Quantifying

variant evidence for precise probabilistic genomics

The quantitative omic epidemiology group, et al. “Quantifying prior probabilities for disease-causing variants reveals the top genetic contributors in inborn errors of immunity” medRxiv preprint (2025).

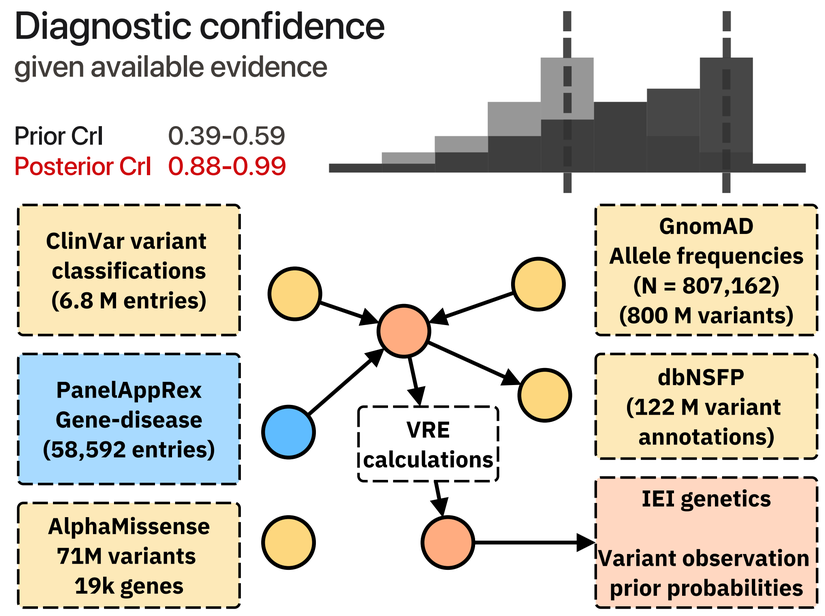
Quant Group 1, Simon Boutry 2, Ali Saadat 2, Maarja Soomann 3, Johannes Trück 3, D. Sean Froese 4, Jacques Fellay 2, Sinisa Savic 5, Luregn J. Schlapbach 6, and Dylan Lawless 6

1 The quantitative omic epidemiology group. 2 Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Switzerland. 3 Division of Immunology and the Children’s Research Center, University Children’s Hospital Zurich, University of Zurich, Zurich, Switzerland. 4 Division of Metabolism and Children’s Research Center, University Children’s Hospital Zürich, University of Zurich, Zurich, Switzerland. 5 Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, Leeds, UK. 6 Department of Intensive Care and Neonatology, University Children’s Hospital Zurich, University of Zurich, Zurich, Switzerland.

Swiss National Science Foundation (SNF) 320030_201060, and NDS-2021-911 (SwissPedHealth) from the Swiss Personalized Health Network and the Strategic Focal Area ‘Personalized Health and Related Technologies’ of the ETH Domain (Swiss Federal Institutes of Technology).

Introduction

Variant interpretation in clinical genetics has long relied on categorical labels such as pathogenic or benign. These labels simplify communication but collapse uncertainty, leaving most of the statistical evidence unquantified. A principled framework should instead account for all possible outcomes, including unobserved but plausible variants, and quantify their contributions to diagnostic probability. Our approach reframes interpretation as a Bayesian inference problem. Each variant is assigned a prior probability based on allele frequency, classification, and inheritance mode, then updated with sequencing evidence to generate a posterior probability distribution. The result is expectation-driven inference: every variant, observed or not, is integrated into a coherent probability model, and uncertainty itself becomes a measurable input to diagnosis.



Data sources and validation studies. CrI (credible interval), IEI (inborn errors of immunity), VRE (variant risk estimate).

Method

1. Target set

The target set defines the scope of inference, from a whole genome or just a small region.

$$P_{\text{tot}} = \sum_{i \in T} p_i$$

2. Priors

Based on population genetics, each variant has a probability of occurrence from allele frequency or *de novo* mutation. Mode of inheritance maps this to disease risk.

$$p_{\text{disease},i} = 2p_i(1-p_i)$$

$$p_i = \frac{1}{\max(\text{AN}) + 1}$$

$$P_{\text{AR}} = P_{\text{tot}}^2, \qquad p_{\text{disease},i} = p_i \cdot P_{\text{tot}}$$

$$p_{\text{male},i} = p_i, \qquad p_{\text{female},i} = 2p_i(1-p_i)$$

3. Weighting

Evidence sources (e.g. ClinVar, ACMG) are converted to weights W_i that scale variant probabilities.

$$p_i^* = W_i \cdot p_i$$

4. Evidence and counterfactuals

q_i encodes presence, absence, or unsequenced status. Missing data are treated as probabilistic rather than ignored.

$$q_i \in \{1, 0, p_i\}$$

$$G_i^{(m)} \sim \text{Bernoulli}(q_i)$$

5. Posterior

Posterior diagnostic probability is obtained by combining weighted priors with observed or possible presence, yielding credible intervals (CrI).

$$P_T^{(m)} = \sum_{i \in T} p_i^* \cdot G_i^{(m)}$$

6. Scenarios

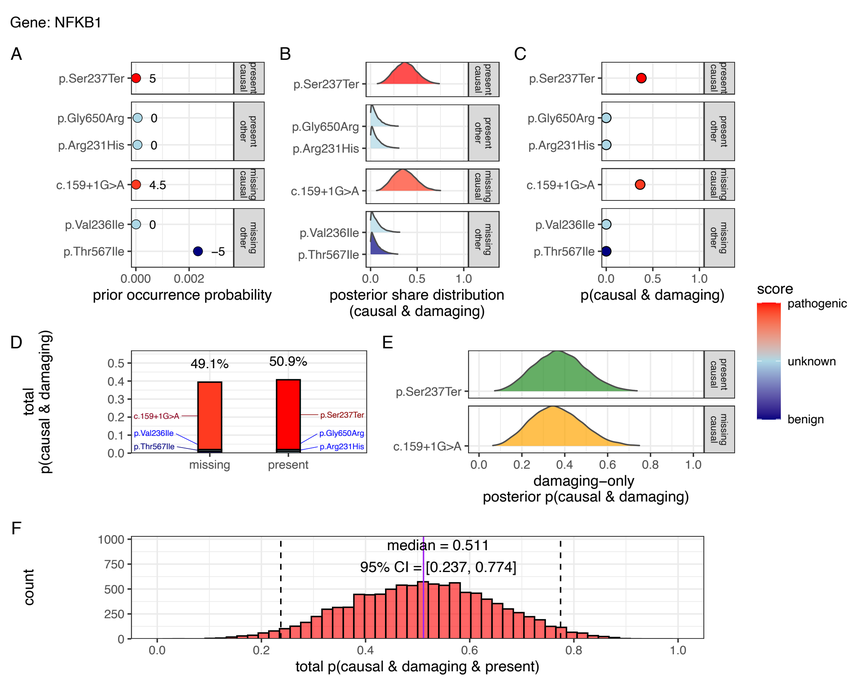
Each variant in a genome, compared to all other candidates, represents a scenario. Scenarios show how observed and missing variants shift the posterior. Both certainty and residual uncertainty are quantified.

$P_T \approx 1$ Causal

$P_T = 0$ Not causal

$P_T \approx 0.54$ (95% CrI 0.26--0.80)

Uncertainty

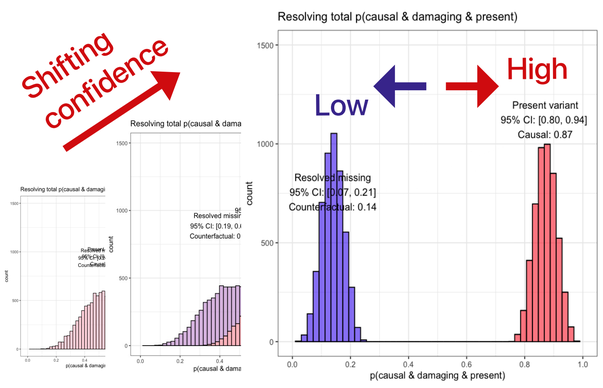


Scenario in NFKB1. Consider six variants as cause of disease. One known pathogenic variant p.Ser237Ter is present. Due to quality, a possible likely-pathogenic c.159+1G>A is unaccounted for. (A) Prior probability of observation, (B-D) posterior weights for each variant, and (E) decomposition of causal probability into observed (green) and missing (orange) sources. The posterior probability of a complete diagnosis is 0.54 (95% CrI 0.26–0.80). We must confirm the missing candidate to maximise the confidence interval.

Result

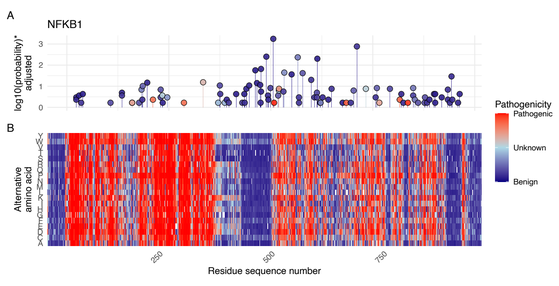
7. The shift

Instead of classifying variants as pathogenic or benign, the method shifts from categorical labels to expectation-driven inference. Every variant, observed or not, is integrated. Updating the interpretation based on complete evidence gives a new high confidence interval for a single diagnosis.



8. Validation

Validation is shown by accurate prediction of the national-scale disease frequency and cases in known causal genetic disease like dominant *NFKB1* or recessive *CFTR*. Tools like AlphaMissense predict pathogenicity. Now we can also model likelihood.



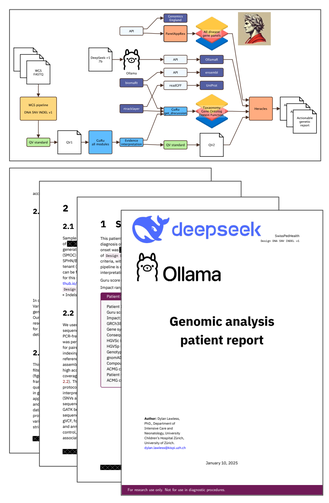
9. Actionable

Stable confidence intervals based on evidence available. We can judge empirically when stronger clinical or mechanistic evidence is required. Suitable for decision-making and AI integration.

Parameter	present	missing
Gene	NFKB1	NFKB1
HGVSc	c.710C>G	c.159+1G>A
HGVSp	p.Ser237Ter	.
Inheritance	AD	AD
Patient sex	Male	Male
gnomAD frequency	6.57e-06	6.57e-06
95% CI lower	0.003	NA
p(median)	0.090	NA
95% CI upper	0.551	NA
Posterior p(causal)	0.377	0.364
Interpretation	Reported causal; variant observed	Reported causal; variant not detected — consider follow-up
Summary	Overall probability of correct causal diagnosis due to SNV/INDEL given the currently available evidence: 0.511 (95% CI 0.237–0.774).	

Verifiable evidence interpretation RAG:

“Multiomic analysis showed that the nonsense variant is technically robust and functionally deleterious. RNA sequencing confirmed that the transcript is preserved without evidence of decay, yet proteomic profiling demonstrated absence of the normal protein product, consistent with truncation and loss of function. Metabolomic readouts supported disruption of ... ”



University of Zurich[UZH] · Universitäts-Kinderspital Zürich · Swiss Personalized Health Network (SPHN) · Strategic Focal Area Personalized Health and Related Technologies · Swiss Ped Health — A Pediatric National Data Stream · EPFL · SMC Swiss Multi-Omics Center · health2030 genome center · SIB Swiss Institute of Bioinformatics