

Quantitative prior probabilities for disease-causing variants reveal the top genetic contributors in inborn errors of immunity

Dylan Lawless^{*1}

¹Department of Intensive Care and Neonatology, University Children's Hospital Zürich,
University of Zürich, Switzerland.

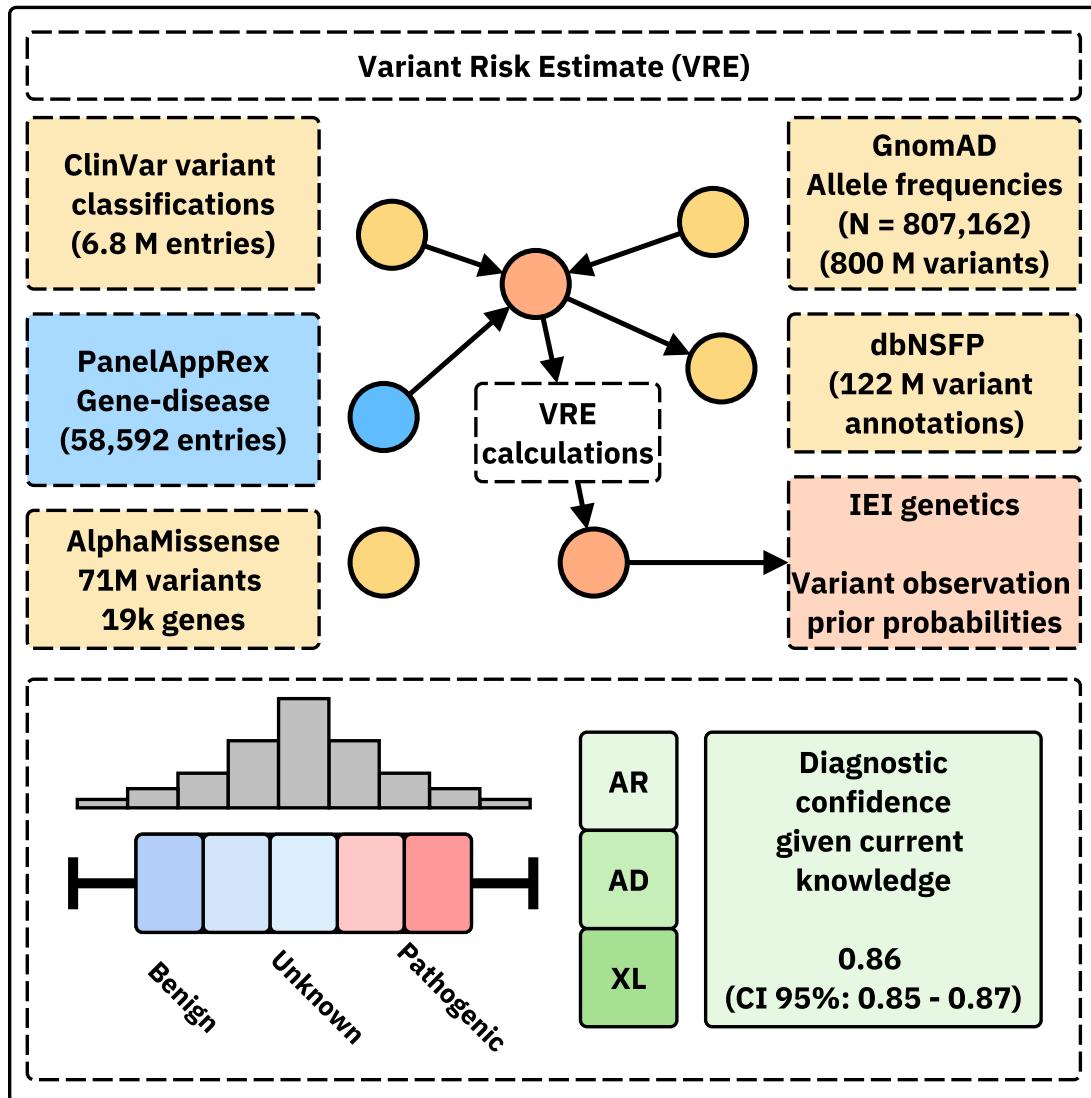
April 20, 2025

Abstract

We present a framework to quantify the prior probability of observing disease-causing variants across all genes and inheritance modes. First, we computed genome-wide occurrence probabilities by integrating population allele frequencies, variant classifications, and Hardy-Weinberg expectations under autosomal dominant, recessive, and X-linked inheritance. Second, both pathogenic variants and missing causal candidates are tested to identify the most likely genetic disease determinant and provide a clear confidence range for the overall diagnosis. This offer a complete and interpretable summary of evidence. Third, we summarised variant probabilities for 557 genes responsible for inborn errors of immunity (IEI), now integrated into a public database. Fourth, we derived new data-driven IEI classifications using protein-protein interactions and curated clinical features, aligned to immunophenotypes. Finally, we validated the framework in national-scale cohorts of autosomal dominant, recessive, and X-linked disorders, showing close concordance with observed case numbers. The resulting datasets support Bayesian variant interpretation and evidence-weighted decision-making in clinical genetics.¹

^{*}Addresses for correspondence: Dylan.Lawless@kispi.uzh.ch

¹ **Availability:** This data is integrated in public panels at <https://iei-genetics.github.io>. The source code and data are accessible as part of the variant risk estimation project at https://github.com/DylanLawless/var_risk_est and IEI-genetics project at <https://github.com/iei-genetics/iei-genetics.github.io>. The variant-level data is available from the Zenodo repository: <https://doi.org/10.5281/zenodo.15111583> (VarRiskEst PanelAppRex ID 398 gene variants.tsv). VarRiskEst is available under the MIT licence.



18

¹⁹ Acronyms

²⁰ ACMG	American College of Medical Genetics and Genomics.....	³⁶
²¹ ACAT	Aggregated Cauchy Association Test	³⁶
²² AD	Autosomal Dominant.....	⁴
²³ ANOVA	Analysis of Variance	¹⁵
²⁴ AR	Autosomal Recessive	⁴
²⁵ BMF	Bone Marrow Failure.....	²⁴
²⁶ CD	Complement Deficiencies	²⁵
²⁷ CI	Confidence Interval.....	⁹
²⁸ CrI	Credible Interval	¹⁰
²⁹ CF	Cystic Fibrosis	¹²
³⁰ CFTR	Cystic Fibrosis Transmembrane Conductance Regulator.....	⁶
³¹ CVID	Common Variable Immunodeficiency.....	¹¹
³² dbNSFP	database for Non-Synonymous Functional Predictions	⁵
³³ GE	Genomics England	⁵
³⁴ gnomAD	Genome Aggregation Database.....	⁵
³⁵ gVCF	genomic variant call format	⁹
³⁶ HGVS	Human Genome Variation Society.....	⁶
³⁷ HPC	High-Performance Computing.....	⁸
³⁸ HSD	Honestly Significant Difference	¹⁵
³⁹ HWE	Hardy-Weinberg Equilibrium.....	⁴
⁴⁰ IEI	Inborn Errors of Immunity	⁴
⁴¹ Ig	Immunoglobulin	²⁹
⁴² InDel	Insertion/Deletion	⁵
⁴³ IUIS	International Union of Immunological Societies	⁶
⁴⁴ LD	Linkage Disequilibrium	²⁷
⁴⁵ LOEUF	Loss-Of-function Observed/Expected Upper bound Fraction	¹⁵
⁴⁶ LOF	Loss-of-Function	²⁴
⁴⁷ MOI	Mode of Inheritance	⁴
⁴⁸ NFKB1	Nuclear Factor Kappa B Subunit 1	⁶
⁴⁹ OMIM	Online Mendelian Inheritance in Man	³³
⁵⁰ PID	Primary Immunodeficiency	⁴
⁵¹ PPI	Protein-Protein Interaction	⁶
⁵² QC	Quality Control	¹⁶
⁵³ SNV	Single Nucleotide Variant	⁴
⁵⁴ SKAT	Sequence Kernel Association Test.....	³⁶
⁵⁵ STRINGdb	Search Tool for the Retrieval of Interacting Genes/Proteins.....	⁶
⁵⁶ TP	true positive	⁹

⁹³ FP	false positive.....	9
⁹⁵ TN	true negative.....	9
⁹⁷ FN	false negative.....	9
¹⁰⁰ TNFAIP3	Tumor necrosis factor, alpha-induced protein 3	18
¹⁰² UMAP	Uniform Manifold Approximation and Projection	24
¹⁰⁴ UniProt	Universal Protein Resource.....	5
¹⁰⁶ VCF	variant call format	9
¹⁰⁸ VEP	Variant Effect Predictor.....	6
¹¹⁰ XL	X-Linked	4

¹¹¹

¹¹² 1 Introduction

¹¹³ In this study, we focused on reporting the probability of disease observation through
¹¹⁴ genome-wide assessments of gene-disease combinations. Our central hypothesis was
¹¹⁵ that by using highly curated annotation data including population allele frequen-
¹¹⁶ cies, disease phenotypes, Mode of Inheritance (MOI) patterns, and variant classi-
¹¹⁷ fications and by applying rigorous calculations based on Hardy-Weinberg Equilib-
¹¹⁸ rium (HWE), we could accurately estimate the expected probabilities of observing
¹¹⁹ disease-associated variants. Among other benefits, this knowledge can be used to
¹²⁰ derive genetic diagnosis confidence by incorporating these new priors.

¹²¹ In this report, we focused on known Inborn Errors of Immunity (IEI) genes, also re-
¹²² ferred to as the Primary Immunodeficiency (PID) or Monogenic Inflammatory Bowel
¹²³ Disease genes ([1–3](#)) to validate our approach and demonstrate its clinical relevance.
¹²⁴ This application to a well-established genotype-phenotype set, comprising over 500
¹²⁵ gene-disease associations, underscores its utility ([1](#)).

¹²⁶ Quantifying the risk that a newborn inherits a disease-causing variant is a fun-
¹²⁷ damental challenge in genomics. Classical statistical approaches grounded in HWE
¹²⁸ ([4; 5](#)) have long been used to calculate genetic MOI probabilities for Single Nucleotide
¹²⁹ Variant (SNV)s. However, applying these methods becomes more complex when ac-
¹³⁰ counting for different MOI, such as Autosomal Recessive (AR) versus Autosomal
¹³¹ Dominant (AD) or X-Linked (XL) disorders. In AR conditions, for example, the
¹³² occurrence probability must incorporate both the homozygous state and compound
¹³³ heterozygosity, whereas for AD and XL disorders, a single pathogenic allele is suffi-
¹³⁴ cient to cause disease. Advances in genetic research have revealed that MOI can be
¹³⁵ even more complex ([6](#)). Mechanisms such as dominant negative effects, haploinsuffi-
¹³⁶ ciency, mosaicism, and digenic or epistatic interactions can further modulate disease
¹³⁷ risk and clinical presentation, underscoring the need for nuanced approaches in risk
¹³⁸ estimation. Karczewski et al. ([7](#)) made significant advances; however, the remain-
¹³⁹ ing challenge lay in applying the necessary statistical genomics data across all MOI
¹⁴⁰ for any gene-disease combination. Similar approaches have been reported for disease

141 such Wilson disease, Mucopolysaccharidoses, Primary ciliary dyskinesia, and treat-
142 able metabolic diseases, (8; 9), as reviewed by Hannah et al. (10).

143 To our knowledge all approaches to date have been limited to single MOI, specific
144 to the given disease, or restricted to a small number of genes. We argue that our
145 integrated approach is highly powerful because the resulting probabilities can serve
146 as informative priors in a Bayesian framework for variant and disease probability
147 estimation; a perspective that is often overlooked in clinical and statistical genetics.
148 Such a framework not only refines classical HWE-based risk estimates but also has
149 the potential to enrich clinicians' understanding of what to expect in a patient and to
150 enhance the analytical models employed by bioinformaticians. The dataset also holds
151 value for AI and reinforcement learning applications, providing an enriched version of
152 the data underpinning frameworks such as AlphaFold (11) and AlphaMissense (12).

153 We introduced PanelAppRex to aggregate gene panel data from multiple sources,
154 including Genomics England (GE) PanelApp, ClinVar, and Universal Protein Re-
155 source (UniProt), thereby enabling advanced natural searches for clinical and research
156 applications (2; 3; 13; 14). It automatically retrieves expert-curated panels, such as
157 those from the NHS National Genomic Test Directory and the 100,000 Genomes
158 Project, and converts them into machine-readable formats for rapid variant discov-
159 ery and interpretation. We used PanelAppRex to label disease-associated variants.
160 We also integrate key statistical genomic resources. The gnomAD v4 dataset com-
161 piles data from 807,162 individuals, encompassing over 786 million SNVs and 122
162 million Insertion/Deletion (InDel)s with detailed population-specific allele frequen-
163 cies (7). database for Non-Synonymous Functional Predictions (dbNSFP) provides
164 functional predictions for over 120 million potential non-synonymous and splicing-
165 site SNVs, aggregating scores from 33 sources alongside allele frequencies from major
166 populations (15). ClinVar offers curated variant classifications such as "Pathogenic",
167 "Likely pathogenic" and "Benign" mapped to HGVS standards and incorporating
168 expert reviews (13).

169 2 Methods

170 2.1 Dataset

171 Data from Genome Aggregation Database (gnomAD) v4 comprised 807,162 indi-
172 viduals, including 730,947 exomes and 76,215 genomes (7). This dataset provided
173 786,500,648 SNVs and 122,583,462 InDels, with variant type counts of 9,643,254 syn-
174 onymous, 16,412,219 missense, 726,924 nonsense, 1,186,588 frameshift and 542,514
175 canonical splice site variants. ClinVar data were obtained from the variant summary
176 dataset (as of: 16 March 2025) available from the NCBI FTP site, and included
177 6,845,091 entries, which were processed into 91,319 gene classification groups and a
178 total of 38,983 gene classifications; for example, the gene *A1BG* contained four vari-
179 ants classified as likely benign and 102 total entries (13). For our analysis phase

we also used dbNSFP which consisted of a number of annotations for 121,832,908 SNVs (15). The PanelAppRex core model contained 58,592 entries consisting of 52 sets of annotations, including the gene name, disease-gene panel ID, diseases-related features, confidence measurements. (2) A Protein-Protein Interaction (PPI) network data was provided by Search Tool for the Retrieval of Interacting Genes/Proteins (STRINGdb), consisting of 19,566 proteins and 505,968 interactions (16). The Human Genome Variation Society (HGVS) nomenclature is used with Variant Effect Predictor (VEP)-based codes for variant IDs. We carried out validations for disease cohorts with Nuclear Factor Kappa B Subunit 1 (*NFKB1*) (17–20) and Cystic Fibrosis Transmembrane Conductance Regulator (*CFTR*) (21–23) to demonstrate applications in AD and AR disease genes, respectively. AlphaMissense includes pathogenicity prediction classifications for 71 million variants in 19 thousand human genes (12; 26). We used these scores to compared against the probability of observing the same given variants. **Box 2.1** list the definitions from the International Union of Immunological Societies (IUIS) IEI for the major disease categories used throughout this study (1).

Box 2.1 Definitions for IEI Major Disease Categories

Major Category	Description
1. CID	Immunodeficiencies affecting cellular and humoral immunity
2. CID+	Combined immunodeficiencies with associated or syndromic features
3. PAD	- Predominantly Antibody Deficiencies
4. PIRD	- Diseases of Immune Dysregulation
5. PD	- Congenital defects of phagocyte number or function
6. IID	- Defects in intrinsic and innate immunity
7. AID	- Autoinflammatory Disorders
8. CD	- Complement Deficiencies
9. BMF	- Bone marrow failure

195

196 2.2 Variant Class Observation Probability

As a starting point, we considered the classical HWE for a biallelic locus:

$$p^2 + 2pq + q^2 = 1,$$

where p is the allele frequency, $q = 1 - p$, p^2 represents the homozygous dominant, $2pq$ the heterozygous, and q^2 the homozygous recessive genotype frequencies. For disease phenotypes, particularly under AR MOI, the risk is traditionally linked to the homozygous state (p^2); however, to account for compound heterozygosity across multiple variants, we allocated the overall gene-level risk proportionally among variants.

Our computational pipeline estimated the probability of observing a disease-associated genotype for each variant and aggregated these probabilities by gene and ClinVar

classification. This approach included all variant classifications, not limited solely to those deemed “pathogenic”, and explicitly conditioned the classification on the given phenotype, recognising that a variant could only be considered pathogenic relative to a defined clinical context. The core calculations proceeded as follows:

1. Allele Frequency and Total Variant Frequency. For each variant i in a gene, the allele frequency was denoted as p_i . For each gene, we defined the total variant frequency (summing across all reported variants in that gene) as:

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

If any of the possible SNV had no observed allele ($p_i = 0$), we assigned a minimal risk:

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

where $\max(AN)$ was the maximum allele number observed for that gene. This adjustment ensured that a nonzero risk was incorporated even in the absence of observed variants.

2. Occurrence Probability Based on MOI. The probability that an individual was affected by a variant depended on the MOI relative to a specific phenotype. Specifically, we calculated the occurrence probability $p_{\text{disease},i}$ for each variant as follows:

- For **AD** and **XL** variants, a single copy was sufficient, so

$$p_{\text{disease},i} = p_i.$$

- For **AR** variants, disease is expected to manifest when two pathogenic alleles were present. In this case, we accounted for both the homozygous state and the possibility of compound heterozygosity. We allocated the overall gene-level risk (P_{tot}^2) proportionally by variant allele frequency:

$$p_{\text{disease},i} = p_i P_{\text{tot}}.$$

221 **3. Expected Case Numbers and Case Detection Probability.** Given a pop-
222 ulation with N births (e.g. as seen in our validation studies, $N = 69\,433\,632$), the
223 expected number of cases attributable to variant i was calculated as:

$$E_i = N \cdot p_{\text{disease},i}.$$

224 The probability of detecting at least one affected individual for that variant was
225 computed as:

$$P(\geq 1)_i = 1 - (1 - p_{\text{disease},i})^N.$$

226 **4. Aggregation by Gene and ClinVar Classification.** For each gene and for
227 each ClinVar classification (e.g. “Pathogenic”, “Likely pathogenic”, “Uncertain sig-
228 nificance”, etc.), we aggregated the results across all variants. The total expected
229 cases for a given group was:

$$E_{\text{group}} = \sum_{i \in \text{group}} E_i,$$

230 and the overall probability of observing at least one case within the group was
231 calculated as:

$$P_{\text{group}} = 1 - \prod_{i \in \text{group}} (1 - p_{\text{disease},i}).$$

232 **5. Data Processing and Implementation.** We implemented the calculations
233 within a High-Performance Computing (HPC) pipeline and provided an example
234 for a single dominant disease gene, *TNFAIP3*, in the source code to enhance repro-
235 ducibility. Variant data were imported in chunks from the annotation database for
236 all chromosomes (1-22, X, Y, M).

237 For each data chunk, the relevant fields were gene name, position, allele number,
238 allele frequency, ClinVar classification, and HGVS annotations. Missing classifica-
239 tions (denoted by “.”) were replaced with zeros and allele frequencies were converted
240 to numeric values. We then retained only the first transcript allele annotation for sim-
241 plicity, as the analysis was based on genomic coordinates. Subsequently, the variant
242 data were merged with gene panel data from PanelAppRex to obtain the disease-
243 related MOI mode for each gene. For each gene, if no variant was observed for a
244 given ClinVar classification (i.e. $p_i = 0$), a minimal risk was assigned as described

245 above. Finally, we computed the occurrence probability, expected cases, and the
246 probability of observing at least one case using the equations presented.

247 The final results were aggregated by gene and ClinVar classification and used to
248 generate summary statistics that reviewed the predicted disease observation proba-
249 bilities.

250 **2.3 Integrating observed true positives and unobserved false**
251 **negatives into a single, actionable conclusion**

252 In this section, we detail our approach to integrating sequencing data with prior clas-
253 sification evidence (e.g. pathogenic on ClinVar) to calculate the posterior probability
254 of a complete successful genetic diagnosis. Our method is designed to account for
255 possible outcomes of true positive (TP), true negative (TN), and false negative (FN),
256 by first ensuring that all nucleotides corresponding to known variant classifications
257 (benign, pathogenic, etc.) have been accurately sequenced. This implies the use of ge-
258 nomic variant call format (gVCF)-style data which refer to variant call format (VCF)s
259 that contain a record for every position in the genome (or interval of interest) regard-
260 less of whether a variant was detected at that site or not. Only after confirming that
261 these positions match the reference alleles (or novel unaccounted variants are classi-
262 fied) do we calculate the probability that additional, alternative pathogenic variants
263 (those not observed in the sequencing data) could be present. Our Confidence In-
264 terval (CI) for pathogenicity thus incorporates uncertainty from the entire process,
265 including the tally of TP, TN, and FN outcomes. We ignore the contribution of false
266 positive (FP)s as a separate task to be tackled in the future.

267 We estimated, for every query (e.g. gene or disease-panel), the posterior proba-
268 bility that at least one constituent allele is both damaging and causal in the proband.
269 The workflow comprises four consecutive stages.

270 **(i) Data pre-processing** All coding and canonical splice-region variants for *NFKB1*
271 were extracted from the gVCF. Sites corresponding to previously reported pathogenic
272 alleles were checked for read depth ≥ 10 and genotype quality ≥ 20 . Positions that
273 failed this check were labelled *missing*, thus separating true reference calls from un-
274 informative sequence.

275 **(ii) Evidence mapping and occurrence probability** PanelAppRex variants
276 were annotated with ClinVar clinical significance. Each label was converted to an
277 ordinal evidence score $S_i \in [-5, 5]$ (Table S2) and rescaled to a pathogenic weight
278 $W_i = \text{rescale}(S_i; -5, 5 \rightarrow 0, 1)$. The HWE-based pipeline of Section 2.2 supplied a
279 per-variant occurrence probability p_i . The adjusted prior was

$$p_i^* = W_i p_i, \quad \text{and} \quad \text{flag}_i \in \{\text{present}, \text{missing}\}.$$

280 **(iii) Prior specification** In a hypothetical cohort of $n = 200$ diploid individuals
 281 the count of allele i follows a Beta–Binomial model. Marginalising the Binomial yields
 282 the Beta prior

$$\pi_i \sim \text{Beta}(\alpha_i, \beta_i), \quad \alpha_i = \text{round}(2np_i^*) + \tilde{w}_i, \quad \beta_i = 2n - \text{round}(2np_i^*) + 1,$$

283 where $\tilde{w}_i = \max(1, S_i + 1)$ contributes an additional pseudo-count whenever $S_i >$
 284 0.

285 **(iv) Posterior simulation and aggregation** For each variant i we drew $M =$
 286 10 000 realisations $\pi_i^{(m)}$ and normalised within each iteration,

$$\tilde{\pi}_i^{(m)} = \frac{\pi_i^{(m)}}{\sum_j \pi_j^{(m)}}.$$

287 Variants with $S_i > 3$ were deemed *causal*. Their mean posterior share $\bar{\pi}_i =$
 288 $M^{-1} \sum_m \tilde{\pi}_i^{(m)}$ and 95% Credible Interval (CrI) were retained. The probability that a
 289 damaging causal allele is physically present was obtained by a second layer:

$$P^{(m)} = \sum_{i: S_i > 3} \tilde{\pi}_i^{(m)} G_i^{(m)}, \quad G_i^{(m)} \sim \text{Bernoulli}(g_i),$$

290 with $g_i = 1$ for present variants, $g_i = 0$ for reference calls, and $g_i = p_i$ for missing
 291 variants. The gene-level estimate is the median of $\{P^{(m)}\}_{m=1}^M$ and its 2.5th/97.5th
 292 percentiles.

293 **Scenario analysis** Three scenarios were explored: (1) observed variants only, in-
 294 cluding only one known TP pathogenic variant, **p.Ser237Ter**, (2) inclusion of the ad-
 295 dditional plausible yet unsequenced splice-donor allele **c.159+1G>A** (likely pathogenic)
 296 as a FN, and (3) where no known causal variants were present for a patient, one rep-
 297 resentative variant from each distinct ClinVar classification was selected and marked
 298 as unsequenced to emulate a range of putative FNs. The selected variants were:
 299 **p.Cys243Arg** (pathogenic), **p.Tyr246Ter** (likely pathogenic), **p.His646Pro** (conflict-
 300 ing interpretations of pathogenicity), **p.Thr635Ile** (uncertain significance), **p.Arg162Trp**
 301 (not provided), **p.Arg280Trp** (likely benign), **p.Ile207Leu** (benign/likely benign),
 302 and **p.Lys304Glu** (benign). All subsequent steps were identical.

303 **2.4 Validation of Autosomal Dominant Estimates Using *NFKB1***

304 To validate our genome-wide probability estimates in an AD gene, we focused on
305 *NFKB1*. Our goal was to compare the expected number of *NFKB1*-related Common
306 Variable Immunodeficiency (CVID) cases, as predicted by our framework, with the
307 reported case count in a well-characterised national-scale PID cohort.

308 **1. Reference Dataset.** We used a reference dataset reported by Tuijnenburg
309 et al. (17) to build a validation model in an AD disease gene. This study performed
310 whole-genome sequencing of 846 predominantly sporadic, unrelated PID cases from
311 the NIHR BioResource-Rare Diseases cohort. There were 390 CVID cases in the
312 cohort. The study identified *NFKB1* as one of the genes most strongly associated
313 with PID. Sixteen novel heterozygous variants including truncating, missense, and
314 gene deletion variants, were found in *NFKB1* among the CVID cases.

2. Cohort Prevalence Calculation. Within the cohort, 16 out of 390 CVID
cases were attributable to *NFKB1*. Thus, the observed cohort prevalence was

$$\text{Prevalence}_{\text{cohort}} = \frac{16}{390} \approx 0.041,$$

315 with a 95% confidence interval (using Wilson's method) of approximately (0.0254, 0.0656).

3. National Estimate Based on Literature. Based on literature, the prevalence
of CVID in the general population was estimated as

$$\text{Prevalence}_{\text{CVID}} = \frac{1}{25\,000}.$$

For a UK population of

$$N_{\text{UK}} \approx 69\,433\,632,$$

the expected total number of CVID cases was

$$E_{\text{CVID}} \approx \frac{69\,433\,632}{25\,000} \approx 2777.$$

Assuming that the proportion of CVID cases attributable to *NFKB1* is equivalent to
the cohort estimate, the literature extrapolated estimate is

$$\text{Estimated NFKB1 cases} \approx 2777 \times 0.041 \approx 114,$$

316 with a median value of approximately 118 and a 95% confidence interval of 70 to 181
317 cases (derived from posterior sampling).

318 **4. Bayesian Adjustment.** Recognising that the clinical cohort likely represents
319 nearly all CVID cases (besides first-second degree relatives), two Bayesian adjust-
320 ments were performed:

1. **Weighted Adjustment (emphasising the cohort, $w = 0.9$):**

$$\text{Adjusted Estimate} = 0.9 \times 16 + 0.1 \times 114 \approx 26,$$

321 with a corresponding 95% confidence interval of approximately 21 to 33 cases.

2. **Mixture Adjustment (equal weighting, $w = 0.5$):** Posterior sampling of
the cohort prevalence was performed assuming

$$p \sim \text{Beta}(16 + 1, 390 - 16 + 1),$$

322 which yielded a Bayesian mixture adjusted median estimate of 67 cases with a
323 95% CrI of approximately 43 to 99 cases.

324 **5. Predicted Total Genotype Counts.** The predicted total synthetic genotype
325 count (before adjustment) was 456, whereas the predicted total genotypes adjusted
326 for `synth_flag` was 0. This higher synthetic count was set based on a minimal risk
327 threshold, ensuring that at least one genotype is assumed to exist (e.g. accounting for
328 a potential unknown de novo variant) even when no variant is observed in gnomAD
329 (as per section 2.2).

330 **6. Validation Test.** Thus, the expected number of *NFKB1*-related CVID cases
331 derived from our genome-wide probability estimates was compared with the observed
332 counts from the UK-based PID cohort. This comparison validates our framework for
333 estimating disease incidence in AD disorders.

334 **2.5 Validation Study for Autosomal Recessive CF Using CFTR**

335 To validate our framework for AR diseases, we focused on Cystic Fibrosis (CF).
336 For comparability sizes between the validation studies, we analysed the most com-
337 mon SNV in the *CFTR* gene, typically reported as “p.Arg117His” (GRCh38 Chr
338 7:117530975 G/A, MANE Select HGVS ENST00000003084.11: p.Arg117His). Our
339 goal was to validate our genome-wide probability estimates by comparing the ex-
340 pected number of CF cases attributable to the p.Arg117His variant in *CFTR* with
341 the nationally reported case count in a well-characterised disease cohort (21–23).

1. Expected Genotype Counts. Let p denote the allele frequency of the p.Arg117His
variant and q denote the combined frequency of all other pathogenic *CFTR* variants,
such that

$$q = P_{\text{tot}} - p \quad \text{with} \quad P_{\text{tot}} = \sum_{i \in \text{CFTR}} p_i.$$

Under Hardy–Weinberg equilibrium for an AR trait, the expected frequencies were:

$$f_{\text{hom}} = p^2 \quad (\text{homozygous for p.Arg117His})$$

and

$$f_{\text{comphet}} = 2pq \quad (\text{compound heterozygotes carrying p.Arg117His and another pathogenic allele}).$$

For a population of size N (here, $N \approx 69\,433\,632$), the expected number of cases were:

$$E_{\text{hom}} = N \cdot p^2, \quad E_{\text{comphet}} = N \cdot 2pq, \quad E_{\text{total}} = E_{\text{hom}} + E_{\text{comphet}}.$$

2. Mortality Adjustment. Since CF patients experience increased mortality, we adjusted the expected genotype counts using an exponential survival model (21–23). With an annual mortality rate $\lambda \approx 0.004$ and a median age of 22 years, the survival factor was computed as:

$$S = \exp(-\lambda \cdot 22).$$

Thus, the mortality-adjusted expected genotype count became:

$$E_{\text{adj}} = E_{\text{total}} \times S.$$

3. Bayesian Uncertainty Simulation. To incorporate uncertainty in the allele frequency p , we modelled p as a beta-distributed random variable:

$$p \sim \text{Beta}(p \cdot AN_{\text{eff}} + 1, AN_{\text{eff}} - p \cdot AN_{\text{eff}} + 1),$$

using a large effective allele count (AN_{eff}) for illustration. By generating 10,000 posterior samples of p , we obtained a distribution of the literature-based adjusted expected counts, E_{adj} .

4. Bayesian Mixture Adjustment. Since the national registry may not capture all nuances (e.g., reduced penetrance) of *CFTR*-related disease, we further combined the literature-based estimate with the observed national count (714 cases from the UK Cystic Fibrosis Registry 2023 Annual Data Report) using a 50:50 weighting:

$$E_{\text{Bayes}} = 0.5 \times (\text{Observed Count}) + 0.5 \times E_{\text{adj}}.$$

5. Validation test. Thus, the expected number of *CFTR*-related CF cases derived from our genome-wide probability estimates was compared with the observed counts from the UK-based CF registry. This comparison validated our framework for estimating disease incidence in AD disorders.

349 **2.6 Validation of SCID-specific Estimates Using PID–SCID
350 Genes**

351 To validate our genome-wide probability estimates for diagnosing a genetic variant in
352 a patient with a PID phenotype, we focused on a subset of genes implicated in Severe
353 Combined Immunodeficiency (SCID). Given that the overall panel corresponds to
354 PID, but SCID represents a rarer subset, the probabilities were converted to values
355 per million PID cases.

356 **1. Incidence Conversion.** Based on literature, PID occurs in approximately 1 in
357 1,000 births, whereas SCID occurs in approximately 1 in 100,000 births. Consequently,
358 in a population of 1,000,000 births there are about 1,000 PID cases and 10 SCID cases.
359 To express SCID-related variant counts on a per-million PID scale, the observed SCID
360 counts were multiplied by 100. For example, if a gene is expected to cause SCID in
361 10 cases within the total PID population, then on a per-million PID basis the count
362 is $10 \times 100 = 1,000$ cases (across all relevant genes).

363 **2. Prevalence Calculation and Data Adjustment.** For each SCID-associated
364 gene (e.g. *IL2RG*, *RAG1*, *DCLRE1C*), the observed variant counts in the dataset were
365 adjusted by multiplying by 100 so that the probabilities reflect the expected number
366 of cases per 1,000,000 PID. In this manner, our estimates are directly comparable to
367 known counts from SCID cohorts, rather than to national population counts as in
368 previous validation studies.

369 **3. Integration with Prior Probability Estimates.** The predicted genotype
370 occurrence probabilities were derived from our framework across the PID gene panel.
371 These probabilities were then converted to expected case counts per million PID
372 cases by multiplying by 1,000,000. For instance, if the probability of observing a
373 pathogenic variant in *IL2RG* is p , the expected SCID-related count becomes $p \times 10^6$.
374 Similar conversions are applied for all relevant SCID genes.

375 **4. Bayesian Uncertainty and Comparison with Observed Data.** To address
376 uncertainty in the SCID-specific estimates, a Bayesian uncertainty simulation was
377 performed for each gene to generate a distribution of predicted case counts on a per-
378 million PID scale. The resulting median estimates and 95% CrIs were then compared
379 against known national SCID counts compiled from independent registries. This
380 comparison permitted a direct evaluation of our framework’s accuracy in predicting
381 the occurrence of SCID-associated variants within a PID cohort.

382 **5. Validation Test.** Thus, by converting the overall probability estimates to a
383 per-million PID scale, our framework was directly validated against observed counts
384 for SCID.

385 2.7 Protein Network and Genetic Constraint Interpretation

386 A PPI network was constructed using protein interaction data from STRINGdb (16).
387 We previously prepared and reported on this dataset consisting of 19,566 proteins and
388 505,968 interactions (<https://github.com/DylanLawless/ProteoMCLustR>). Node
389 attributes were derived from log-transformed score-positive-total values, which in-
390 formed both node size and colour. Top-scoring nodes (top 15 based on score) were
391 labelled to highlight prominent interactions. To evaluate group differences in score-
392 positive-total across major disease categories, one-way Analysis of Variance (ANOVA)
393 was performed followed by Tukey Honestly Significant Difference (HSD) post hoc tests
394 (and non-parametric Dunn's test for confirmation). GnomAD v4.1 constraint metrics
395 data was used for the PPI analysis and was sourced from Karczewski et al. (7). This
396 provided transcript-level metrics, such as observed/expected ratios, Loss-Of-function
397 Observed/Expected Upper bound Fraction (LOEUF), pLI, and Z-scores, quantifying
398 loss-of-function and missense intolerance, along with confidence intervals and related
399 annotations for 211,523 observations.

400 2.8 Gene Set Enrichment Test

401 To test for overrepresentation of biological functions, the prioritised genes were com-
402 pared against gene sets from MsigDB (including hallmark, positional, curated, motif,
403 computational, GO, oncogenic, and immunologic signatures) and WikiPathways using
404 hypergeometric tests with FUMA (24; 25). The background set consisted of 24,304
405 genes. Multiple testing correction was applied per data source using the Benjamini-
406 Hochberg method, and gene sets with an adjusted P-value ≤ 0.05 and more than one
407 overlapping gene are reported.

408 2.9 Deriving novel PID classifications by genetic PPI and 409 clinical features

410 We recategorised 315 immunophenotypic features from the original IUIS IEI anno-
411 tations, reducing the original multi-level descriptors (e.g. "decreased cd8, normal or
412 decreased cd4") first to minimal labels (e.g."low") and second to binary outcomes (nor-
413 mal vs. not-normal) for T cells, B cells, neutrophils, and immunoglobulins Each gene
414 was mapped to its PPI cluster derived from STRINGdb and UMAP embeddings from
415 previous steps. We first tested for non-random associations between these four binary
416 immunophenotypes and PPI clusters using χ^2 tests. To generate a data-driven PID
417 classification, we trained a decision tree (rpart) to predict PPI cluster membership
418 from the four immunophenotypic features plus the traditional IUIS Major and Subcat-
419 egory labels. Hyperparameters (complexity parameter = 0.001, minimum split = 10,
420 minimum bucket = 5, maximum depth = 30) were optimised via five-fold cross vali-
421 dation using the caret framework. Terminal node assignments were then relabelled
422 according to each group's predominant abnormal feature profile.

423 2.10 Probability of observing AlphaMissense pathogenicity

424 We obtained the subset pathogenicity predictions from AlphaMissense via the Al-
425 phaFold database and whole genome data from the studies data repository(12; 26).
426 The AlphaMissense data (genome-aligned and amino acid substitutions) were merged
427 with the panel variants based on genomic coordinate and HGVS annotation. Occur-
428 rence probabilities were log-transformed and adjusted (y-axis displaying $\log_{10}(\text{occurrence}$
429 $\text{prob} + 1e-5) + 5$), to visualise the distribution of pathogenicity scores across the
430 residue sequence. A Kruskal-Wallis test was used to compare the observed disease
431 probability across clinical classification groups.

432 3 Results

433 3.1 Observation probability across disease genes

434 Our study integrated large-scale annotation databases with gene panels from Pan-
435 elAppRex to systematically assess disease genes by MOI. By combining population
436 allele frequencies with ClinVar clinical classifications, we computed an expected obser-
437 vation probability for each SNV, representing the likelihood of encountering a variant
438 of a specific pathogenicity for a given phenotype. We report these probabilities for
439 54,814 ClinVar variant classifications across 557 genes (linked dataset (27)).

440 In practice, our approach computed a simple observation probability for every
441 SNV across the genome and was applicable to any disease-gene panel. Here, we fo-
442 cused on panels related to Primary Immunodeficiency or Monogenic Inflammatory
443 Bowel Disease, using PanelAppRex panel ID 398 as a case study. **Figure 1** dis-
444 plays all reported ClinVar variant classifications for this panel. The resulting natural
445 scaling system (-5 to +5) accounts for the frequently encountered combinations of
446 classification labels (e.g. benign to pathogenic). The resulting data set (27) is briefly
447 shown in **Table 1** to illustrate that our method yielded estimations of the probability
448 of observing a variant with a particular ClinVar classification.

449 3.2 Integrating observed true positives and unobserved false 450 negatives into a single, actionable conclusion

451 Having previously established a probabilistic framework for estimating the prior prob-
452 ability of observing disease-associated variants under different inheritance modes, we
453 then applied this model to a specific patient to demonstrate it's potential for clini-
454 cal genetics. For each gene, we used the computed prior probabilities for all variant
455 classifications, (e.g. benign, uncertain, and pathogenic). We verified that all known
456 pathogenic positions have been sequenced and observed as reference (true negatives),
457 and identify any positions that were either observed as variant (true positives) or
458 not assessable due to missing sequence data of failed Quality Control (QC). These

Table 1: **Example of the first several rows from our main results for 557 genes of PanelAppRex’s panel: (ID 398) Primary immunodeficiency or monogenic inflammatory bowel disease.** “ClinVar Significance” indicates the pathogenicity classification assigned by ClinVar, while “Occurrence Prob” represents our calculated probability of observing the corresponding variant class for a given phenotype. MOI shows the gene-disease-specific mode of inheritance. Additional columns, such as population allele frequency, are not shown. (27)

Gene	Panel ID	ClinVar Clinical Significance	GRCh38 Pos	HGVSc	HGVSp	MOI	Occurrence Prob
ABI3	398	Uncertain significance	49210771	c.47G>A	p.Arg16Gln	AR	0.000000007
ABI3	398	Uncertain significance	49216678	c.265C>T	p.Arg89Cys	AR	0.000000005
ABI3	398	Uncertain significance	49217742	c.289G>A	p.Val97Met	AR	0.000000002
ABI3	398	Uncertain significance	49217781	c.328G>A	p.Gly110Ser	AR	0.000000002
ABI3	398	Uncertain significance	49217844	c.391C>T	p.Pro131Ser	AR	0.000000015
ABI3	398	Uncertain significance	49220257	c.733C>G	p.Pro245Ala	AR	0.000000022
...

missing sites represent potential false negatives. By jointly modelling the observed and unobserved space, the method yielded a calibrated, evidence-weighted probability that at least one damaging causal variant could be present in the gene.

We present the results from three scenarios for an example single-case patient being investigated for the genetic diagnosis of IEI. **Figure S1** shows the results of the first scenario, in which only one known pathogenic variant, *NFKB1* p.Ser237Ter, was observed and all other previously reported pathogenic positions were successfully sequenced and confirmed as reference. In this setting, the model assigned the full posterior probability to the observed allele, yielding 100 % confidence that all present evidence supported a single, true positive causal explanation. The most strongly supported observed variant was p.Ser237Ter (posterior: 0.594). The strongest (probability of observing) unsequenced variant was a benign variant p.Thr567Ile (posterior: 0). The total probability of a causal diagnosis given the available evidence was 1 (95% CI: 1–1) (**Table S1**).

Figure 2 shows the second scenario, where the same pathogenic variant *NFKB1* p.Ser237Ter was present, but coverage was incomplete at three additional sites known classified variants. Among these was the likely-pathogenic splice-site variant c.159+1G>A, which was not captured in the sequencing data. The panels of **Figure 2 (A–F)** illustrate the stepwise integration of observed and missing evidence, culminating in a posterior probability that reflects both confirmed findings and residual uncertainty. **Table 2** lists the metrics used to reach the conclusion for reporting the clinical genetics results.

Bayesian integration of every annotated allele yielded the quantitative CrIs for pathogenic attribution that (i) preserve Hardy-Weinberg expectations, (ii) accommodate AD, AR, XL inheritance, and (iii) carry explicit uncertainty for non-sequenced (or failed QC) genomic positions. **Figure 2 (A)** depicts the prior landscape where occurrence probabilities are partitioned by observed or missing status and by causal or non-causal evidence, with colour reflecting the underlying ClinVar score. **Fig-**

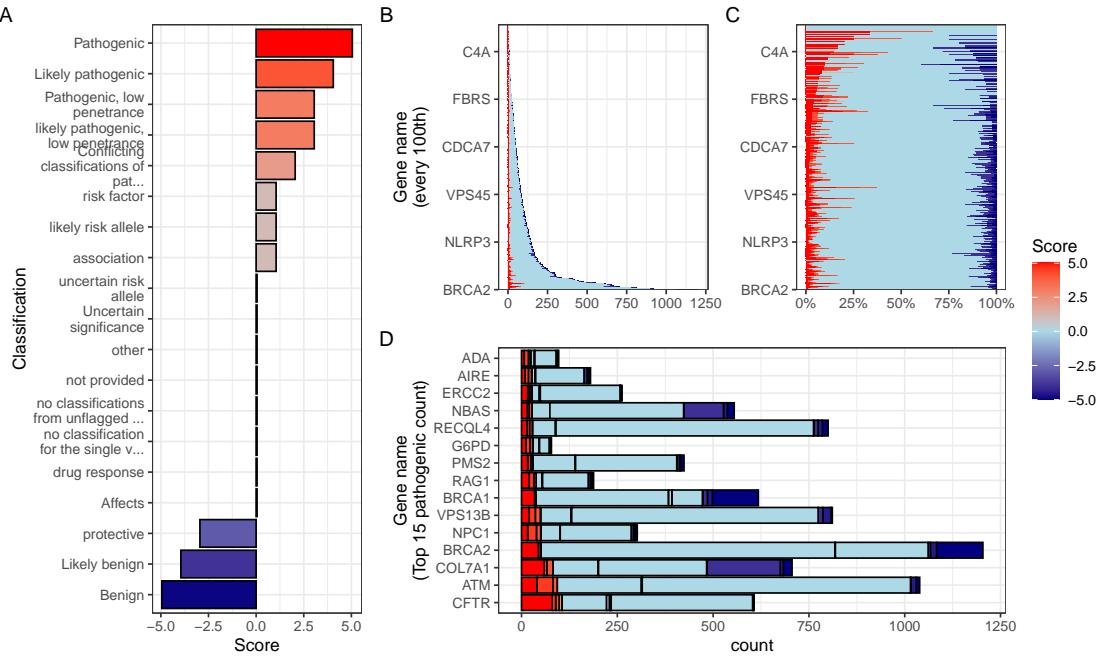


Figure 1: Summary of ClinVar clinical significance classifications in the PID gene panel. (A) Shows the numeric score coding for each classification. Panels (B) and (C) display the tally of classifications per gene as absolute counts and as percentages, respectively. (D) Highlights the top 15 genes with the highest number of reported pathogenic classifications (score 5).

487 **Figure 2 (B)** shows posterior normalisation which concentrates probability density on
 488 two high-confidence (high evidence score) alleles since the benign variants are, by
 489 definition, non-causal. **Figure 2 (C)** shows the resulting per-variant probability
 490 of being simultaneously damaging and causal; only the confirmed present (true posi-
 491 tive) nonsense variant p.Ser237Ter and the (false negative) hypothetical splice-donor
 492 c.159+1G>A retain substantial support. Restricting the view to causal candidates in
 493 **Figure 2 (D)** confirms that posterior mass is distributed across these two variants.
 494 **Figure 2 (E)** decomposes the total damaging probability into observed (approxi-
 495 mately 40 %) and missing (approximately 34 %) sources, whereas **Figure 2 (F)** sum-
 496 marises the gene-level posterior: inclusion of the splice-site allele (scenario 2) produces
 497 a median probability of 0.542 with a 95 % CrI of 0.264–0.8, compared with 1 (1–1)
 498 when the analysis is limited to sequenced alleles (scenario 1). Numerically, the present
 499 variant p.Ser237Ter accounts for 0.399 of the posterior share, and the potentially
 500 causal but missing splice-donor allele c.159+1G>A contributes 0.339. The remaining
 501 alleles together explain a negligible share (**Table 2**).

502 **Figure S2** shows the third scenario, in which no observed variants were detected
 503 in the proband (including *NFKB1*). Instead, unsequenced alleles from each major
 504 ClinVar classification emulated a broad range of plausible FN for the gene Tumor
 505 necrosis factor, alpha-induced protein 3 (*TNFAIP3*). The strongest (probability of

506 observing and pathogenic) unsequenced variant was p.Cys243Arg (posterior: 0.347).
 507 However, the total probability of a causal diagnosis for the patient given the available
 508 evidence was 0 (95% CI: 0–0) since these missing variants must be accounted for
 509 (**Table S2**).

Table 2: Result of clinical genetics diagnosis scenario 2. The most strongly supported observed variant was p.Ser237Ter (posterior: 0.399). The strongest unsequenced variant was c.159+1G>A (posterior: 0.339). The total probability of a causal diagnosis given the available evidence was 0.542 (95% CI: 0.264–0.8).

Variant	Flag	Class	Evidence Score	Occurrence Prob	Adj Occ Prob	Alpha	Beta	Lower	Median	Upper	Posterior Share	Prob Causal
p.Ser237Ter	present	causal	5.0	0.000	0	6.0	345	0.003	0.092	0.576	0.399	0.399
c.159+1G>A	missing	causal	4.5	0.000	0	5.5	373	NA	NA	NA	0.339	0.339
p.Thr567Ile	missing	other	-5.0	0.002	0	1.0	353	NA	NA	NA	0.000	0.000
p.Arg231His	present	other	0.0	0.000	0	1.0	347	0.003	0.092	0.576	0.000	0.000
p.Gly650Arg	present	other	0.0	0.000	0	1.0	359	0.003	0.092	0.576	0.000	0.000
p.Val236Ile	missing	other	0.0	0.000	0	1.0	361	NA	NA	NA	0.000	0.000
Total	NA	NA	NA	NA	NA	NA	NA	0.264	0.542	0.800	NA	0.542

Gene: NFKB1

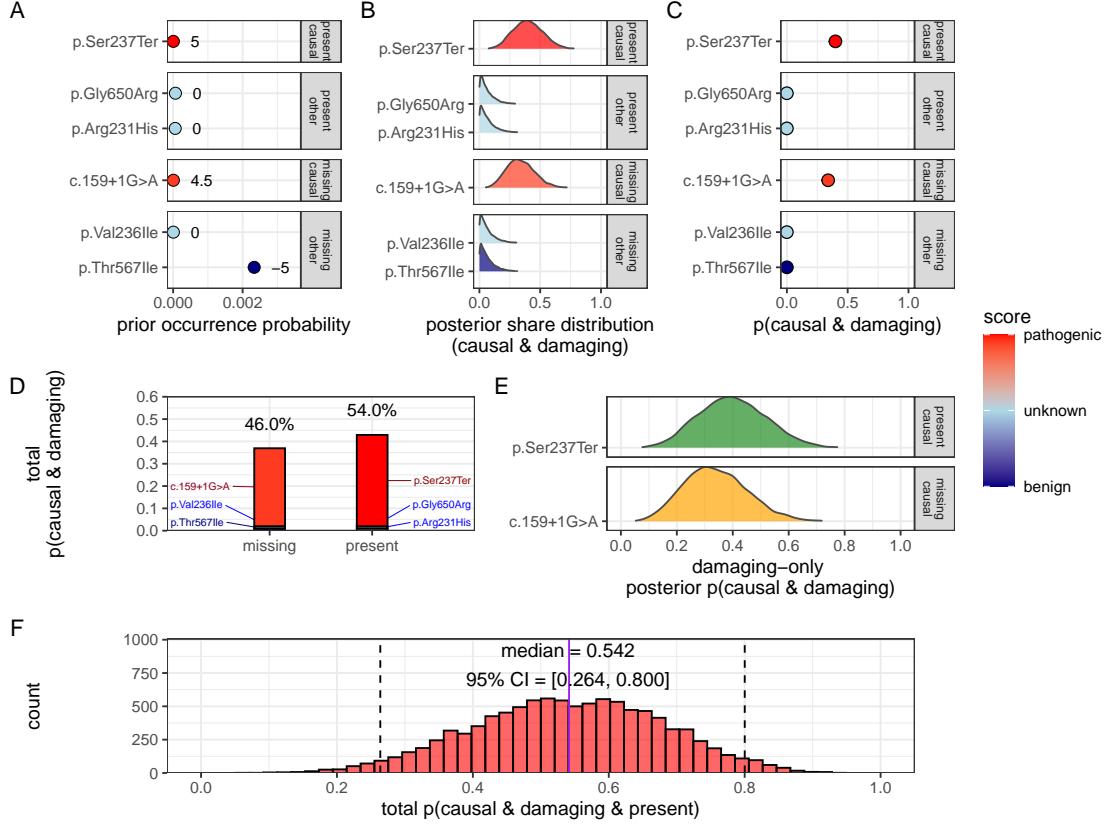


Figure 2: **Quantification of present (TP) and missing (FN) causal genetic variants for disease in *NFKB1* (scenario 2).** The example proband carried three known heterozygous variants, including pathogenic p.Ser237Ter, and had incomplete coverage at three additional loci, including likely-pathogenic splice-site variant c.159+1G>A. The sequential steps towards the posterior probability of complete genetic diagnosis are shown: (A) Prior occurrence probabilities, stratified by observed/missing and causal/non-causal status. Pathogenicity scores (-5 to +5) are annotated. (B) Posterior distributions of normalised variant weights $\tilde{\pi}_i$. (C) Per-variant posterior probability of being both damaging and causal. (D) Posterior distributions for causal variants only. (E) Decomposition of total pathogenic probability into observed (green) and missing (orange) sources. (F) Gene-level posterior showing the probability that at least one damaging causal allele is present; median 0.54, 95 % CrI 0.26-0.80

510 **3.3 Validation studies**

511 **3.3.1 Validation of dominant disease occurrence with *NFKB1***

512 To validate our genome-wide probability estimates for AD disorders, we focused
513 on *NFKB1*. We used a reference dataset from Tuijnenburg et al. (17), in which
514 whole-genome sequencing of 846 PID patients identified *NFKB1* as one of the genes
515 most strongly associated with the disease, with 16 *NFKB1*-related CVID cases at-
516 tributed to AD heterozygous variants. Our goal was to compare the predicted num-
517 ber of *NFKB1*-related CVID cases with the reported count in this well-characterised
518 national-scale cohort.

519 Our model calculated 0 known pathogenic variant *NFKB1*-related CVID cases
520 in the UK with a minimal risk of 456 unknown de novo variants. In the reference
521 cohort, 16 *NFKB1* CVID cases were reported. We additionally wanted to account for
522 potential under-reporting in the reference study. We used an extrapolated national
523 CVID prevalence which yielded a median estimate of 118 cases (95% CI: 70–181),
524 while a Bayesian-adjusted mixture estimate produced a median of 67 cases (95% CI:
525 43–99). **Figure S3 (A)** illustrates that our predicted values reflect these ranges and
526 are closer to the observed count. This case supports the validity of our integrated
527 probability estimation framework for AD disorders, and represents a challenging ex-
528 ample where pathogenic SNV are not reported in the reference population of gnomAD.
529 Our min-max values successfully contained the true reported values.

530 **3.3.2 Validation of recessive disease occurrence with *CFTR***

531 Our analysis predicted the number of CF cases attributable to carriage of the p.Arg117His
532 variant (either as homozygous or as compound heterozygous with another pathogenic
533 allele) in the UK. Based on HWE calculations and mortality adjustments, we pre-
534 dicted approximately 648 cases arising from biallelic variants and 160 cases from
535 homozygous variants, resulting in a total of 808 expected cases.

536 In contrast, the nationally reported number of CF cases was 714, as recorded
537 in the UK Cystic Fibrosis Registry 2023 Annual Data Report (21). To account for
538 factors such as reduced penetrance and the mortality-adjusted expected genotype,
539 we derived a Bayesian-adjusted estimate via posterior simulation. Our Bayesian ap-
540 proach yielded a median estimate of 740 cases (95% CI: 696, 786) and a mixture-
541 based estimate of 727 cases (95% CI: 705, 750). **Figure S3 (B)** illustrates the close
542 concordance between the predicted values, the Bayesian-adjusted estimates, and the
543 national report supports the validity of our approach for estimating disease.

544 **Figure S4** shows the final values for these genes of interest in a given population
545 size and phenotype. It reveals that an allele frequency threshold of approximately
546 0.000007 is required to observe a single heterozygous disease-causing variant carrier in
547 the UK population for both genes. However, owing to the AR MOI pattern of *CFTR*,
548 this threshold translates into more than 100,000 heterozygous carriers, compared to

549 only 456 carriers for the AD gene *NFKB1*. Note that this allele frequency threshold,
550 being derived from the current reference population, represents a lower bound that
551 can become more precise as public datasets continue to grow. This marked difference
552 underscores the significant impact of MOI patterns on population carrier frequencies
553 and the observed disease prevalence.

554 **3.3.3 Interpretation of ClinVar variant observations**

555 **Figure S11** shows the two validation study PID genes, representing AR and domi-
556 nant MOI. **Figure S11 (A)** illustrates the overall probability of an affected birth by
557 ClinVar variant classification, whereas **Figure S11 (B)** depicts the total expected
558 number of cases per classification for an example population, here the UK, of approx-
559 imately 69.4 million.

560 **3.3.4 Validation of SCID-specific disease occurrence**

561 Given that SCID is a subset of PID, our probability estimates reflect the likelihood of
562 observing a genetic variant as a diagnosis when the phenotype is PID. However, we
563 additionally tested our results against SCID cohorts in **Figure S6**. The summarised
564 raw cohort data for SCID-specific gene counts are summarised and compared across
565 countries in **Figure S5**. True counts for *IL2RG* and *DCLRE1C* from ten distinct
566 locations yielded 95% confidence intervals surrounding our predicted values. For
567 *IL2RG*, the prediction was low (approximately 1 case per 1,000,000 PID), as expected
568 since loss-of-function variants in this X-linked gene are highly deleterious and rarely
569 observed in gnomAD. In contrast, the predicted value for *RAG1* was substantially
570 higher (553 cases per 1,000,000 PID) than the observed counts (ranging from 0 to
571 200). We attributed this discrepancy to the lower penetrance and higher background
572 frequency of *RAG1* variants in recessive inheritance, whereby reference studies may
573 underreport the true national incidence. Overall, we argued that agreement within
574 an order of magnitude was tolerable given the inherent uncertainties from reference
575 studies arising from variable penetrance and allele frequencies.

576 **3.4 Genetic constraint in high-impact protein networks**

577 We next examined genetic constraint in high-impact protein networks across the whole
578 IEI gene set of over 500 known disease-gene phenotypes (1). By integrating ClinVar
579 variant classification scores with PPI data, we quantified the pathogenic burden per
580 gene and assessed its relationship with network connectivity and genetic constraint
581 (7; 16).

582 **3.4.1 Score-positive-total within IEI PPI network**

583 The ClinVar classifications reported in **Figure 1** were scaled -5 to +5 based on their
 584 pathogenicity. We were interested in positive (potentially damaging) but not negative
 585 (benign) scoring variants, which are statistically incidental in this analysis. We tallied
 586 gene-level positive scores to give the score positive total metric. **Figure 3 (A)** shows
 587 the PPI network of disease-associated genes, where node size and colour encode the
 588 score positive total (log-transformed). The top 15 genes with the highest total prior
 589 probabilities of being observed with disease are labelled (as per **Figure 1**).

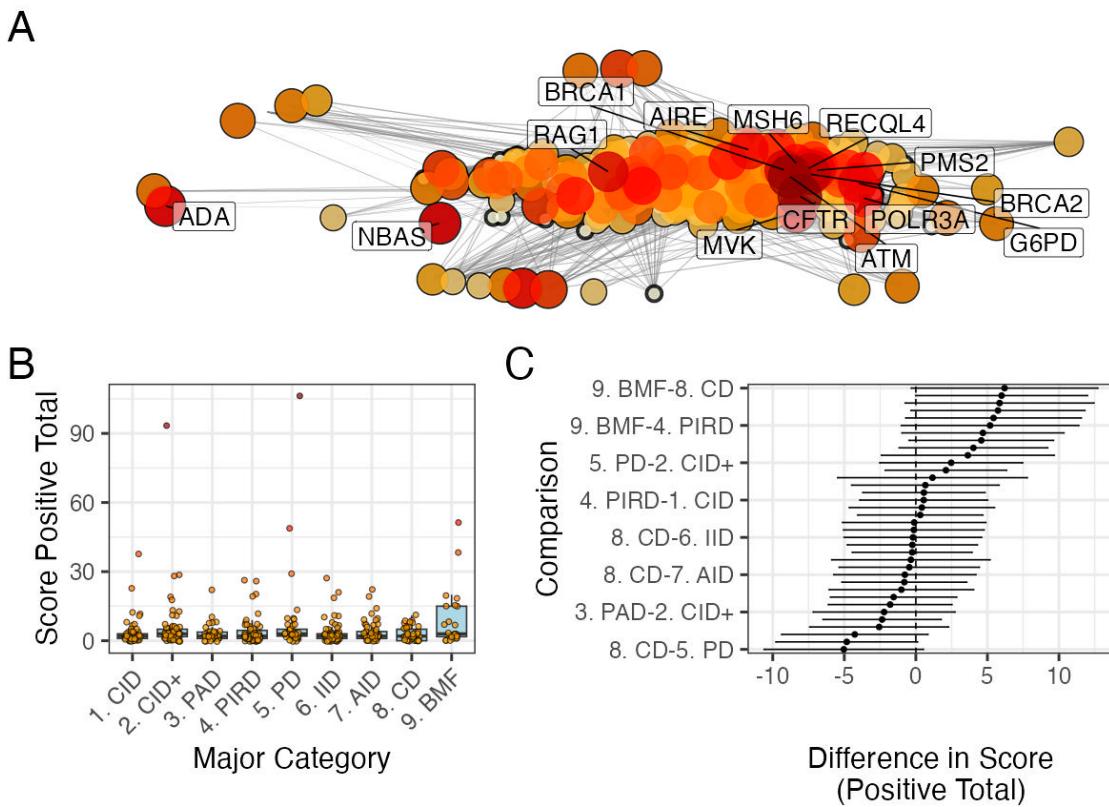


Figure 3: **PPI network and score positive total ClinVar significance variants.** (A) PPI network of disease-associated genes. Node size and colour represent the log-transformed score positive total, the top 15 genes/proteins with the highest probability of being observed in disease are labelled. (B) Distribution of score positive total across the major IEI disease categories. (C) Tukey HSD comparisons of mean differences in score positive total among all pairwise disease categories. Every 5th label is shown on y-axis.

590 **3.4.2 Association analysis of score-positive-total across IEI categories**

591 We checked for any statistical enrichment in score positive totals, which represents
 592 the expected observation of pathogenicity, between the IEI categories. The one-way

593 ANOVA revealed an effect of major disease category on score positive total ($F(8, 500) =$
594 2.82, $p = 0.0046$), indicating that group means were not identical, which we observed
595 in **Figure 3 (B)**. However, despite some apparent differences in median scores across
596 categories (i.e. 9. Bone Marrow Failure (BMF)), the Tukey HSD post hoc compar-
597 isons **Figure 3 (C)** showed that all pairwise differences had 95% CIs overlapping
598 zero, suggesting that individual group differences were not significant.

599 **3.4.3 UMAP embedding of the PPI network**

600 To address the density of the PPI network for the IEI gene panel, we applied Uniform
601 Manifold Approximation and Projection (UMAP) (**Figure 4**). Node sizes reflect
602 interaction degree, a measure of evidence-supported connectivity (16). We tested
603 for a correlation between interaction degree and score positive total. In **Figure**
604 **4**, gene names with degrees above the 95th percentile are labelled in blue, while
605 the top 15 genes by score positive total are labelled in yellow (as per **Figure 1**).
606 Notably, genes with high pathogenic variant loads segregated from highly connected
607 nodes, suggesting that Loss-of-Function (LOF) in hub genes is selectively constrained,
608 whereas damaging variants in lower-degree genes yield more specific effects. This
609 observation was subsequently tested empirically.

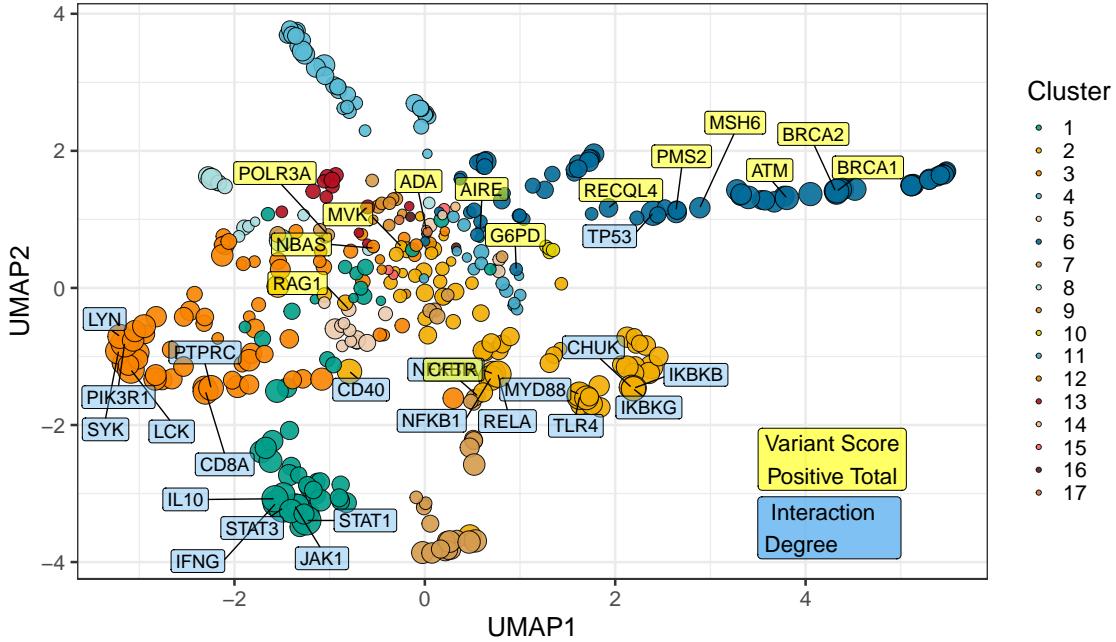


Figure 4: **UMAP embedding of the PPI network (p_umap).** The plot projects the high-dimensional protein-protein interaction network into two dimensions, with nodes coloured by cluster and sized by interaction degree. Blue labels indicate hub genes (degree above the 95th percentile) and yellow labels mark the top 15 genes by score positive total (damaging ClinVar classifications). The spatial segregation suggests that genes with high pathogenic variant loads are distinct from highly connected nodes.

610 3.4.4 Hierarchical clustering of enrichment scores for major disease cate-
611 gories

612 Figure S7 presents a heatmap of standardised residuals for major disease categories
613 across network clusters, as per Figure 4. A dendrogram clusters similar disease cate-
614 gories, while the accompanying bar plot displays the maximum absolute standardised
615 residual for each category. Notably, (8) Complement Deficiencies (CD) shows the
616 highest maximum enrichment, followed by (9) BMF. While all maximum values
617 exceed 2, the threshold for significance, this likely reflects the presence of protein
618 clusters with strong damaging variant scores rather than uniform significance across
619 all categories (i.e. genes from cluster 4 in 8 CD).

620 3.4.5 PPI connectivity, LOEUF constraint and enriched network cluster
621 analysis

622 Based on the preliminary insight from Figure S7, we evaluated the relationship
623 between network connectivity (PPI degree) and LOEUF constraint (LOEUF upper rank)
624 Karczewski et al. (7) using Spearman's rank correlation. Overall, there was a weak

625 but significant negative correlation ($\rho = -0.181$, $p = 0.00024$) at the global scale,
 626 indicating that highly connected genes tend to be more constrained. A supplementary
 627 analysis (**Figure S8**) did not reveal distinct visual associations between network
 628 clusters and constraint metrics, likely due to the high network density. However
 629 once stratified by gene clusters, the natural biological scenario based on quantitative
 630 PPI evidence (16), some groups showed strong correlations; for instance, cluster 2
 631 ($\rho = -0.375$, $p = 0.000994$) and cluster 4 ($\rho = -0.800$, $p < 0.000001$), while others did
 632 not. This indicated that shared mechanisms within pathway clusters may underpin
 633 genetic constraints, particularly for LOF intolerance. We observe that the score
 634 positive total metric effectively summarises the aggregate pathogenic burden across
 635 IEI genes, serving as a robust indicator of genetic constraint and highlighting those
 636 with elevated disease relevance.

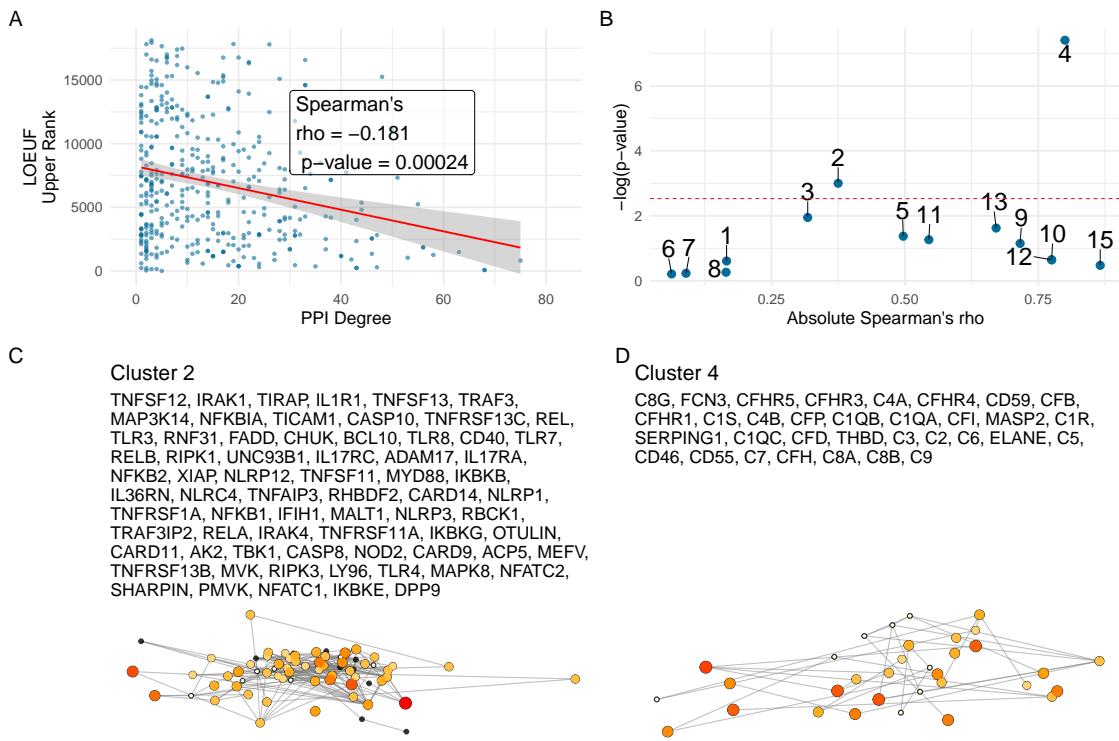


Figure 5: **Correlation between PPI degree and LOEUF upper rank.** (A) Ananlysis across all genes revealed a weak, significant negative correlation between PPI degree and LOEUF upper rank. (B) The cluster-wise analysis showed that clusters 2 and 4 exhibited moderate to strong correlations, while other clusters display weak or non-significant relationships. (C) and (D) Shows the new network plots for the significantly enriched clusters based on gnomAD constraint metrics.

637 **Figure 5 (C, D)** shows the re-plotted PPI networks for clusters with significant
 638 correlations between PPI degree and LOEUF upper rank. In these networks, node
 639 size is scaled by a normalised variant score, while node colour reflects the variant
 640 score according to a predefined palette.

641 **3.5 New insight from functional enrichment**

642 To interpret the functional relevance of our prioritised IEI gene sets with the highest
643 load of damaging variants (i.e. clusters 2 and 4 in **Figure 5**), we performed func-
644 tional enrichment analysis for known disease associations using MsigDB with FUMA
645 (i.e. GWAScatalog and Immunologic Signatures) (24). Composite enrichment pro-
646 files (**Figure S9**) reveal that our enriched PPI clusters were associated with distinct
647 disease-related phenotypes, providing functional insights beyond traditional IUIS IEI
648 groupings (1). The gene expression profiles shown in **Figure S10** (GTEx v8 54
649 tissue types) offer the tissue-specific context for these associations. Together, these
650 results enable the annotation of IEI gene sets with established disease phenotypes,
651 supporting a data-driven classification of IEI.

652 Based on these independent sources of interpretation, we observed that genes
653 from cluster 2 were independently associated with specific inflammatory phenotypes,
654 including ankylosing spondylitis, psoriasis, inflammatory bowel disease, and rheuma-
655 toid arthritis, as well as quantitative immune traits such as lymphocyte and neutrophil
656 percentages and serum protein levels. In contrast, genes from Cluster 4 were linked
657 to ocular and complement-related phenotypes, notably various forms of age-related
658 macular degeneration (e.g. geographic atrophy and choroidal neovascularisation) and
659 biomarkers of the complement system (e.g. C3, C4, and factor H-related proteins),
660 with additional associations to nephropathy and pulmonary function metrics.

661 **3.6 Genome-wide gene distribution and locus-specific variant**
662 **occurrence**

663 **Figure 6 (A)** shows a genome-wide karyoplot of all IEI panel genes across GRCh38,
664 with colour-coding based on MOI. Figures **(B)** and **(C)** display zoomed-in locus plots
665 for *NFKB1* and *CFTR*, respectively. In **Figure 6 (B)**, the probability of observing
666 variants with known classifications is high only for variants such as p.Ala475Gly,
667 which are considered benign in the AD *NFKB1* gene that is intolerant to LOF. In
668 **Figure 6 (C)**, high probabilities of observing patients with pathogenic variants in
669 *CFTR* are evident, reproducing this well-established phenomenon. Furthermore, the
670 analysis of Linkage Disequilibrium (LD) using R^2 shows that high LD regions can be
671 modelled effectively, allowing independent variant signals to be distinguished.

672 **3.7 Novel PID classifications derived from genetic PPI and**
673 **clinical features**

674 We recategorised 315 immunophenotypic features from the original IUIS IEI annota-
675 tions, reducing detailed descriptions (e.g. “decreased cd8, normal or decreased cd4”),
676 first to minimal labels (e.g.“low”), and second to binary outcomes (normal vs. not-
677 normal) for T cells, B cells, neutrophils, and immunoglobulins (**Figure 7**). These

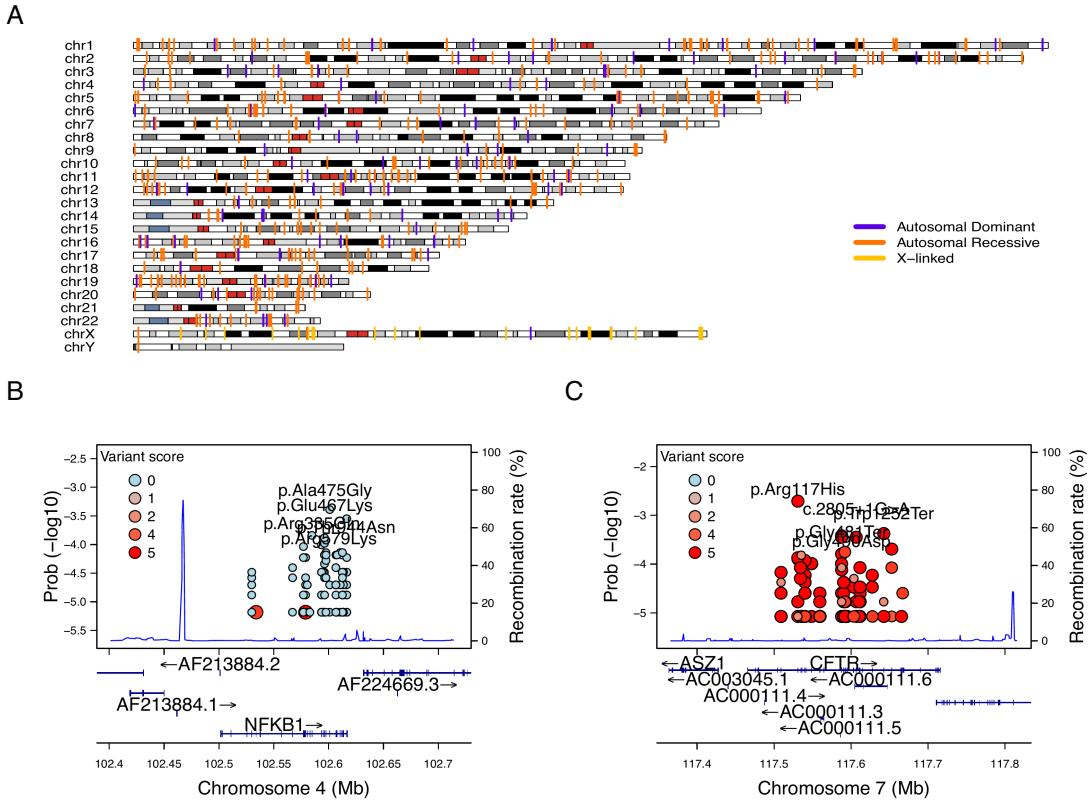


Figure 6: Genome-wide IEI, variant occurrence probability and LD by R^2 .
 (A) Genome-wide karyoplot of all IEI panel genes mapped to GRCh38, with colours indicating MOI. (B) Zoomed-in locus plot for *NFKB1* showing variant observation probabilities; only benign variants such exhibit high probabilities in this AD gene intolerant to LOF. (C) Locus plot for *CFTR* displaying high probabilities for pathogenic variants; due to the dense clustering of pathogenic variants, score filter >0 was applied. Top five variant are labelled per gene.

simplified profiles were integrated with PPI network clustering from STRINGdb to refine PID gene groupings. Chi-square analyses confirmed significant associations between specific clinical abnormalities and PPI clusters (**Figure ??**). A decision tree classifier, with hyperparameters optimised via 5-fold cross validation, demonstrated high sensitivity and specificity, as shown in the confusion matrices and variable importance metrics (**Figure S12**). The resulting novel PID classifications, illustrated by the decision tree and gene group distributions (**Figure 10**), provide a more coherent and data-driven framework for categorising PID genes.

686 **3.8 Novel PID classifications derived from genetic PPI and**
687 **clinical features**

688 We recategorised 315 immunophenotypic features from the original IUIS IEI annotations,
689 reducing detailed descriptions (e.g. “decreased CD8, normal or decreased CD4”) to minimal labels (e.g. “low”) and then binarising them (normal vs. not-normal) for T
690 cells, B cells, Immunoglobulin (Ig) and neutrophils (**Figure 7**). These simplified profiles
691 were mapped onto STRINGdb PPI clusters, revealing non-random distributions
692 ($\chi^2 < 1e-13$; **Figure 8**), indicating that network context captures key immunopheno-
693 notypic variation.

695 We next compared four classifiers under 5-fold cross-validation to determine which
696 features predicted PPI clustering. As shown in **Figure 9**, the fully combined model
697 achieved the highest accuracy among the four: (i) phenotypes only (33 %) (i.e. T
698 cell, B cell, Ig, Neutrophil); (ii) phenotypes + IUIS major category (50 %) (e.g. CID.
699 See **Box 2.1** for more); (iii) IUIS major + subcategory only (59 %) (e.g. CID, T-B+
700 SCID); and (iv) phenotypes + IUIS major + subcategory (61 %). This demonstrated
701 that incorporating both traditional IUIS classifications and core immunophenotypic
702 markers into the PPI-based framework yielded the most robust discrimination of PID
703 gene clusters. Variable importance analysis highlighted abnormality status for Ig and
704 T cells were among the top ten features in addition to the other IUIS major and sub
705 categories. Per-class specificity remained uniform across the classes while sensitivity
706 dropped.

707 The PPI and immunophenotype model yielded 17 data-driven PID groups, whereas
708 incorporating the full complement of IUIS categories expanded this to 33 groups. For
709 clarity, we only demonstrate the decision tree from the smaller 17-group model in
710 **Figure 10**. Each terminal node is annotated by its predominant immunophenotypic
711 signature (for example, “group 65 with abnormal T cell and B cell features”), and the
712 full resulting gene counts per 33 class are plotted in **Figure 10**. Although, less user-
713 friendly, this data-driven taxonomy both aligns with and refines traditional IUIS IEI
714 classifications to provide a scaffold for large-scale computational analyses. Because
715 this framework is fully reproducible, alternative PPI embeddings or incorporate additional
716 molecular annotations can readily swapped to continue building on these PID
717 classification schemes.

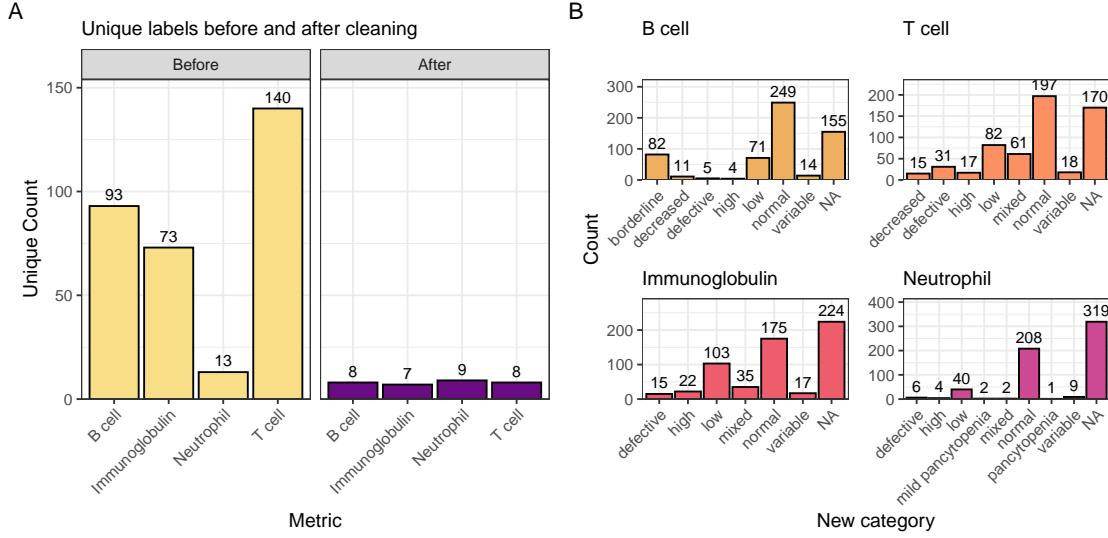


Figure 7: Distribution of immunophenotypic features before and after recategorisation. The original IUIS IEI descriptions contain information such as T cell-related “decreased cd8, normal or decreased cd4 cells” which we recategorise as “low”. The bar plot shows the count of unique labels for each status (normal, not_normal) across the T cell, B cell, Ig, and Neutrophil features.

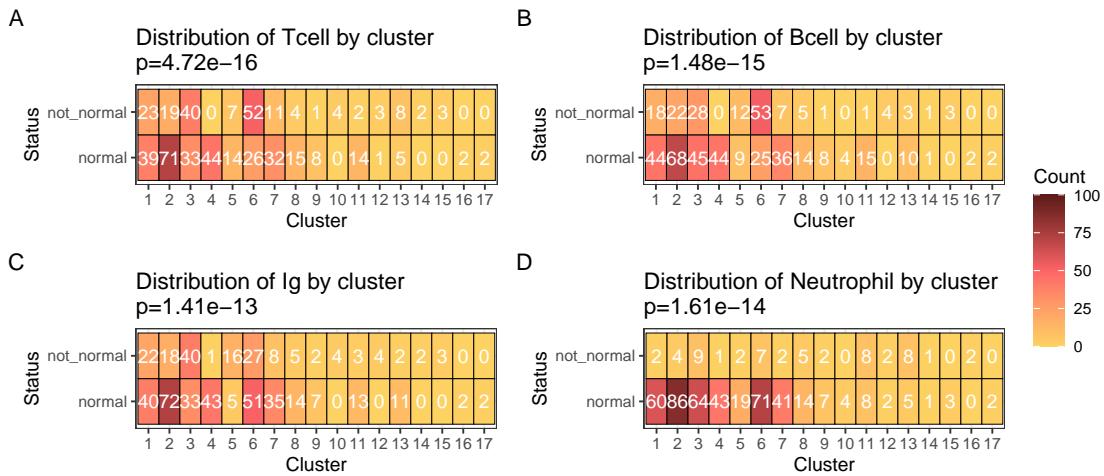


Figure 8: Heatmaps of clinical feature distributions by PPI cluster. The heatmaps display the count of observations for abnormality of each clinical feature (A) T cell, (B) B cell, (C) Immunoglobulin, (D) Neutrophil, in relation to the PPI clusters, with p-values from chi-square tests annotated in the titles.

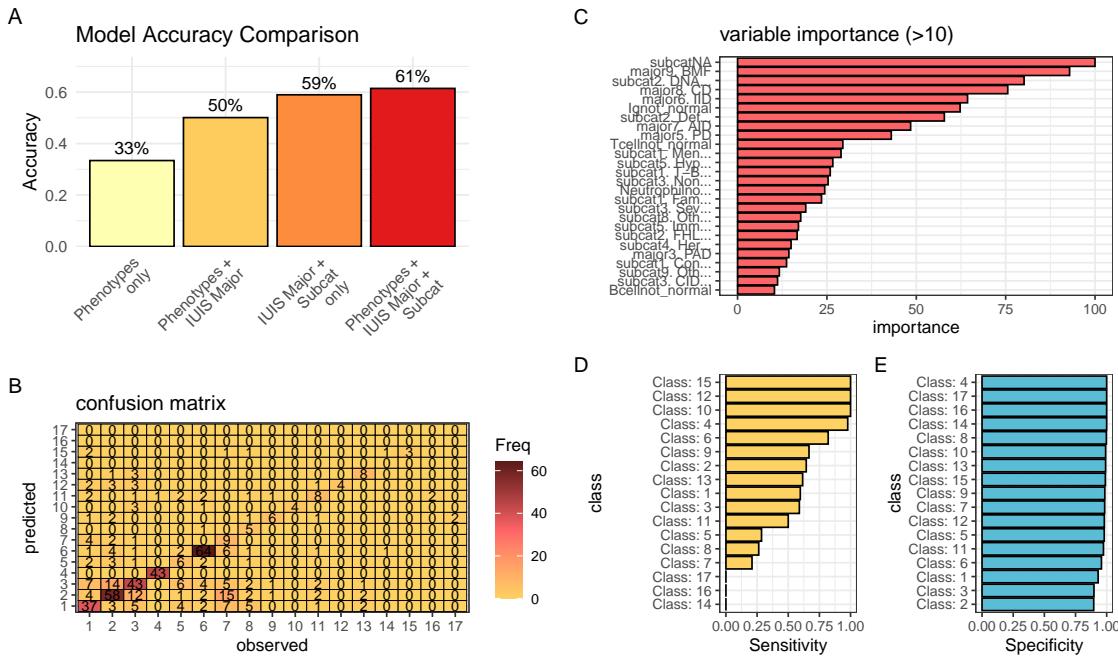


Figure 9: Performance comparison of PID classifiers. Classification predicting PPI cluster membership from IUIS major category, subcategory, and immunological features. (A) Overall accuracy for four rpart models used to predict PPI clustering. The combined model achieves 61.4 % accuracy, exceeding all simpler approaches. Nodes were split to minimize Gini impurity, pruned by cost-complexity ($cp = 0.001$), and validated via 5-fold cross-validation. (B-E) The summary statistics from the top model are detailed.

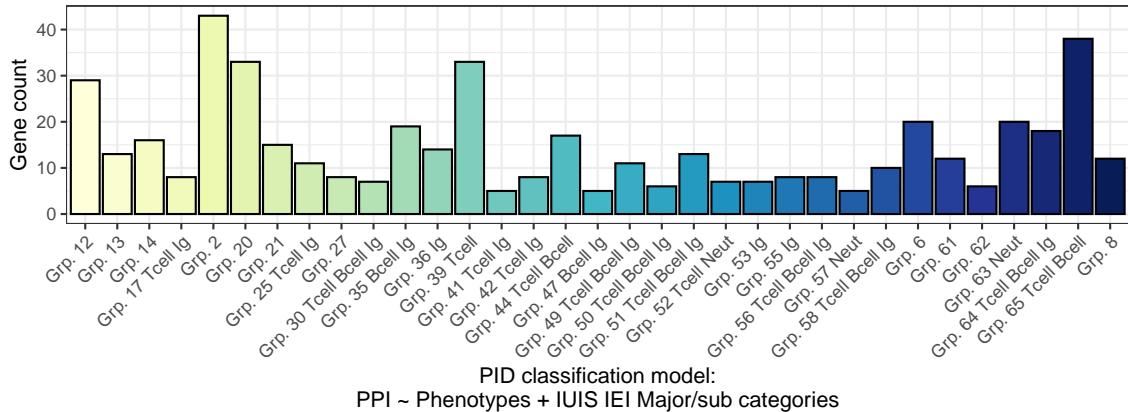
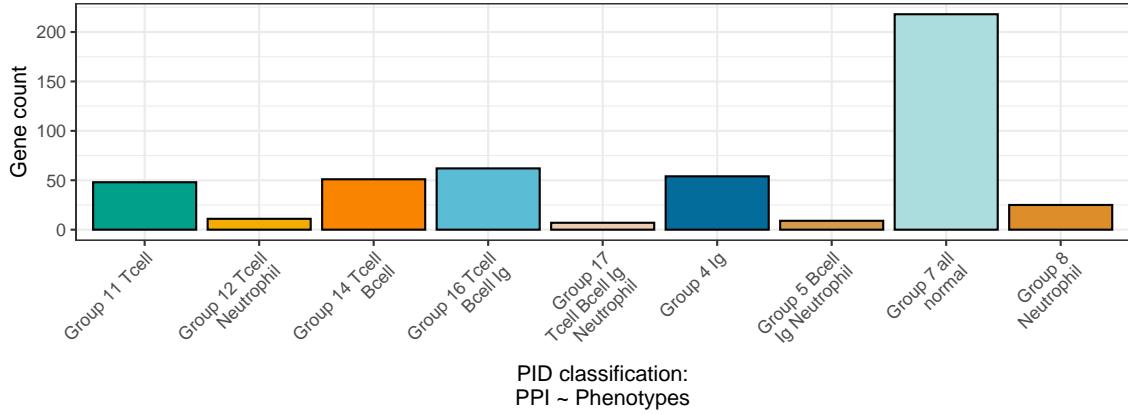
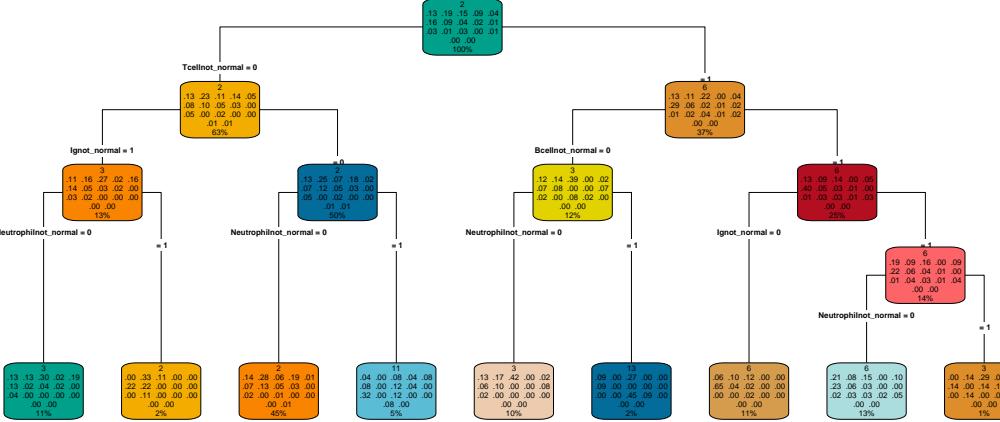


Figure 10: **Fine-tuned model for PID classification.** (Top) In each terminal node, the top block indicates the number of genes in the node; the middle block shows the fitted class probabilities (which sum to 1); and the bottom block displays the percentage of the total sample in that node. These metrics summarise the model's assignment based on immunophenotypic and PPI features. (Middle) Bar plot presenting the distribution of novel PID classifications, where group labels denote the predominant abnormal clinical feature(s) (e.g. T cell, B cell, Ig, Neutrophil) characterising each group. (Bottom) The complete model including the traditional IUIS IEI categories.

718 **3.8.1 Integration of Variant Probabilities into IEI Genetics Data**

719 We integrated the computed prior probabilities for observing variants in all known
 720 genes associated with a given phenotype (1), across AD, AR, and XL MOI, into
 721 our IEI genetics framework. These calculations, derived from gene panels in Pan-
 722 elAppRex, have yielded novel insights for the IEI disease panel. The final result
 723 comprised of machine- and human-readable datasets, including the table of variant
 724 classifications and priors available via a the linked repository (27), and a user-friendly
 725 web interface that incorporates these new metrics.

726 **Figure 11** shows the interface summarising integrated variant data. Server-side
 727 pre-calculation of summary statistics minimises browser load, while clinical signifi-
 728 cance is converted to numerical metrics. Key quantiles (min, Q1, median, Q3, max)
 729 for each gene are rendered as sparkline box plots, and dynamic URLs link table entries
 730 to external databases (e.g. ClinVar, Online Mendelian Inheritance in Man (OMIM),
 731 AlphaFold).

The screenshot shows a table titled "Viewer Zoom" with a search bar at the top. The table has 13 columns: Major category, Subcategory, Disease, Genetic defect, Inheritance, Gene score, Prior prob of pathogenicity, ClinVar SNV classification, ClinVar all variant reports, OMIM, Alpha Missense / Uniprot ID, HPO combined, and HPO term. The "Gene score" column contains sparkline box plots. The "Prior prob of pathogenicity" column also contains sparkline box plots. The "ClinVar SNV classification" column lists variant types like "benign", "pathogenic", and "unknown". The "ClinVar all variant reports" column lists report IDs like 186780, P04234, P07766, P31146, 308380, P16871, P27987, P52333, and O43561. The "OMIM" column lists OMIM IDs like P20963, P04234, P07766, P31146, 308380, P16871, P27987, P52333, and O43561. The "Alpha Missense / Uniprot ID" column lists UniProt IDs like P20963, P04234, P07766, P31146, 308380, P16871, P27987, P52333, and O43561. The "HPO combined" and "HPO term" columns list Human Phenotype Ontology terms like "Abnormalit" and "T lymphocit". A footer at the bottom left says "1-10 of 591 rows Show 10". A footer at the bottom right says "Previous 1 2 3 4 5 ... 60 Next".

Figure 11: **Integration of variant probabilities into the IEI genetics framework.** The interface summarises the condensed variant data, with pre-calculated summary statistics and dynamic links to external databases. This integration enables immediate access to detailed variant classifications and prior probabilities for each gene.

732 **3.9 Probability of observing AlphaMissense pathogenicity**

733 AlphaMissense provides pathogenicity scores for all possible amino acid substitutions;
 734 however, our results in **Figure 12** show that the most probable observations in pa-
 735 tients occur predominantly for benign or unknown variants. This finding places the
 736 likelihood of disease-associated substitutions into perspective and offers a data-driven
 737 foundation for future improvements in variant prediction. The values in **Figure 12**
 738 (A) can be directly compared to **Figure 1 (D)** to view the distribution of classifi-
 739 cations. A Kruskal-Wallis test was used to compare the observed disease probability

740 across clinical classification groups and no significant differences were detected. In
 741 general, most variants in patients are classified as benign or unknown, indicating
 742 limited discriminative power in the current classification, such that pathogenicity
 743 prediction does not infer observation prediction (**Figure S13**).

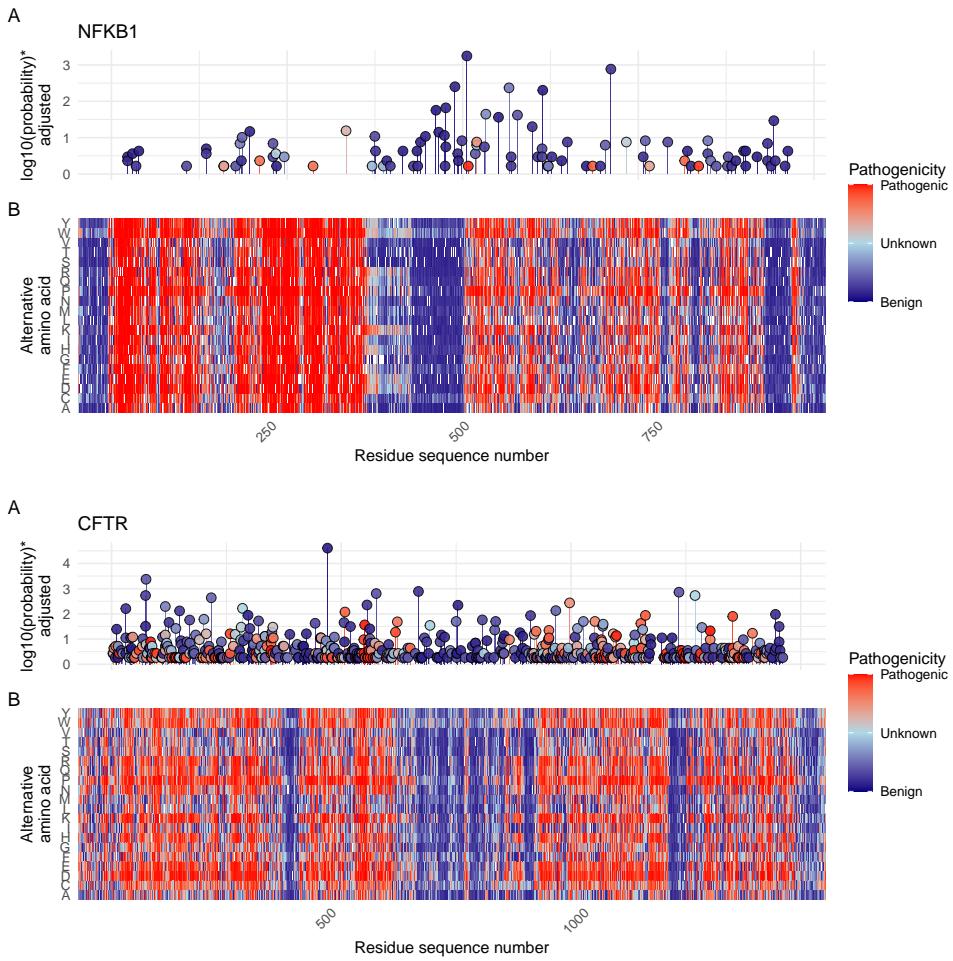


Figure 12: **(A) Probabilities of observing a patient with (B) AlphaMissense-derived pathogenicity scores.** Although AlphaMissense provides scores for every possible amino acid substitution, the most frequently observed variants in patients tend to be classified as benign or of unknown significance. This juxtaposition contextualises the likelihood of disease-associated substitutions and underlines prospects for refining predictive models. *Axis scaled for visibility near zero. Higher point indicates higher probability.

744 4 Discussion

745 Our study presents, to our knowledge, the first comprehensive framework for calculating
746 prior probabilities of observing disease-associated variants. By integrating large-
747 scale genomic annotations, including population allele frequencies from gnomAD (7),
748 variant classifications from ClinVar (13), and functional annotations from resources
749 such as dbNSFP, with classical Hardy-Weinberg-based calculations, we derived robust
750 estimates for 54,814 ClinVar variant classifications across 557 IEI genes implicated in
751 PID and monogenic inflammatory bowel disease (1; 2).

752 Our approach yielded two key results. First, our detailed, per-variant pre-calculated
753 results provide prior probabilities of observing disease-associated variants across all
754 MOI for any gene-disease combination. Second, the score positive total metric effec-
755 tively summarises the aggregate pathogenic burden across genes, serving as a robust
756 indicator of genetic constraint and highlighting those with elevated disease relevance.

Estimating disease risk in genetic studies is complicated by uncertainties in key parameters such as variant penetrance and the fraction of cases attributable to specific variants (6). In the simplest model, where a single, fully penetrant variant causes disease, the lifetime risk $P(D)$ is equivalent to the genotype frequency $P(G)$. For an allele with frequency p , this translates to:

$$\begin{aligned} \text{Recessive: } P(D) &= p^2, \\ \text{Dominant: } P(D) &= 2p(1 - p) \approx 2p. \end{aligned}$$

When penetrance is incomplete, defined as $P(D | G)$, the risk becomes:

$$P(D) = P(G) P(D | G).$$

In more realistic scenarios where multiple variants contribute to disease, $P(G | D)$ denotes the fraction of cases attributable to a given variant. This leads to:

$$P(D) = \frac{P(G) P(D | G)}{P(G | D)}.$$

757 Because both penetrance and $P(G | D)$ are often uncertain, solving this equation
758 systematically poses a major challenge.

759 Our framework addresses this challenge by combining variant classifications, pop-
760 ulation allele frequencies, and curated gene-disease associations. While imperfect on
761 an individual level, these sources exhibit predictable aggregate behaviour, supported
762 by James-Stein estimation principles (28). Curated gene-disease associations help
763 identify genes that explainable for most disease cases, allowing us to approximate
764 $P(G | D)$ close to one. In this way, we obtain robust estimates of $P(G)$ (the fre-
765 quency of disease-associated genotypes), even when exact values of penetrance and
766 case attribution remain uncertain.

This approach allows us to pre-calculate priors and summarise the overall pathogenic burden using our *score positive total* metric. By focusing on a subset \mathcal{V} of variants

that pass stringent filtering, where each $P(G_i | D)$ is the probability that a case of disease D is attributable to variant i , we assume that, in aggregate,

$$\sum_{i \in \mathcal{V}} P(G_i | D) \approx 1.$$

Even if the cumulative contribution is slightly less than one, the resultant risk estimates remain robust within the broad CIs typical of epidemiological studies. By incorporating these pre-calculated priors into a Bayesian framework, our method refines risk estimates and enhances clinical decision-making despite inherent uncertainties.

Our results focused on IEI, but the genome-wide approach accommodates the distinct MOI patterns of AD, AR, and XL disorders. Whereas AD and XL conditions require only a single pathogenic allele, AR disorders necessitate the consideration of both homozygous and compound heterozygous states. These classical HWE-based estimates provide an informative baseline for predicting variant occurrence and serve as robust priors for Bayesian models of variant and disease risk estimation. This is an approach that has been underutilised in clinical and statistical genetics. As such, our framework refines risk calculations by incorporating MOI complexities and enhances clinicians' understanding of expected variant occurrences, thereby improving diagnostic precision.

Moreover, our method complements existing statistical approaches for aggregating variant effects with methods like Sequence Kernel Association Test (SKAT) and Aggregated Cauchy Association Test (ACAT) (29–32)) and multi-omics integration techniques (33; 34), while remaining consistent with established variant interpretation guidelines from the American College of Medical Genetics and Genomics (ACMG) (35) and complementary frameworks (36; 37), as well as quality control protocols (38; 39). Standardised reporting for qualifying variant sets, such as ACMG Secondary Findings v3.2 (40), further contextualises the integration of these probabilities into clinical decision-making.

We acknowledge that our current framework is restricted to SNVs and does not incorporate numerous other complexities of genetic disease, such as structural variants, de novo variants, hypomorphic alleles, overdominance, variable penetrance, tissue-specific expression, the Wahlund effect, pleiotropy, and others (6). In certain applications, more refined estimates would benefit from including factors such as embryonic lethality, condition-specific penetrance, and age of onset (10). Our analysis also relies on simplifying assumptions of random mating, an effectively infinite population, and the absence of migration, novel mutations, or natural selection.

Future work will incorporate additional variant types and models to further refine these probability estimates. By continuously updating classical estimates with emerging data and prior knowledge, we aim to enhance the precision of genetic diagnostics and ultimately improve patient care.

The results presented in the *NFKB1* case study demonstrate how our Bayesian framework quantifies the final posterior probability that a damaging causal variant is present, by integrating both observed variant calls and unobserved, but potentially

805 pathogenic, sites into a single conclusion. This result applies across all disease-genes.
806 The model partitions posterior probability between a true positive (*p.Ser237Ter*) and
807 a likely causal splice-site variant that was not captured by sequencing. Despite the
808 latter being unobserved, the analysis attributes 48 % of total pathogenic probability
809 to this missing position—an insight that would be unavailable under conventional
810 approaches limited to called variants.

811 This framework enables calibrated probabilistic conclusions at two levels: per-
812 variant and per-gene. Rather than issuing a binary classification, the method produces
813 interpretable posterior intervals that reflect both sequencing completeness and prior
814 pathogenic evidence. For clinicians, this means that residual uncertainty—arising
815 from false negatives or incomplete genomic coverage—is made visible, quantifiable,
816 and actionable.

817 A key output is the ability to generate structured variant reports (e.g. Table ??) that
818 transparently communicate which alleles are supported, which are excluded, and
819 which remain unassessed but potentially disease-causing. These summaries facilitate
820 informed consent, prioritise follow-up sequencing (e.g. long-read or targeted assays),
821 and provide a reproducible, evidence-weighted foundation for diagnostic decision-
822 making.

823 More broadly, the same approach generalises to genome-wide applications. The
824 precomputed priors for over 50,000 variants across 557 immune-related genes now
825 enable quantitative confidence statements for any patient–phenotype combination.
826 Because the model explicitly incorporates allele frequency, clinical classification, in-
827 heritance mode, and sequence quality, it provides a scalable template for delivering
828 uncertainty-aware interpretation in clinical genomics.

829 5 Conclusion

830 Our work generates prior probabilities for observing any variant classification in IEI
831 genetic disease, providing a quantitative resource to enhance Bayesian variant inter-
832 pretation and clinical decision-making.

833 Acknowledgements

834 We acknowledge Genomics England for providing public access to the PanelApp data.
835 The use of data from Genomics England panelapp was licensed under the Apache
836 License 2.0. The use of data from UniProt was licensed under Creative Commons
837 Attribution 4.0 International (CC BY 4.0). ClinVar asks its users who distribute or
838 copy data to provide attribution to them as a data source in publications and websites
839 (13). dbNSFP version 4.4a is licensed under the Creative Commons Attribution-
840 NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0); while we cite
841 this dataset as used our research publication, it is not used for the final version which

842 instead used ClinVar and gnomAD directly. GnomAD is licensed under Creative
843 Commons Zero Public Domain Dedication (CC0 1.0 Universal). GnomAD request
844 that usages cites the gnomAD flagship paper (7) and any online resources that include
845 the data set provide a link to the browser, and note that tool includes data from the
846 gnomAD v4.1 release. AlphaMissense asks to cite Cheng et al. (12) for usage in
847 research, with data available from Cheng et al. (26).

848 Competing interest

849 We declare no competing interest.

850 References

- 851 [1] Stuart G. Tangye, Waleed Al-Herz, Aziz Bousfiha, Charlotte Cunningham-
852 Rundles, Jose Luis Franco, Steven M. Holland, Christoph Klein, Tomohiro Morio,
853 Eric Oksenhendler, Capucine Picard, Anne Puel, Jennifer Puck, Mikko R. J.
854 Seppänen, Raz Somech, Helen C. Su, Kathleen E. Sullivan, Troy R. Torger-
855 son, and Isabelle Meyts. Human Inborn Errors of Immunity: 2022 Update
856 on the Classification from the International Union of Immunological Societies
857 Expert Committee. *Journal of Clinical Immunology*, 42(7):1473–1507, October
858 2022. ISSN 0271-9142, 1573-2592. doi: 10.1007/s10875-022-01289-3. URL
859 <https://link.springer.com/10.1007/s10875-022-01289-3>.
- 860 [2] Dylan Lawless. PanelAppRex aggregates disease gene panels and facilitates
861 sophisticated search. March 2025. doi: 10.1101/2025.03.20.25324319. URL
862 <http://medrxiv.org/lookup/doi/10.1101/2025.03.20.25324319>.
- 863 [3] Antonio Rueda Martin, Eleanor Williams, Rebecca E. Foulger, Sarah Leigh,
864 Louise C. Daugherty, Olivia Niblock, Ivone U. S. Leong, Katherine R. Smith,
865 Oleg Gerasimenko, Eik Haraldsdottir, Ellen Thomas, Richard H. Scott, Emma
866 Baple, Arianna Tucci, Helen Brittain, Anna De Burca, Kristina Ibañez, Dalia
867 Kasperaviciute, Damian Smedley, Mark Caulfield, Augusto Rendon, and Ellen M.
868 McDonagh. PanelApp crowdsources expert knowledge to establish consensus
869 diagnostic gene panels. *Nature Genetics*, 51(11):1560–1565, November 2019.
870 ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-019-0528-2. URL <https://www.nature.com/articles/s41588-019-0528-2>.
- 872 [4] Oliver Mayo. A Century of Hardy–Weinberg Equilibrium. *Twin Research
873 and Human Genetics*, 11(3):249–256, June 2008. ISSN 1832-4274, 1839-
874 2628. doi: 10.1375/twin.11.3.249. URL https://www.cambridge.org/core/product/identifier/S1832427400009051/type/journal_article.
- 876 [5] Nikita Abramovs, Andrew Brass, and May Tassabehji. Hardy–Weinberg Equi-
877 librium in the Large Scale Genomic Sequencing Era. *Frontiers in Genetics*,

- 878 11:210, March 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.00210. URL
879 <https://www.frontiersin.org/article/10.3389/fgene.2020.00210/full>.
- 880 [6] Johannes Zschocke, Peter H. Byers, and Andrew O. M. Wilkie. Mendelian
881 inheritance revisited: dominance and recessiveness in medical genetics. *Nature
882 Reviews Genetics*, 24(7):442–463, July 2023. ISSN 1471-0056, 1471-0064.
883 doi: 10.1038/s41576-023-00574-0. URL <https://www.nature.com/articles/s41576-023-00574-0>.
- 885 [7] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings,
886 Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea
887 Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified
888 from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.
- 889 [8] Sarah L. Bick, Aparna Nathan, Hannah Park, Robert C. Green, Monica H. Wo-
890 jcik, and Nina B. Gold. Estimating the sensitivity of genomic newborn screen-
891 ing for treatable inherited metabolic disorders. *Genetics in Medicine*, 27(1):
892 101284, January 2025. ISSN 10983600. doi: 10.1016/j.gim.2024.101284. URL
893 <https://linkinghub.elsevier.com/retrieve/pii/S1098360024002181>.
- 894 [9] Benjamin D. Evans, Piotr Słowiński, Andrew T. Hattersley, Samuel E. Jones,
895 Seth Sharp, Robert A. Kimmitt, Michael N. Weedon, Richard A. Oram,
896 Krasimira Tsaneva-Atanasova, and Nicholas J. Thomas. Estimating disease
897 prevalence in large datasets using genetic risk scores. *Nature Communications*,
898 12(1):6441, November 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-26501-7.
899 URL <https://www.nature.com/articles/s41467-021-26501-7>.
- 900 [10] William B. Hannah, Mitchell L. Drumm, Keith Nykamp, Tiziano Prampano,
901 Robert D. Steiner, and Steven J. Schrodi. Using genomic databases to de-
902 termine the frequency and population-based heterogeneity of autosomal reces-
903 sive conditions. *Genetics in Medicine Open*, 2:101881, 2024. ISSN 29497744.
904 doi: 10.1016/j.gimo.2024.101881. URL <https://linkinghub.elsevier.com/retrieve/pii/S2949774424010276>.
- 906 [11] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov,
907 Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek,
908 Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J.
909 Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishabh
910 Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy,
911 Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer,
912 Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray
913 Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate pro-
914 tein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August
915 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03819-2. URL
916 <https://www.nature.com/articles/s41586-021-03819-2>.

- 917 [12] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Tay-
918 lor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias
919 Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hass-
920 abis, Pushmeet Kohli, and Žiga Avsec. Accurate proteome-wide missense vari-
921 ant effect prediction with AlphaMissense. *Science*, 381(6664):eadg7492, Septem-
922 ber 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adg7492. URL
923 <https://www.science.org/doi/10.1126/science.adg7492>.
- 924 [13] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao,
925 Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee
926 Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adri-
927 ana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou,
928 J Bradley Holmes, Brandi L Kattman, and Donna R Maglott. ClinVar: im-
929 proving access to variant interpretations and supporting evidence. *Nucleic Acids
930 Research*, 46(D1):D1062–D1067, January 2018. ISSN 0305-1048, 1362-4962. doi:
931 10.1093/nar/gkx1153. URL [http://academic.oup.com/nar/article/46/D1/
932 D1062/4641904](http://academic.oup.com/nar/article/46/D1/D1062/4641904).
- 933 [14] The UniProt Consortium, Alex Bateman, Maria-Jesus Martin, Sandra Orchard,
934 Michele Magrane, Aduragbemi Adesina, Shadab Ahmad, Bowler-Barnett, and
935 Others. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic
936 Acids Research*, 53(D1):D609–D617, January 2025. ISSN 0305-1048, 1362-4962.
937 doi: 10.1093/nar/gkae1010. URL [https://academic.oup.com/nar/article/
938 53/D1/D609/7902999](https://academic.oup.com/nar/article/53/D1/D609/7902999).
- 939 [15] Xiaoming Liu, Chang Li, Chengcheng Mou, Yibo Dong, and Yicheng Tu.
940 dbNSFP v4: a comprehensive database of transcript-specific functional pre-
941 dictions and annotations for human nonsynonymous and splice-site SNVs.
942 *Genome Medicine*, 12(1):103, December 2020. ISSN 1756-994X. doi: 10.
943 1186/s13073-020-00803-9. URL [https://genomemedicine.biomedcentral.
944 com/articles/10.1186/s13073-020-00803-9](https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00803-9).
- 945 [16] Damian Szklarczyk, Katerina Nastou, Mikaela Koutrouli, Rebecca Kirsch, Far-
946 rokh Mehryary, Radja Hachilif, Dewei Hu, Matteo E Peluso, Qingyao Huang,
947 Tao Fang, et al. The string database in 2025: protein networks with directional-
948 ity of regulation. *Nucleic Acids Research*, 53(D1):D730–D737, 2025.
- 949 [17] Paul Tijnenburg, Hana Lango Allen, Siobhan O. Burns, Daniel Greene,
950 Machiel H. Jansen, and Others. Loss-of-function nuclear factor B subunit
951 1 (NFKB1) variants are the most common monogenic cause of common vari-
952 able immunodeficiency in Europeans. *Journal of Allergy and Clinical Im-
953 munology*, 142(4):1285–1296, October 2018. ISSN 00916749. doi: 10.1016/
954 j.jaci.2018.01.039. URL [https://linkinghub.elsevier.com/retrieve/pii/
955 S0091674918302860](https://linkinghub.elsevier.com/retrieve/pii/S0091674918302860).
- 956 [18] WHO Scientific Group et al. Primary immunodeficiency diseases: report of a
957 who scientific group. *Clin. Exp. Immunol.*, 109(1):1–28, 1997.

- 958 [19] Charlotte Cunningham-Rundles and Carol Bodian. Common variable immunodeficiency: clinical and immunological features of 248 patients. *Clinical immunology*, 92(1):34–48, 1999.
- 959
- 960
- 961 [20] Eric Oksenhendler, Laurence Gérard, Claire Fieschi, Marion Malphettes, Gael
962 Mouillot, Roland Jaussaud, Jean-François Viallard, Martine Gardembas, Lionel
963 Galicier, Nicolas Schleinitz, et al. Infections in 252 patients with common variable
964 immunodeficiency. *Clinical Infectious Diseases*, 46(10):1547–1554, 2008.
- 965
- 966 [21] Y Naito, F Adams, S Charman, J Duckers, G Davies, and S Clarke. Uk cystic
967 fibrosis registry 2023 annual data report. *London: Cystic Fibrosis Trust*, 2023.
- 968
- 969 [22] Carlo Castellani, CFTR2 team, et al. Cftr2: how will it help care? *Paediatric
970 respiratory reviews*, 14:2–5, 2013.
- 971
- 972 [23] Hartmut Grasemann and Felix Ratjen. Cystic fibrosis. *New England Journal
973 of Medicine*, 389(18):1693–1707, 2023. doi: 10.1056/NEJMra2216474. URL
974 <https://www.nejm.org/doi/full/10.1056/NEJMra2216474>.
- 975
- 976 [24] Kyoko Watanabe, Erdogan Taskesen, Arjen Van Bochoven, and Danielle
977 Posthuma. Functional mapping and annotation of genetic associations with
978 FUMA. *Nature Communications*, 8(1):1826, November 2017. ISSN 2041-1723.
979 doi: 10.1038/s41467-017-01261-5. URL <https://www.nature.com/articles/s41467-017-01261-5>.
- 980
- 981
- 982 [25] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir,
983 Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (MSigDB)
984 3.0. *Bioinformatics*, 27(12):1739–1740, June 2011. ISSN 1367-4811, 1367–
985 4803. doi: 10.1093/bioinformatics/btr260. URL <https://academic.oup.com/bioinformatics/article/27/12/1739/257711>.
- 986
- 987 [26] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Tay-
988 lor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias
989 Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hass-
990 abis, Pushmeet Kohli, and Žiga Avsec. Predictions for alphanonsense, September
991 2023. URL <https://doi.org/10.5281/zenodo.8208688>.
- 992
- 993 [27] Dylan Lawless. Variant risk estimate probabilities for ie genes. March 2025. doi:
994 10.5281/zenodo.15111584. URL <https://doi.org/10.5281/zenodo.15111584>.
- 995
- 996 [28] Bradley Efron and Carl Morris. Stein’s Estimation Rule and Its Competitors—
An Empirical Bayes Approach. *Journal of the American Statistical Association*,
68(341):117, March 1973. ISSN 01621459. doi: 10.2307/2284155. URL <https://www.jstor.org/stable/2284155?origin=crossref>.
- 997
- 998 [29] Yaowu Liu, Sixing Chen, Zilin Li, Alanna C Morrison, Eric Boerwinkle, and
999 Xihong Lin. Acat: a fast and powerful p value combination method for rare-
999 variant analysis in sequencing studies. *The American Journal of Human Genetics*,
104(3):410–421, 2019.

- 997 [30] Xihao Li, Zilin Li, Hufeng Zhou, Sheila M Gaynor, Yaowu Liu, Han Chen, Ryan
998 Sun, Rounak Dey, Donna K Arnett, Stella Aslibekyan, et al. Dynamic incorpora-
999 tion of multiple *in silico* functional annotations empowers rare variant association
1000 analysis of large whole-genome sequencing studies at scale. *Nature genetics*, 52
1001 (9):969–983, 2020.
- 1002 [31] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xi-
1003 hong Lin. Rare-variant association testing for sequencing data with the sequence
1004 kernel association test. *The American Journal of Human Genetics*, 89(1):82–93,
1005 2011.
- 1006 [32] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J
1007 Rieder, Deborah A Nickerson, David C Christiani, Mark M Wurfel, and Xihong
1008 Lin. Optimal unified approach for rare-variant association testing with applica-
1009 tion to small-sample case-control whole-exome sequencing studies. *The American
1010 Journal of Human Genetics*, 91(2):224–237, 2012.
- 1011 [33] Augustine Kong, Gudmar Thorleifsson, Michael L Frigge, Bjarni J Vilhjalmsson,
1012 Alexander I Young, Thorgeir E Thorgeirsson, Stefania Benonisdottir, Asmundur
1013 Oddsson, Bjarni V Halldorsson, Gisli Masson, et al. The nature of nurture:
1014 Effects of parental genotypes. *Science*, 359(6374):424–428, 2018.
- 1015 [34] Laurence J Howe, Michel G Nivard, Tim T Morris, Ailin F Hansen, Humaira
1016 Rasheed, Yoonsu Cho, Geetha Chittoor, Penelope A Lind, Teemu Palviainen,
1017 Matthijs D van der Zee, et al. Within-sibship gwas improve estimates of direct
1018 genetic effects. *BioRxiv*, pages 2021–03, 2021.
- 1019 [35] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-
1020 Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al.
1021 Standards and guidelines for the interpretation of sequence variants: a joint
1022 consensus recommendation of the american college of medical genetics and ge-
1023 nomics and the association for molecular pathology. *Genetics in medicine*, 17
1024 (5):405–423, 2015.
- 1025 [36] Sean V Tavtigian, Steven M Harrison, Kenneth M Boucher, and Leslie G
1026 Biesecker. Fitting a naturally scaled point system to the acmg/amp variant
1027 classification guidelines. *Human mutation*, 41(10):1734–1737, 2020.
- 1028 [37] Quan Li and Kai Wang. Intervar: clinical interpretation of genetic variants by
1029 the 2015 acmg-amp guidelines. *The American Journal of Human Genetics*, 100
1030 (2):267–280, 2017.
- 1031 [38] Brent S Pedersen, Joe M Brown, Harriet Dashnow, Amelia D Wallace, Matt
1032 Velinder, Martin Tristani-Firouzi, Joshua D Schiffman, Tatiana Tvrzik, Rong
1033 Mao, D Hunter Best, et al. Effective variant filtering and expected candidate
1034 variant yield in studies of rare human disease. *NPJ Genomic Medicine*, 6(1):1–8,
1035 2021.

- 1036 [39] Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon,
1037 Andrew P Morris, and Krina T Zondervan. Data quality control in genetic
1038 case-control association studies. *Nature protocols*, 5(9):1564–1573, 2010. URL
1039 <https://doi.org/10.1038/nprot.2010.116>.
- 1040 [40] David T Miller, Kristy Lee, Noura S Abul-Husn, Laura M Amendola, Kyle Broth-
1041 ers, Wendy K Chung, Michael H Gollob, Adam S Gordon, Steven M Harrison,
1042 Ray E Hershberger, et al. Acmg sf v3. 2 list for reporting of secondary findings
1043 in clinical exome and genome sequencing: a policy statement of the american
1044 college of medical genetics and genomics (acmg). *Genetics in Medicine*, 25(8):
1045 100866, 2023.

¹⁰⁴⁶ **6 Supplemental**

¹⁰⁴⁷ **6.1 Integrating observed true positives and unobserved false**
¹⁰⁴⁸ **negatives into a single, actionable conclusion**

Table S1: Result of clinical genetics diagnosis scenario 1. The most strongly supported observed variant was p.Ser237Ter (posterior: 0.594). The strongest unsequenced variant was p.Thr567Ile (posterior: 0). The total probability of a causal diagnosis given the available evidence was 1 (95% CI: 1–1).

Variant	Flag	Class	Evidence Score	Occurrence Prob	Adj Occ Prob	Alpha	Beta	Lower	Median	Upper	Posterior Share	Prob Causal
p.Ser237Ter	present	causal	5	0.000	0	6	371	0.004	0.142	0.803	0.594	0.594
p.Thr567Ile	missing	other	-5	0.002	0	1	363	NA	NA	NA	0.000	0.000
p.Arg231His	present	other	0	0.000	0	1	361	0.004	0.142	0.803	0.000	0.000
p.Gly650Arg	present	other	0	0.000	0	1	379	0.004	0.142	0.803	0.000	0.000
p.Val236Ile	missing	other	0	0.000	0	1	351	NA	NA	NA	0.000	0.000
Total	NA	NA	NA	NA	NA	NA	NA	1.000	1.000	1.000	NA	1.000

Table S2: Result of clinical genetics diagnosis scenario 3. No observed variants were detected in this scenario. The strongest unsequenced variant was p.Cys243Arg (posterior: 0.347). The total probability of a causal diagnosis given the available evidence was 0 (95% CI: 0–0).

Variant	Flag	Class	Evidence Score	Occurrence Prob	Adj Occ Prob	Alpha	Beta	Lower	Median	Upper	Posterior Share	Prob Causal
p.Cys243Arg	missing	causal	5.0	0.000	0.000	6	371	NA	NA	NA	0.347	0.347
p.Tyr246Ter	missing	causal	4.0	0.000	0.000	5	365	NA	NA	NA	0.296	0.296
p.Lys304Glu	missing	other	-5.0	0.000	0.000	1	367	NA	NA	NA	0.000	0.000
p.Ile207Leu	missing	other	-4.5	0.000	0.000	1	359	NA	NA	NA	0.000	0.000
p.His646Pro	missing	other	0.0	0.002	0.001	1	371	NA	NA	NA	0.000	0.000
p.Arg280Trp	missing	other	-4.0	0.000	0.000	1	353	NA	NA	NA	0.000	0.000
p.Thr635Ile	missing	other	0.0	0.000	0.000	1	349	NA	NA	NA	0.000	0.000
p.Arg162Trp	missing	other	0.0	0.000	0.000	1	371	NA	NA	NA	0.000	0.000
Total	NA	NA	NA	NA	NA	NA	NA	0	0	0	NA	0.000

¹⁰⁴⁹ **6.2 Validation studies**

Gene: *NFKB1*

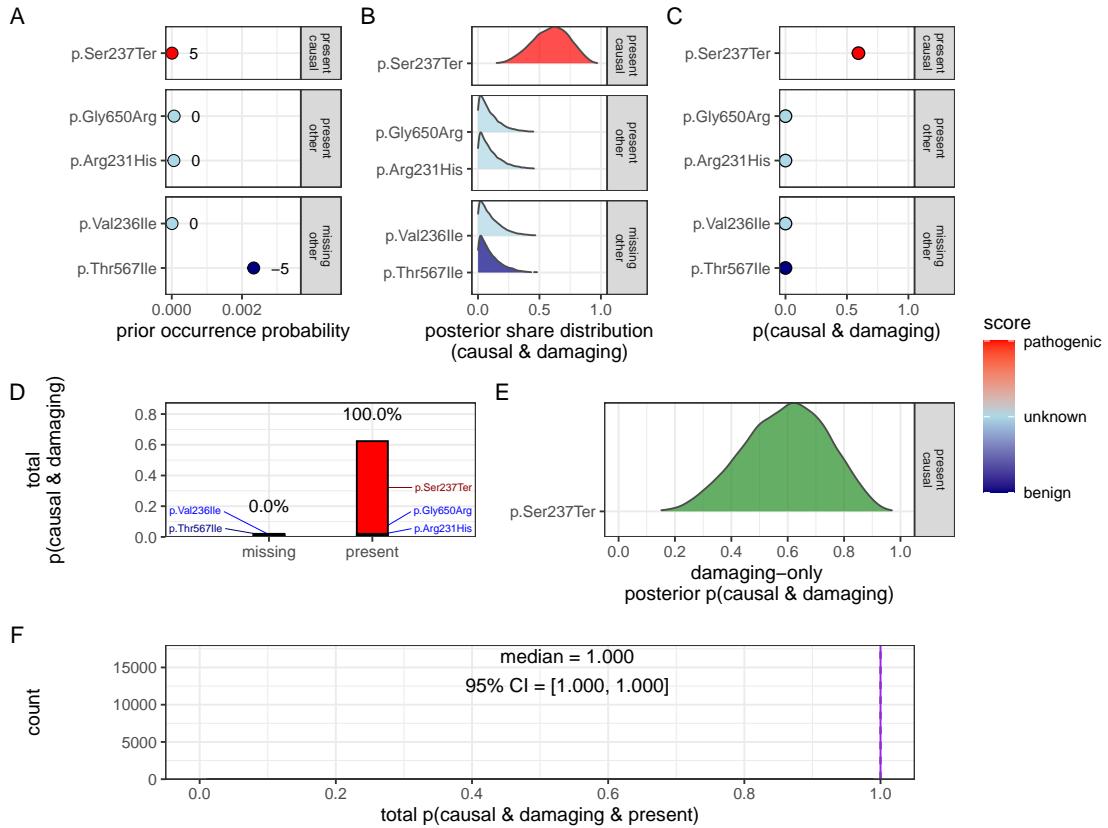


Figure S1: Quantification of present (TP) and no missing (FN) causal genetic variants for disease in *NFKB1* (scenario 1). Only one known pathogenic variant, p.Ser237Ter, was observed and all previously reported pathogenic positions were successfully sequenced and confirmed as reference (true negatives). Panels (A–F) follow the same structure as scenario 2 described in **Figure 2**, culminating in a gene-level posterior probability of 1 (95 % CrI: 0.99–1.00), with full support assigned to the observed allele given the available evidence. Pathogenicity scores (-5 to +5) are annotated.

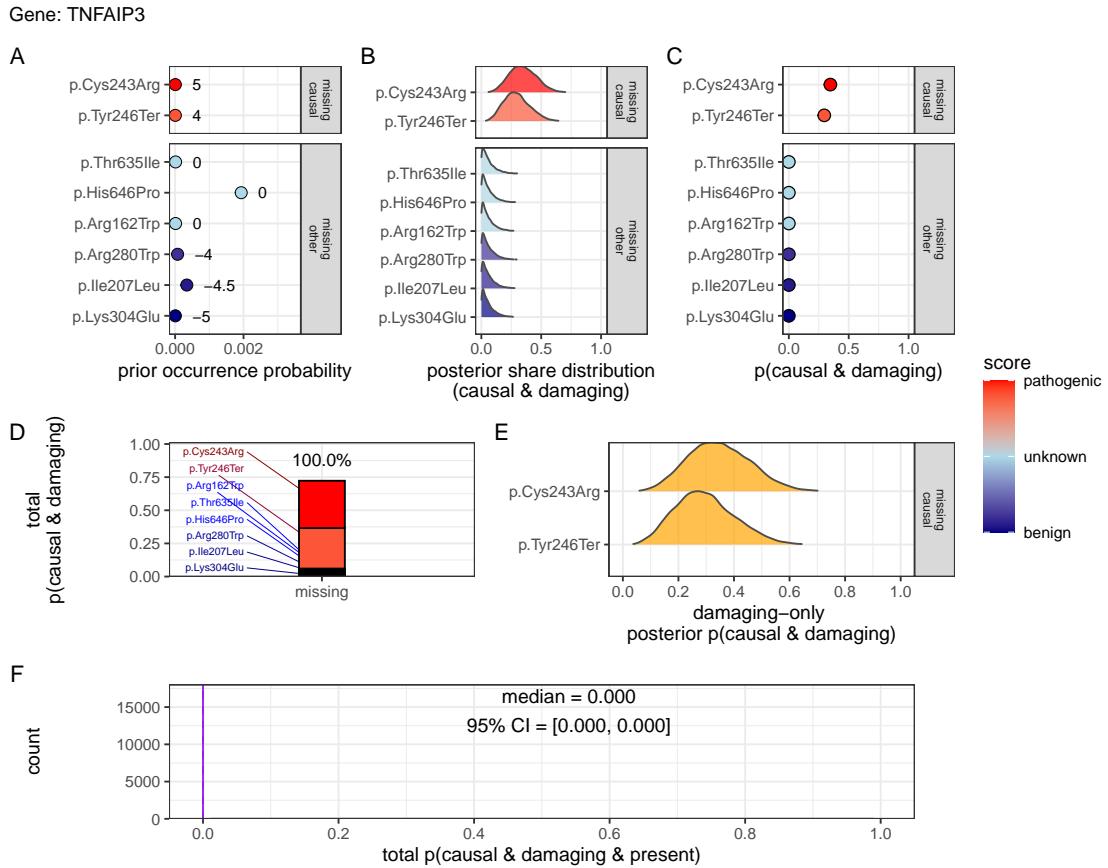


Figure S2: **Quantification of no present (TP) in *NFKB1* and only missing (FN) causal genetic variants for disease in *TNFAIP3* (scenario 3).** No known causal variants were observed in *NFKB1*, but one representative unsequenced allele was selected from each distinct ClinVar classification and treated as a potential false negative. Panels (A–F) follow the same structure as scenario 2 described in **Figure 2**. The posterior reflects uncertainty across multiple plausible but unobserved variants, resulting in low CrI (0–0) and 100% missing overall attribution in contrast to scenarios where known pathogenic variants were observed. For this patient, we have no evidence of a causal variant since the only top candidates are not yet accounted for. Pathogenicity scores (-5 to +5) are annotated in (A).

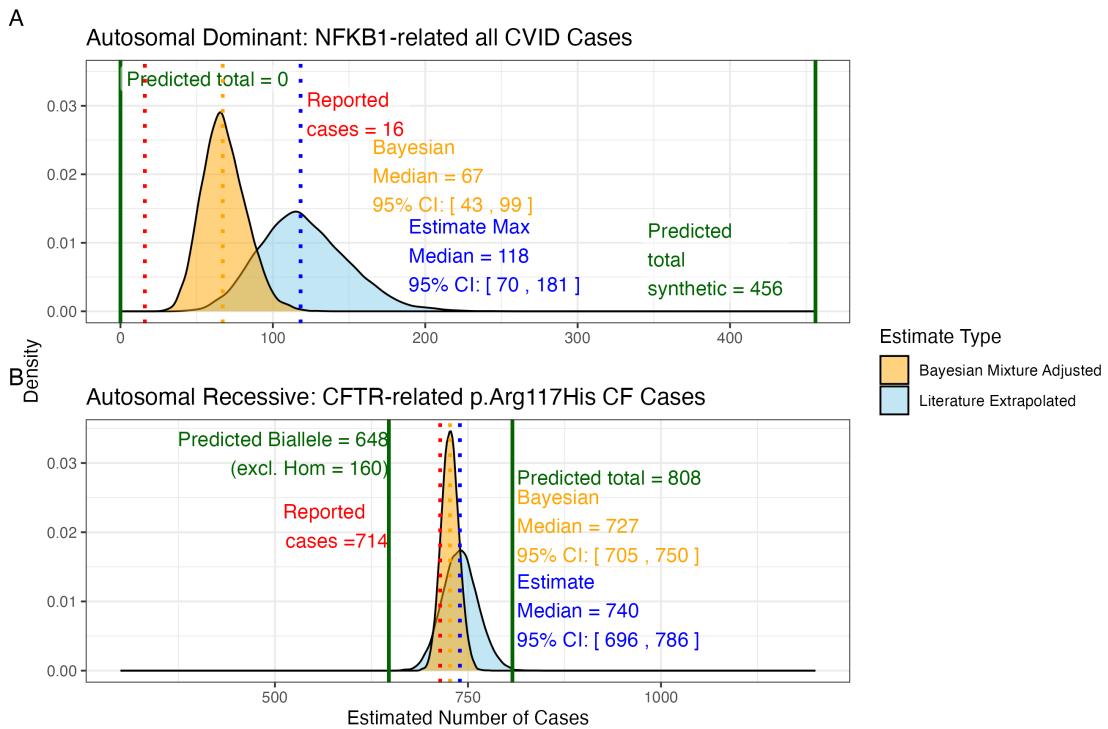


Figure S3: Prior probabilities compared to validation disease cohort metrics.

(A) Density distributions for the number of *NFKB1*-related CVID cases in the UK. Our model (green) predicted 456 cases, which falls between the observed cohort count (red) of 390 and the upper extrapolated values. The blue curve represents maximum count of 1280, and the orange curve shows the Bayesian-adjusted mixture estimate of 835. (B) Density distributions for *CFTR*-related p.Arg117His CF cases. Our model (green) predicted 648 biallelic cases and 808 total cases. The nationally reported case count (red) was 714. The blue curve represents maximum extrapolated count of 740, and the orange curve shows the Bayesian-adjusted mixture estimate of 727. We observed close agreement among the reported disease cases and our integrated probability estimation framework.

Condition: population size 69433632, phenotype PID-related, genes CFTR and NFKB1.

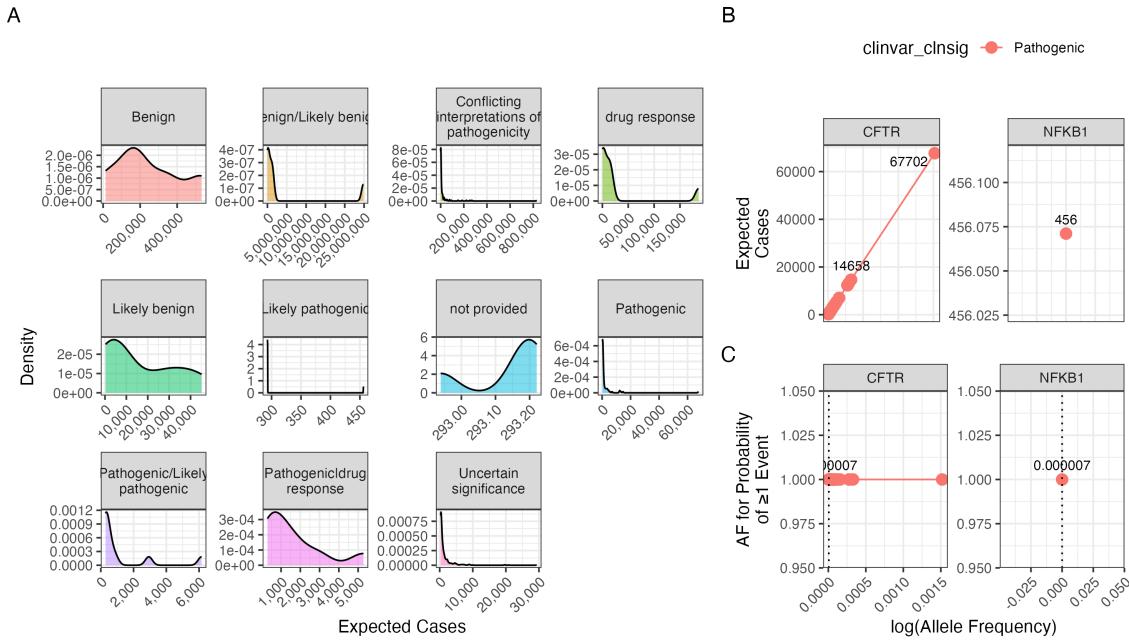


Figure S4: Interpretation of probability of observing a variant classification.
The result from the chosen validation genes *CFTR* and *NFKB1* are shown. Case counts are dependant on the population size and phenotype. (A) The density plots of expected observations by ClinVar clinical significance. We then highlight the values for pathogenic variants specifically showing; (B) the allele frequency versus expected cases in this population size and (C) the probability of observing at least one event in this population size.

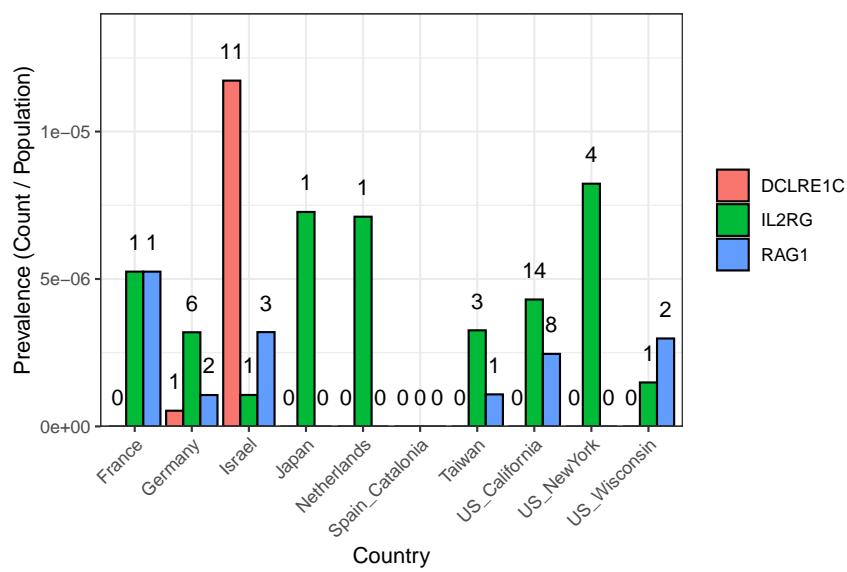


Figure S5: SCID-specific gene comparison across regions. The bar plot shows the prevalence of SCID-related cases (count divided by population) for each gene and country (or region), with numbers printed above the bars representing the actual counts in the original cohort (ranging from 0 to 11 per region and gene).

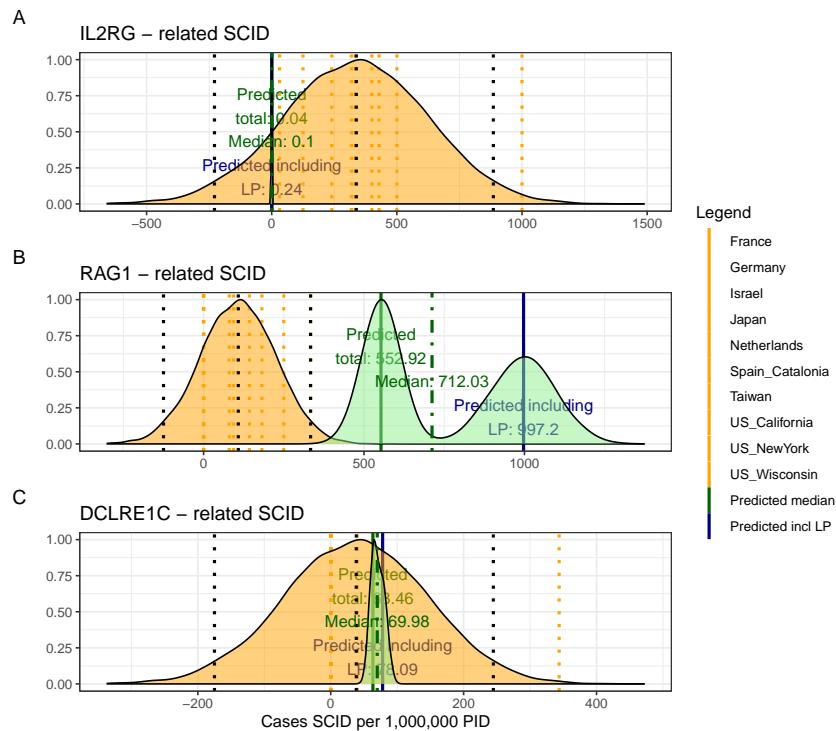


Figure S6: Combined SCID-specific Predictions and Observed Rates per 1,000,000 PID. The figure presents density distributions for the predicted SCID case counts (per 1,000,000 PID) for three genes: *IL2RG*, *RAG1*, and *DCLRE1C*. Country-specific rates (displayed as dotted vertical lines) are overlaid with the overall predicted distributions for pathogenic and likely pathogenic variants (solid lines with annotated medians). For *IL2RG*, the low predicted value is consistent with the high deleteriousness of loss-of-function variants in this X-linked gene, while *RAG1* exhibits considerably higher predicted counts, reflecting its lower penetrance in an autosomal recessive context.

1050 **6.3 Hierarchical Clustering of Enrichment Scores for Major**
 1051 **Disease Categories**

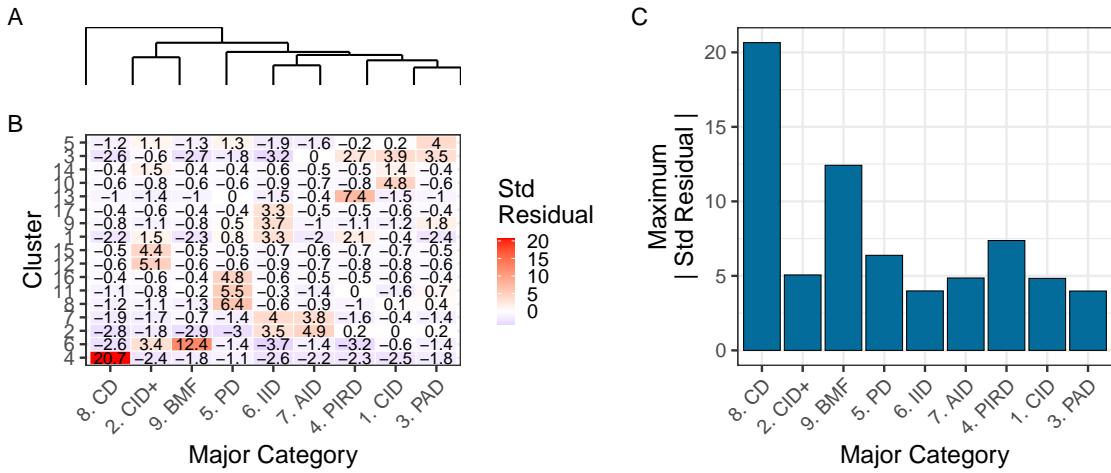


Figure S7: **Hierarchical clustering of enrichment scores.** The heatmap displays standardised residuals for major disease categories (x-axis) across network clusters (y-axis). A dendrogram groups similar disease categories, and the bar plot shows the maximum absolute residual per category. (8) CD and (9)BMF show the highest values, indicating significant enrichment or depletion ($\text{residuals} > |2|$). Definitions in **Box 2.1**.

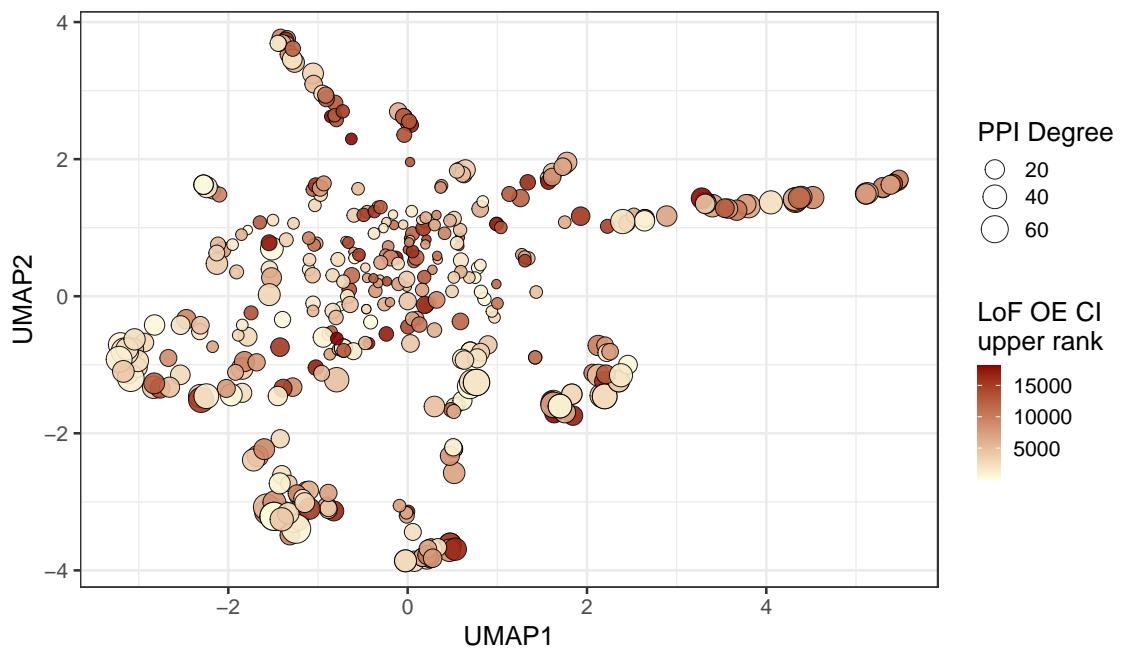


Figure S8: **Analysis of PPI degree versus LOEUF upper rank with UMAP embedding of the PPI network.** The relationship between PPI degree (size) and LOEUF upper rank (color) across gene clusters. No clear patterns are evident.

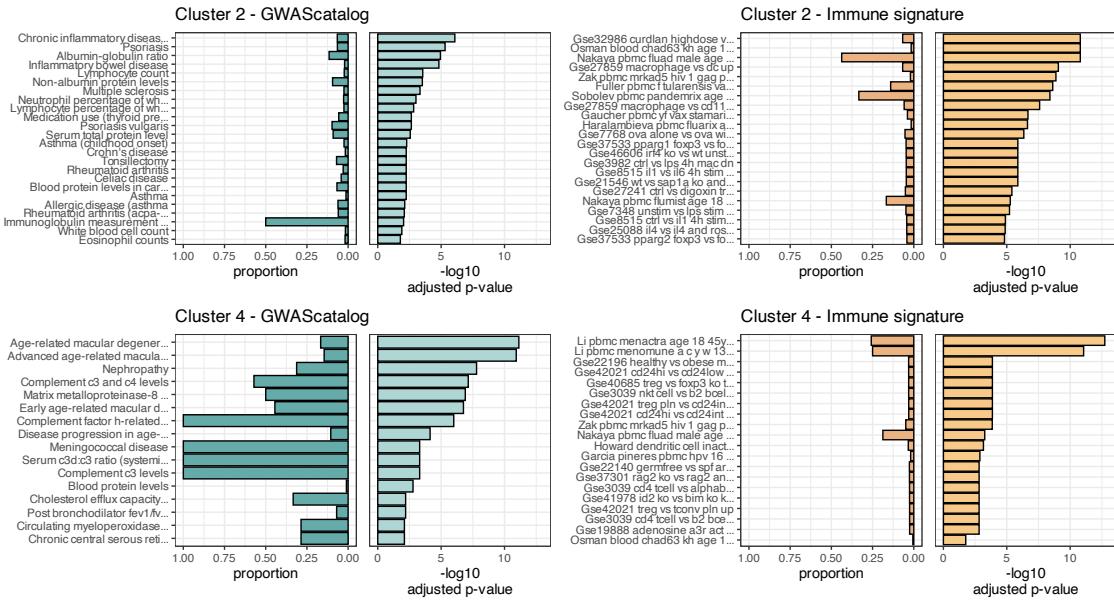


Figure S9: Composite Enrichment Profiles for IEI Gene Sets. We selected the top two enriched clusters (as per [Figure 5](#)) and performed functional enrichment analysis derived from known disease associations. For each gene set, the left panel displays the proportion of input genes overlapping with a curated gene set, and the right panel shows the $-\log_{10}$ adjusted p-value from hypergeometric testing. These profiles, stratified by cluster (Cluster 2 and Cluster 4) and by gene set category (GWAScatalog and Immunologic Signatures), highlight distinct enrichment patterns that reflect differential pathogenic variant loads in the IEI gene panels.

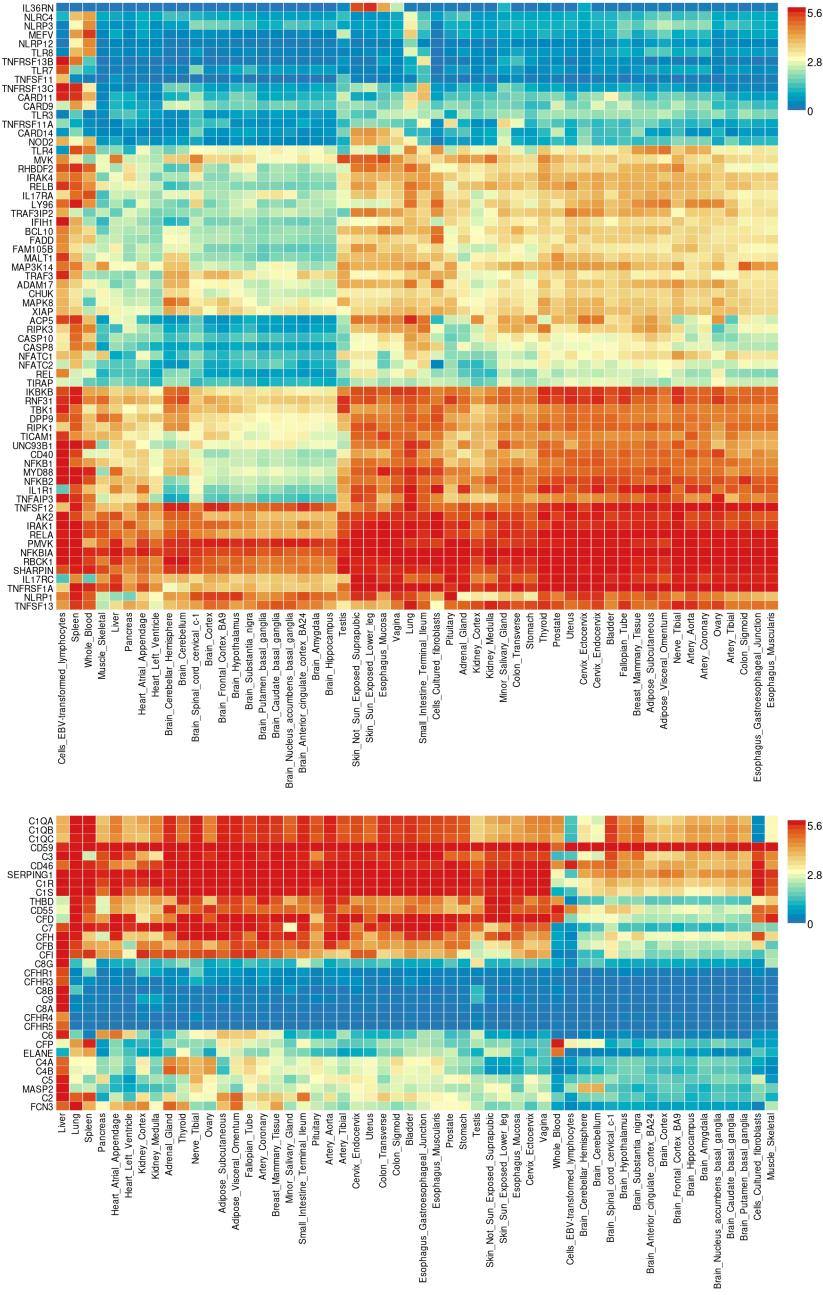


Figure S10: **Gene Expression Heatmaps for IEI Genes.** GTEx v8 data from 54 tissue types display the average expression per tissue label (log₂ transformed) for the IEI gene panels. Top: Cluster 2; Bottom: Cluster 4.

6.4 Interpretation of ClinVar Variant Observations

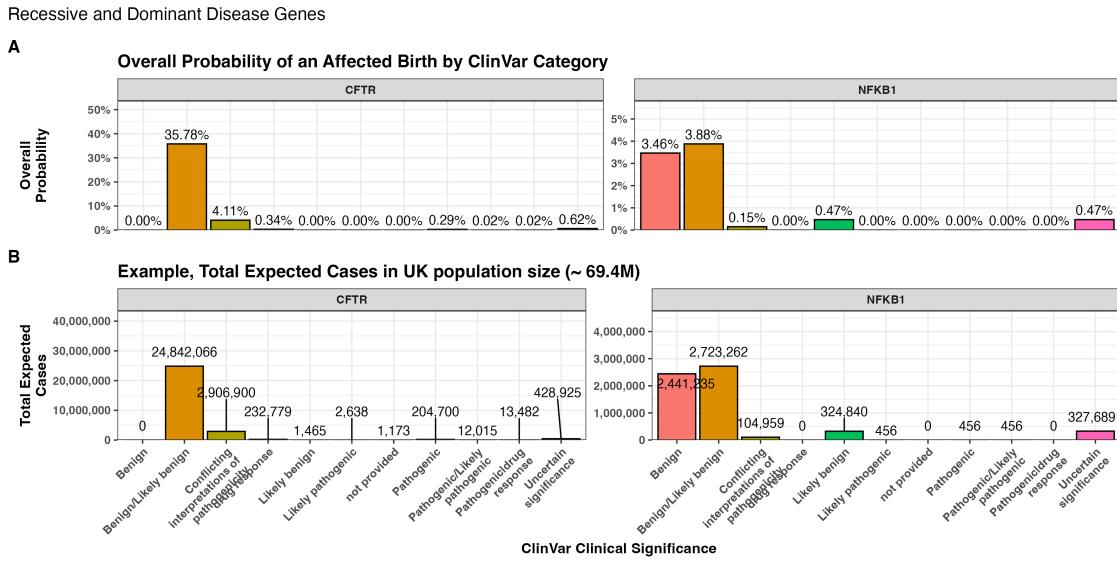


Figure S11: Combined bar charts summarizing the genome-wide analysis of ClinVar clinical significance for the PID gene panel. Panel (A) shows the overall probability of an affected birth by variant classification, and (B) displays the total expected number of cases per classification, both stratified by gene. These integrated results illustrate the variability in variant observations across genes and underpin our validation of the probability estimation framework.

6.5 Novel PID classifications

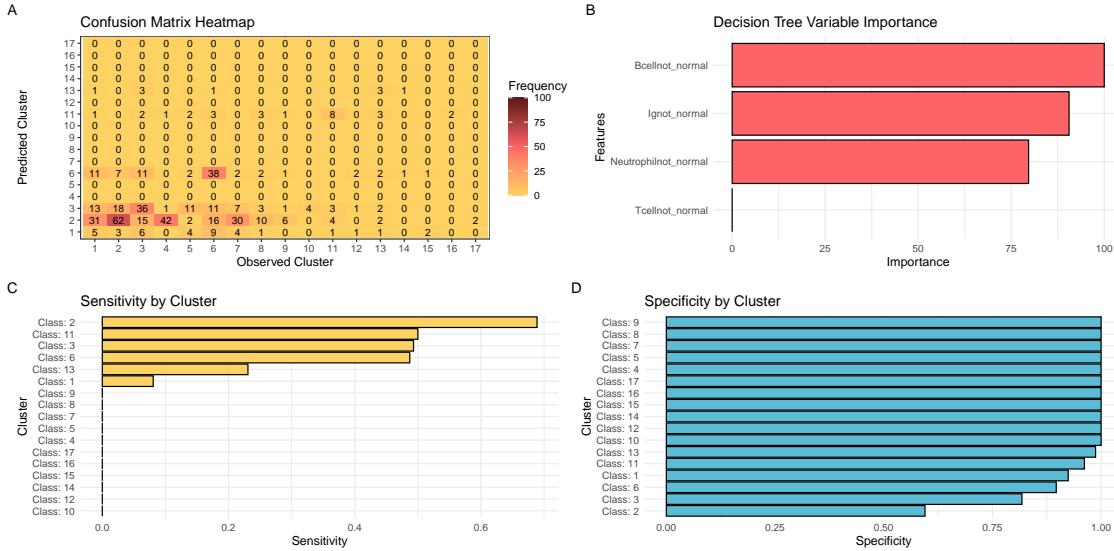


Figure S12: Model performance for fine-tuned PID classification. (A) Confusion matrix heatmap comparing observed and predicted PPI clusters. (B) Variable importance plot ranking immunophenotypic features contributing to the classifier. (C) Per-class sensitivity and (D) per-class specificity bar plots. These panels collectively demonstrate the performance of the decision tree classifier in stratifying PID genes based on immunophenotypic and PPI features.

1054 6.6 Probability of observing AlphaMissense pathogenicity

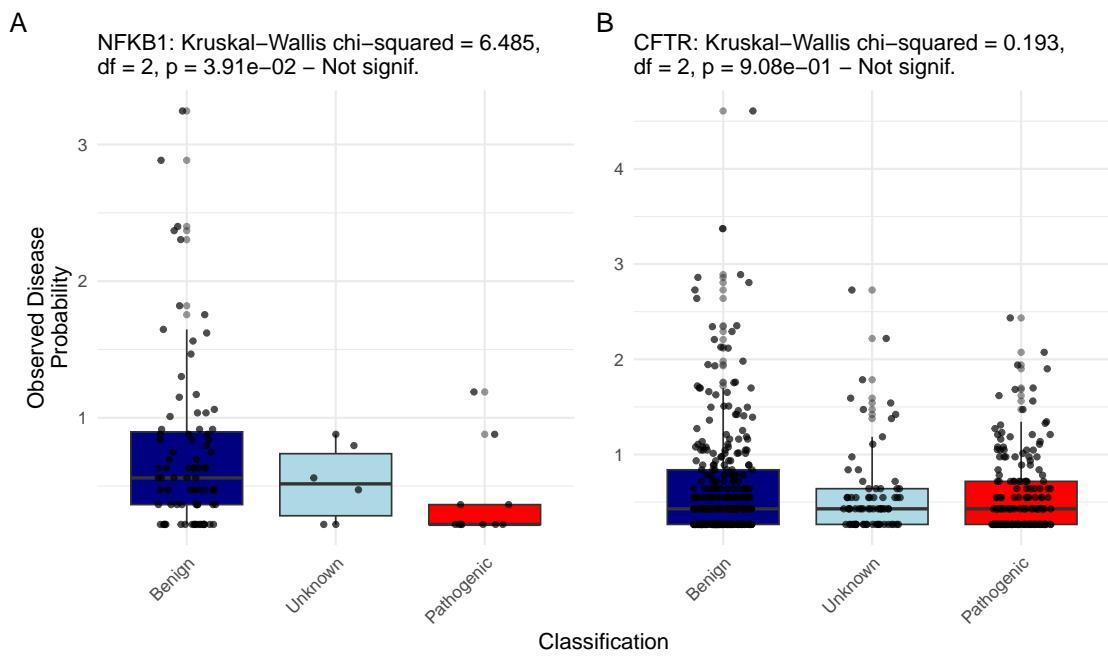


Figure S13: Observed Disease Probability by Clinical Classification with AlphaMissense. The figure displays the Kruskal-Wallis test results for NFKB1 and CFTR, showing no significant differences.