. However, the accuracy was 0.531, PPV was 0.910, sensitivity was 0.302, specificity was
291 0.945, and LR^+ was 5.531 (95% CI, 4.575–6.818), reaching the supporting evidence level

292 (LR^+_LB ≥ 2.658).

293 **Calibration of in-silico tools for predicting pathogenicity for the *PTEN* gene**

294 We systematically evaluated 38 variant impact prediction algorithms curated in the dbNSFP v4.2a
295 database to establish calibrated thresholds for *PTEN* missense variant pathogenicity classification,
296 enhancing precision in clinical interpretation aligned with ACMG/AMP guidelines. The
297 distribution of prediction scores between non-pathogenic and disease-causing variants in the *PTEN*
298 dataset across all 38 tools was shown in Figure S2. Twelve tools (SiPhy, phastCons, phyloP,
299 GERP++, fitCons, GenoCanyon, fathmm_MKL, DANN, LIST_S2, fathmm_XF, MutationTaster,
300 and LRT) were excluded from downstream analysis due to over-mixed prediction scores for both
301 non-pathogenic and disease-causing variants. The remaining 26 tools showed an area under the
302 curve (AUC) values ranging from 0.782 to 0.946 (Figure S3), with the Evolutionary model of
303 Variant Effect (EVE; AUC =0.946) exhibiting the best performance, followed closely by Rare
304 Exome Variant Ensemble Learner (REVEL; AUC = 0.945) and Variant Effect Scoring Tool (VEST;
305 AUC = 0.943).

306 To assess the clinical utility of these prediction scores for variant curation based on ACMG/AMP
307 recommendations, we calibrated the top 10 performing tools. The local posterior probability-based
308 approach proposed by ClinGen SVI allowed for identifying thresholds corresponding to different
309 levels of evidence. This approach was applied to the following 10 tools: AlphaMissense[26],
310 BayesDel[27], ClinPred[28], EVE[29], M-CAP[30], MetaLR[31], MetaRNN[32], MVP[33],
311 REVEL[34], and SIFT[35]. We obtained variant scores from these tools, calculated local posterior
312 probabilities, and assigned corresponding strengths of evidence based on the thresholds outlined

313 in Table 1.

314 We identified thresholds for supporting and moderate evidence of pathogenicity (PP3) and
315 benignity (BP4) for all tools, except for MetaLR and MVP, which did not yield moderate evidence
316 for BP4. Interestingly, the local posterior probability curves demonstrated that, at appropriate
317 thresholds, several tools could provide strong evidence for pathogenicity (ClinPred, M-CAP,
318 MetaLR, MetaRNN) or benignity (AlphaMissense), as illustrated in Figure 4 and Table 2.

319 **Discussion**

320 The ClinGen Bayesian framework offers a solid mathematical basis for refining ACMG/AMP
321 criteria within clinical laboratory practice. However, to our knowledge, there is a lack of freely
322 available software tools that provide similar functionalities for the academic community. In this
323 article, we present *BayesQuantify*, an R package designed to enhance the flexibility and
324 convenience of quantifying the strength of evidence and establishing corresponding thresholds for
325 ACMG/AMP criteria based on the Bayesian framework. It is important to note that while
326 *BayesQuantify* is a computational tool, it necessitates human input to ensure the accuracy of its
327 results. By expanding the evidence base of the ACMG/AMP criteria, *BayesQuantify* has the
328 potential to elevate the classification of VUS to definitive interpretations.

329 *BayesQuantify* offers an objective assessment of whether various forms of evidence align with the
330 relative weights specified by the Bayesian framework. Our analysis indicated that the PM2
331 criterion should be downgraded from moderate to supporting evidence, which is consistent with
332 the ClinGen SVI recommendations (https://clinicalgenome.org/docs/pm2-recommendation-for-
333 absence-rarity/). This adjustment addresses the historical overestimation of the predictive value of

334 PM2 as population databases reveal high rates of ultra-rare variants. To mitigate the potential
335 impact of PM2 downgrading in variant classification, the ClinGen SVI also proposes novel
336 evidence combinations beyond the 2015 ACMG/AMP guidelines, combining a Very Strong (e.g.
337 PVS1) and a Supporting criterion (e.g. PM2_Supporting) to achieve a Likely Pathogenic
338 classification (Bayesian posterior probability=0.988). This approach ensures that novel loss-of-
339 function variants with strong mechanistic evidence (PVS1) and an adjusted PM2 weight retain
340 actionable classifications, while mitigating over-reliance on rarity alone.

341 The Bayesian framework implemented in *BayesQuantify* extends the quantification of evidence
342 strength from categorical to continuous variables, including computational predictors. We
343 corroborates Pejaver et al.'s findings, demonstrating that BayesDel, MutPred2, REVEL and
344 VEST4 achieve strong evidence thresholds for pathogenicity (PP3) and moderate evidence for
345 benignity (BP4). The slight discrepancies between our results and those of Pejaver et al. can be
346 attributed to variations in the bootstrap process. Although these findings are applicable across
347 various genes, a genome-wide approach may mask the heterogeneous performance of predictors
348 within distinct genes[36, 37]. As such, gene-specific calibrations are crucial for improving
349 accuracy and precision in variant classification.

350 Our study attempted to calibrate the PP3 and BP4 criteria for missense variants in the *PTEN* gene,
351 thereby refining their application in variant classification. Germline pathogenic variants in *PTEN*
352 are associated with a highly heterogeneous clinical spectrum and are observed in individuals with
353 conditions such as Cowden syndrome, Bannayan-Riley-Ruvalcaba syndrome, and other
354 phenotypes, including macrocephaly and autism, collectively referred to as PTEN Hamartoma

355 Tumor Syndrome[38, 39]. In 2018, the ClinGen PTEN Expert Panel introduced gene-specific
356 criteria for *PTEN* variant curation, incorporating disease-specific modifications and adjustments
357 to the strength of evidence across 19 criteria[40]. However, the guidelines did not establish a set
358 of thresholds for applying PP3 or BP4 criteria to *PTEN* missense variants, leaving a critical gap in
359 variant interpretation. According to our results, we recommend using MetaRNN or AlphaMissense
360 to determine evidence for codes PP3 and BP4, as these tools provide strong or moderate evidence
361 for pathogenicity and benignity[26, 32]. Both tools demonstrated high AUC values (above 0.93),
362 maximizing the strength of evidence while minimizing false positives in the supporting and
363 moderate categories. Given the absence of a dedicated validation dataset, future studies are
364 necessary to validate these thresholds.
365 A limitation worth discussing is the assumption within the Bayesian framework that each
366 categorical type of evidence at the same evidence strength level had the same mathematical support.
367 Given the complexity of interpreting genetic evidence, this simplifying assumption is reasonable
368 but may not fully capture the nuances involved. Previous studies have identified dependencies and
369 correlations between evidence rules, such as those observed in BS3/BP4 and PVS1/PP3[41, 42].
370 To address this limitation, extending the Bayesian framework by estimating more precise evidence
371 weights using real-world datasets containing true positives and true negatives could provide
372 greater accuracy. Such methods have already been implemented for some evidence types in genes
373 related to breast and colorectal cancers[16, 43]. Furthermore, given the potential misclassification
374 in public databases[44-46], we emphasize the importance of user-curated datasets in refining
375 evidence when utilizing *BayesQuantify*.

Conclusions

In summary, we have developed *BayesQuantify*, an R package for calibrating and refining ACMG/AMP criteria for variant classification according to the Bayesian framework. This tool is user-friendly and significantly facilitates the shift from a predominantly subjective assessment to a more quantitative and objective approach. By expanding the evidence base of the ACMG/AMP criteria, *BayesQuantify* can enhance the existing ACMG criteria and improve the reclassification of VUSs.

Abbreviations

ACMG/AMP: American College of Medical Genetics and Genomics/ Association for Molecular Pathology

AF: Allele frequency

B: Benign

BL-VUS: Benign-leaning VUS

CI: Confidence interval

gnomAD: The Genome Aggregation Database

LB: Likely Benign

LP: Likely Pathogenic

LR^+: Positive likelihood ratio

lr^+: Local positive likelihood ratio

OP: Odds of pathogenicity

O_{PVSt}: Odds of pathogenicity for "very strong" evidence

P: Pathogenic

Post_P: Posterior probability of pathogenicity

Prior_P: Prior probability of pathogenicity

VUS: Variants with Uncertain Significance

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The ClinGen Curated Variants data was downloaded from https://erepo.clinicalgenome.org/evrepo/. The ClinVar 2019 set was sourced from https://zenodo.org/records/8347415. The three datasets included in this study are provided in the supplementary tables and are also included in the *BayesQuantify* package.

Competing interests

The authors declare no conflicts of interest.

Funding

This work was supported by the Natural Science Foundation of Sichuan Province, China (Grant No. 2024NSFSC1767), the National Natural Science Foundation of China (Grant No. 82402185 and Grant No. 82171836) and the 1·3·5 project for disciplines of excellence, West China Hospital,

Sichuan University (Grant No. ZYJC20002).

Authors' contributions

Sihan Liu: Conceptualization, Methodology, Validation, Investigation, Visualization, Writing – original draft, Writing – review & editing, Supervision. **Xiaoshu Feng**: Methodology, Writing – original draft, Writing – review & editing. **Fengxiao Bu**: Supervision, Writing – original draft, Writing – review & editing. All authors have read and approved the final manuscript.

Acknowledgements

We would like to extend a special thanks to Information Center of West China Hospital for technical support. Figure 1 created with BioRender.com.

References

1. Giles HH, Hegde MR, Lyon E, Stanley CM, Kerr ID, Garlapow ME, Eggington JM: **The Science and Art of Clinical Genetic Variant Classification and Its Impact on Test Accuracy**. *Annu Rev Genomics Hum Genet* 2021, **22**:285-307.
2. Marian AJ: **Clinical Interpretation and Management of Genetic Variants**. *JACC Basic Transl Sci* 2020, **5**(10):1029-1042.
3. Pepin MG, Murray ML, Bailey S, Leistritz-Kessler D, Schwarze U, Byers PH: **The challenge of comprehensive and consistent sequence variant interpretation between clinical laboratories**. *Genet Med* 2016, **18**(1):20-24.
4. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E *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* 2015, **17**(5):405-424.
5. Niehaus A, Azzariti DR, Harrison SM, DiStefano MT, Hemphill SE, Senol-Cosar O, Rehm HL: **A survey assessing adoption of the ACMG-AMP guidelines for interpreting sequence variants and identification of areas for continued improvement**. *Genet Med* 2019, **21**(8):1699-1701.
6. Amendola LM, Jarvik GP, Leo MC, McLaughlin HM, Akkari Y, Amaral MD, Berg JS, Biswas S, Bowling KM, Conlin LK *et al*: **Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium**. *Am J Hum Genet* 2016, **98**(6):1067-1076.
7. Patel MJ, DiStefano MT, Oza AM, Hughes MY, Wilcox EH, Hemphill SE, Cushman BJ, Grant AR, Siegert RK, Shen J *et al*: **Disease-specific ACMG/AMP guidelines improve sequence variant interpretation for hearing loss**. *Genet Med* 2021, **23**(11):2208-2212.
8. Amendola LM, Muenzen K, Biesecker LG, Bowling KM, Cooper GM, Dorschner MO, Driscoll C, Foreman AKM, Golden-Grant K, Greally JM *et al*: **Variant Classification Concordance using the ACMG-AMP Variant Interpretation Guidelines across Nine Genomic Implementation Research Studies**. *Am J Hum Genet* 2020, **107**(5):932-941.
9. Tavtigian SV, Greenblatt MS, Harrison SM, Nussbaum RL, Prabhu SA, Boucher KM, Biesecker LG, ClinGen Sequence Variant Interpretation Working G: **Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework**. *Genet Med* 2018, **20**(9):1054-1060.
10. Pejaver V, Byrne AB, Feng BJ, Pagel KA, Mooney SD, Karchin R, O'Donnell-Luria A, Harrison SM, Tavtigian SV, Greenblatt MS *et al*: **Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria**. *Am J Hum Genet* 2022, **109**(12):2163-2177.
11. Villani RM, McKenzie ME, Davidson AL, Spurdle AB: **Regional-specific calibration enables application of computational evidence for clinical classification of 5' cis-regulatory variants in Mendelian disease**. *Am J Hum Genet* 2024, **111**(7):1301-1315.
12. Brünger T, Ivaniuk A, Pérez-Palma E, Montanucci L, Cohen S, Smith L, Parthasarathy S,

Helbig I, Nothnagel M, May P *et al*: **Pathogenic paralogous variants can be used to apply the ACMG PS1 and PM5 variant interpretation criteria**. *medRxiv* 2023:2023.2008.2022.23294353.

13. Walker LC, Hoya M, Wiggins GAR, Lindy A, Vincent LM, Parsons MT, Canson DM, Bis-Brewer D, Cass A, Tchourbanov A *et al*: **Using the ACMG/AMP framework to capture evidence related to predicted and observed impact on splicing: Recommendations from the ClinGen SVI Splicing Subgroup**. *Am J Hum Genet* 2023, **110**(7):1046-1067.
14. Brnich SE, Abou Tayoun AN, Couch FJ, Cutting GR, Greenblatt MS, Heinen CD, Kanavy DM, Luo X, McNulty SM, Starita LM *et al*: **Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework**. *Genome Med* 2019, **12**(1):3.
15. Wilcox EH, Sarmady M, Wulf B, Wright MW, Rehm HL, Biesecker LG, Abou Tayoun AN: **Evaluating the impact of in silico predictors on clinical variant classification**. *Genet Med* 2022, **24**(4):924-930.
16. Clark KA, Paquette A, Tao K, Bell R, Boyle JL, Rosenthal J, Snow AK, Stark AW, Thompson BA, Unger J *et al*: **Comprehensive evaluation and efficient classification of BRCA1 RING domain missense substitutions**. *Am J Hum Genet* 2022, **109**(6):1153-1174.
17. Liu S, Zhong M, Huang Y, Zhang Q, Chen T, Xu X, Peng W, Wang X, Feng X, Kang L *et al*: **Quantitative thresholds for variant enrichment in 13,845 cases: improving pathogenicity classification in genetic hearing loss**. *Genome Med* 2023, **15**(1):116.
18. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL *et al*: **ClinGen--the Clinical Genome Resource**. *N Engl J Med* 2015, **372**(23):2235-2242.
19. Marill KA, Chang Y, Wong KF, Friedman AB: **Estimating negative likelihood ratio confidence when test sensitivity is 100%: A bootstrapping approach**. *Stat Methods Med Res* 2017, **26**(4):1936-1948.
20. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP *et al*: **The mutational constraint spectrum quantified from variation in 141,456 humans**. *Nature* 2020, **581**(7809):434-443.
21. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W *et al*: **ClinVar: improving access to variant interpretations and supporting evidence**. *Nucleic Acids Res* 2018, **46**(D1):D1062-D1067.
22. Stenson PD, Mort M, Ball EV, Chapman M, Evans K, Azevedo L, Hayden M, Heywood S, Millar DS, Phillips AD *et al*: **The Human Gene Mutation Database (HGMD((R))): optimizing its use in a clinical diagnostic or research setting**. *Hum Genet* 2020, **139**(10):1197-1207.
23. 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**(1):103.
24. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F:

The Ensembl Variant Effect Predictor. *Genome Biol* 2016, **17**(1):122.

25. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M: **pROC: an open-source package for R and S+ to analyze and compare ROC curves**. *BMC Bioinformatics* 2011, **12**:77.
26. Cheng J, Novati G, Pan J, Bycroft C, Zemgulyte A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Sargeant T *et al*: **Accurate proteome-wide missense variant effect prediction with AlphaMissense**. *Science* 2023, **381**(6664):eadg7492.
27. Feng BJ: **PERCH: A Unified Framework for Disease Gene Prioritization**. *Hum Mutat* 2017, **38**(3):243-251.
28. 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**(4):474-483.
29. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, Gal Y, Marks DS: **Disease variant prediction with deep generative models of evolutionary data**. *Nature* 2021, **599**(7883):91-95.
30. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, Bernstein JA, Bejerano G: **M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity**. *Nat Genet* 2016, **48**(12):1581-1586.
31. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X: **Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies**. *Hum Mol Genet* 2015, **24**(8):2125-2137.
32. Li C, Zhi D, Wang K, Liu X: **MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning**. *Genome Med* 2022, **14**(1):115.
33. Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, Guan Y, Shen Y: **MVP predicts the pathogenicity of missense variants by deep learning**. *Nat Commun* 2021, **12**(1):510.
34. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D *et al*: **REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants**. *Am J Hum Genet* 2016, **99**(4):877-885.
35. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function**. *Nucleic Acids Res* 2003, **31**(13):3812-3814.
36. Brock DC, Wang M, Hussain HMJ, Rauch DE, Marra M, Pennesi ME, Yang P, Everett L, Ajlan RS, Colbert J *et al*: **Comparative analysis of in-silico tools in identifying pathogenic variants in dominant inherited retinal diseases**. *Hum Mol Genet* 2024, **33**(11):945-957.
37. Tejura M, Fayer S, McEwen AE, Flynn J, Starita LM, Fowler DM: **Calibration of variant effect predictors on genome-wide data masks heterogeneous performance across genes**. *Am J Hum Genet* 2024, **111**(9):2031-2043.
38. Mester J, Eng C: **When overgrowth bumps into cancer: the PTEN-opathies**. *Am J Med Genet C Semin Med Genet* 2013, **163C**(2):114-121.
39. Yehia L, Keel E, Eng C: **The Clinical Spectrum of PTEN Mutations**. *Annu Rev Med* 2020, **71**:103-116.

40. Mester JL, Ghosh R, Pesaran T, Huether R, Karam R, Hruska KS, Costa HA, Lachlan K, Ngeow J, Barnholtz-Sloan J *et al*: **Gene-specific criteria for PTEN variant curation: Recommendations from the ClinGen PTEN Expert Panel**. *Hum Mutat* 2018, **39**(11):1581-1592.
41. Kanavy DM, McNulty SM, Jairath MK, Brnich SE, Bizon C, Powell BC, Berg JS: **Comparative analysis of functional assay evidence use by ClinGen Variant Curation Expert Panels**. *Genome Med* 2019, **11**(1):77.
42. Garrett A, Durkie M, Callaway A, Burghel GJ, Robinson R, Drummond J, Torr B, Cubuk C, Berry IR, Wallace AJ *et al*: **Combining evidence for and against pathogenicity for variants in cancer susceptibility genes: CanVIG-UK consensus recommendations**. *J Med Genet* 2021, **58**(5):297-304.
43. Easton DF, Deffenbaugh AM, Pruss D, Frye C, Wenstrup RJ, Allen-Brady K, Tavtigian SV, Monteiro AN, Iversen ES, Couch FJ *et al*: **A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes**. *Am J Hum Genet* 2007, **81**(5):873-883.
44. Wright CF, Eberhardt RY, Constantinou P, Hurles ME, FitzPatrick DR, Firth HV, Study DDD: **Evaluating variants classified as pathogenic in ClinVar in the DDD Study**. *Genet Med* 2021, **23**(3):571-575.
45. Sharo AG, Zou Y, Adhikari AN, Brenner SE: **ClinVar and HGMD genomic variant classification accuracy has improved over time, as measured by implied disease burden**. *Genome Med* 2023, **15**(1):51.
46. Shah N, Hou YC, Yu HC, Sainger R, Caskey CT, Venter JC, Telenti A: **Identification of Misclassified ClinVar Variants via Disease Population Prevalence**. *Am J Hum Genet* 2018, **102**(4):609-619.

Figures

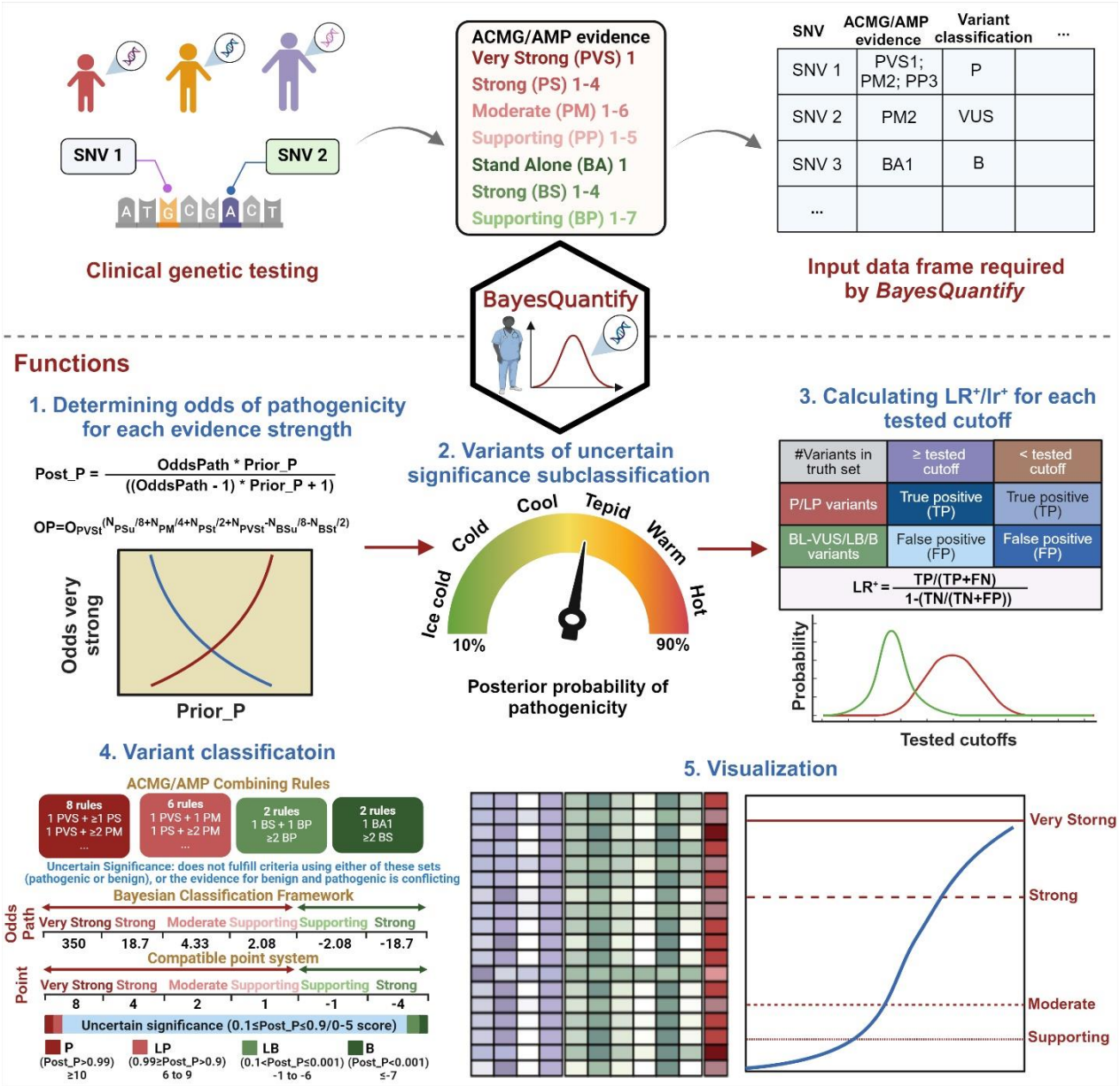
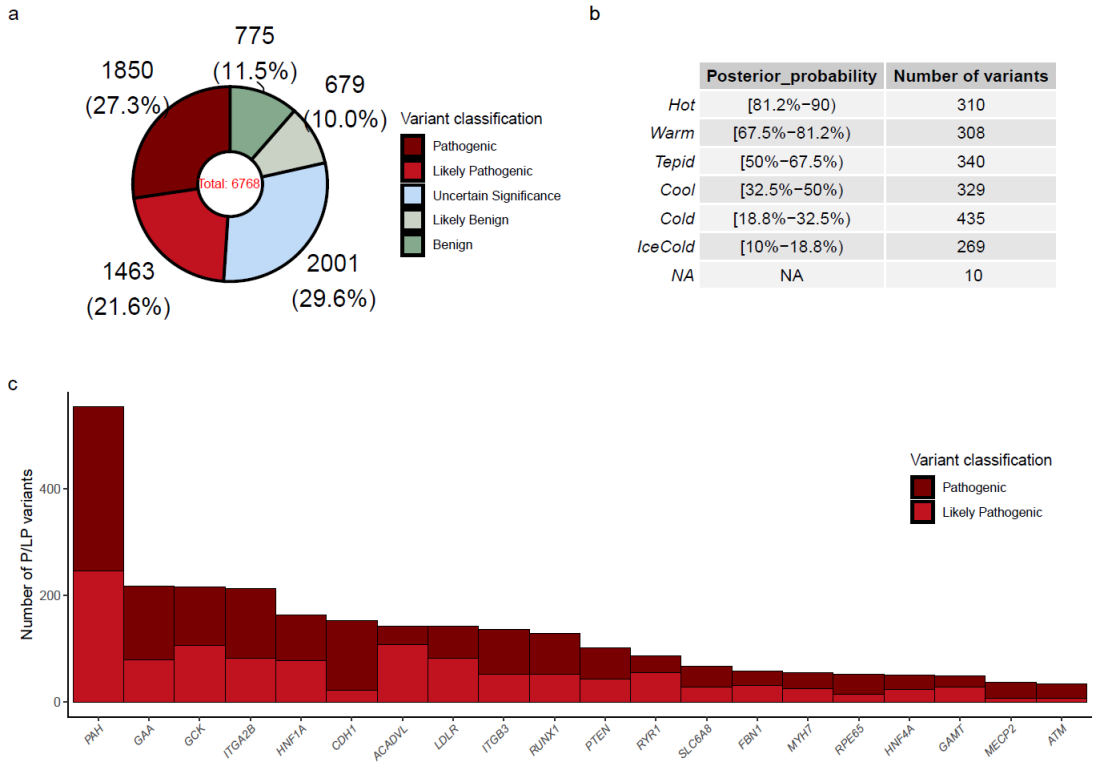


Figure 1: Schematic overview of *BayesQuantify*. *BayesQuantify* begins with the input of variant classification results dataset. Subsequently, it computes the positive likelihood ratio (LR^+) and local positive likelihood ratio (lr^+), which are then aligned with predefined odds of pathogenicity (OP) for each evidence strength to refine evidence. Prior_P: prior probability of pathogenicity; Post_P: posterior probability of pathogenicity.



	Posterior_probability	Number of variants
Hot	[81.2%−90)	310
Warm	[67.5%−81.2%)	308
Tepid	[50%−67.5%)	340
Cool	[32.5%−50%)	329
Cold	[18.8%−32.5%)	435
IceCold	[10%−18.8%)	269
NA	NA	10

Figure 2: Characteristics of the 6,768 variants in the ClinGen Curated Variants dataset. (a) Summary of classification for the 6,768 variants. (b) Further subclassification of variants of uncertain significance (VUS) to hot, warm, tepid, cool, cold, and ice-cold groups, according to the ACGS Best Practice Guidelines. Variants categorized as "cool/cold/ice cold" were defined as benign-leaning VUS (BL-VUS), while those classified as "hot/warm/tepid" were defined as pathogenic-leaning VUS (PL-VUS). (c) Distribution of pathogenic (P) and likely pathogenic (LP) variants across the top twenty genes.

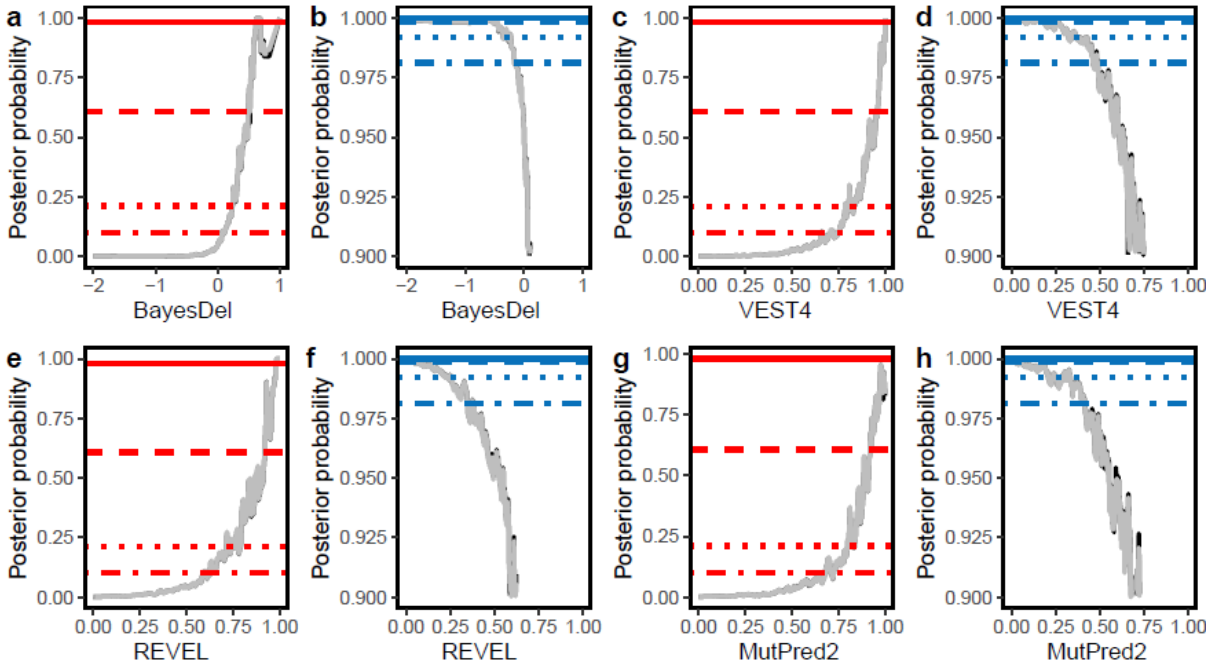


Figure 3: Local posterior probability curves of pathogenicity and benignity for various BayesDel, VEST4, REVEL and MutPred2 scores in the ClinVar 2019 dataset. The horizontal lines depict the posterior probability thresholds for supporting, moderate, strong, and very strong evidence. The black curves represent the posterior probability estimated from ClinVar 2019 dataset, while the grey curves represent one-sided 95% confidence intervals calculated from 10,000 bootstrap samples of this dataset.

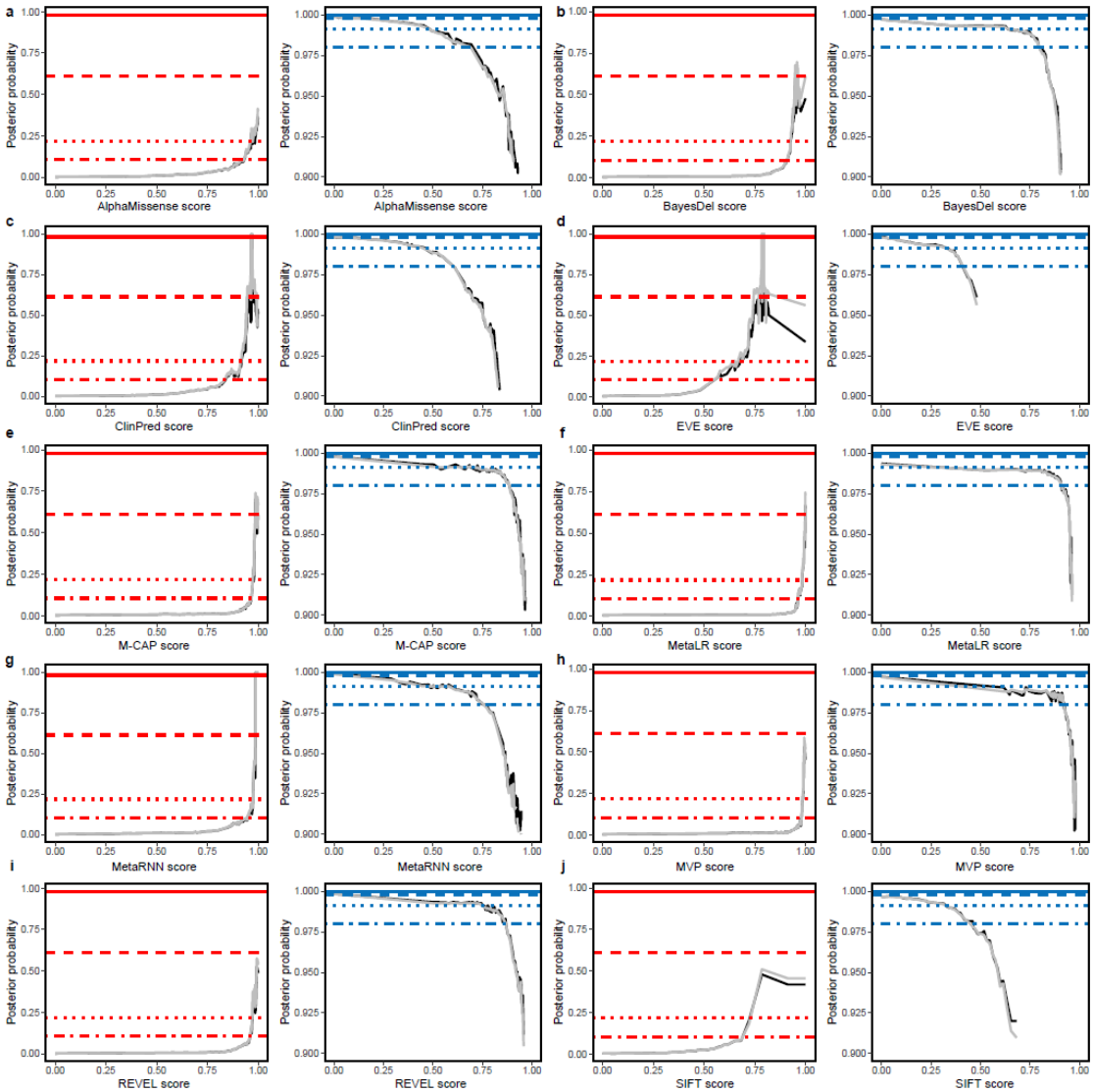


Figure 4: Local posterior probability curves ten in-silico tools in the *PTEN* dataset. The horizontal lines depict the posterior probability thresholds for supporting, moderate, strong, and very strong evidence. The black curves represent the posterior probability estimated from *PTEN* dataset, while the grey curves represent one-sided 95% confidence intervals calculated from 10,000 bootstrap samples of this dataset. The points at which the gray curves intersect the horizontal lines represent the thresholds for the relevant intervals.

Table 1: Posterior probability and positive likelihood ratio values corresponding to each evidence strength in the three datasets included in this study.

Evidence Strength	ClinGen Curated Variants dataset		ClinVar 2019 dataset		*PTEN* dataset	
	OP	**Posterior probability**	**OP**	**Posterior probability**	**OP**	**Posterior probability**
PVS	2490	0.985	1132	0.981	1042	0.981
PS	49.90	0.560	33.645	0.608	32.28	0.613
PM	7.064	0.153	5.800	0.211	5.682	0.218
PP	2.658	0.064	2.408	0.100	2.384	0.105
BP	0.376	0.990	0.415	0.981	0.420	0.980
BM	0.142	0.996	0.172	0.992	0.176	0.991
BS	0.020	0.999	0.030	0.999	0.031	0.998
BVS	0.0004	1.000	0.001	1.000	0.001	1.000

Table 2: Estimated threshold ranges corresponding to the four pathogenic and four benign intervals for the *PTEN* gene.

Method	Benign (BP4)				Pathogenic (PP3)			
	Very Strong	Strong	Moderate	Supporting	Supporting	Moderate	Strong	Very Strong
AlphaMissense	-	≤ 0.060	(0.060, 0.453]	(0.453, 0.643]	[0.935, 0.974)	≥ 0.974	-	-
BayesDel	-	-	≤ 0.638	(0.638, 0.794]	[0.908, 0.925)	≥ 0.925	-	-
ClinPred	-	-	≤ 0.447	(0.447, 0.564]	[0.836, 0.917)	[0.917, 0.999)	≥ 0.999	-
EVE	-	-	≤ 0.324	(0.324, 0.385]	[0.567, 0.668)	≥ 0.668	-	-
M-CAP	-	-	≤ 0.686	(0.686, 0.883]	[0.967, 0.973)	[0.973, 0.998)	≥ 0.998	-
MetaLR	-	-	-	≤0.907	[0.965, 0.983)	[0.983, 0.999)	≥ 0.999	-
MetaRNN	-	-	≤ 0.581	(0.581, 0.745]	[0.949, 0.976)	[0.976, 0.986)	[0.986, 0.987)	≥ 0.987
MVP	-	-	-	≤ 0.912	[0.980, 0.985)	≥ 0.985	-	-
REVEL	-	-	≤ 0.744	(0.744, 0.865]	[0.959, 0.972)	≥ 0.972	-	-
SIFT	-	-	≤ 0.328	(0.328, 0.440]	[0.722, 0.785)	≥ 0.785	-	-