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

Dylan Lawless*[1]

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

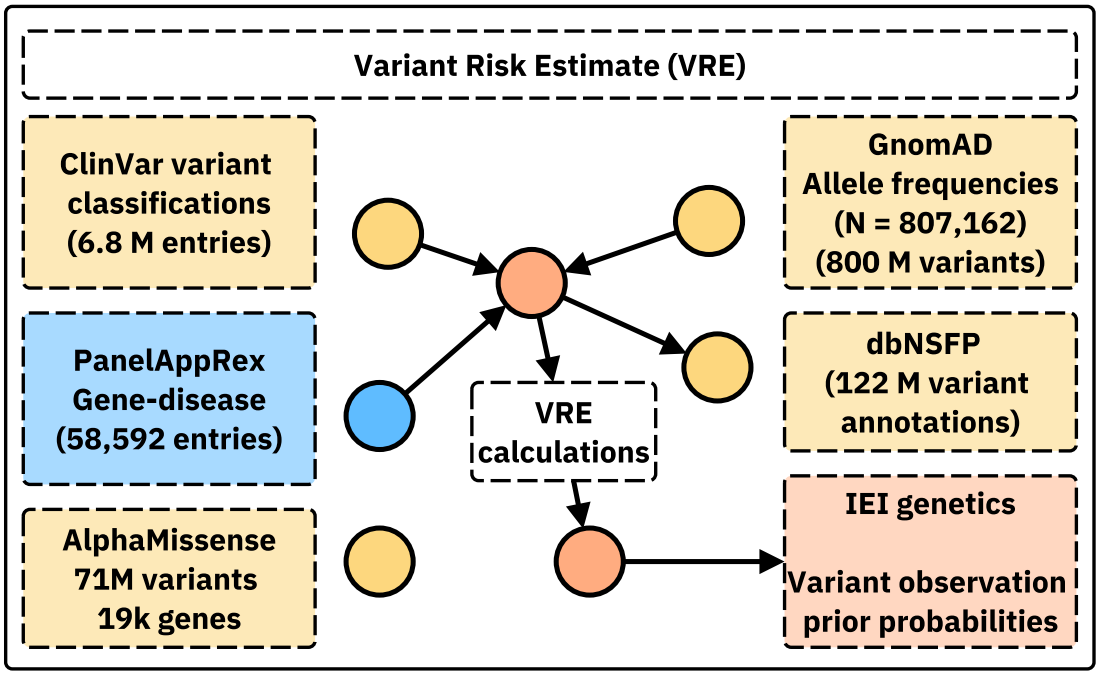
April 17, 2025

Abstract

We present a novel framework for quantifying the prior probability of observing disease-associated variants in any gene for a given phenotype. By integrating large-scale genomic annotations, including population allele frequencies and ClinVar variant classifications, with Hardy-Weinberg-based calculations, our method estimates per-variant observation probabilities under autosomal dominant (AD), autosomal recessive (AR), and X-linked modes of inheritance. Applied to 557 genes implicated in primary immunodeficiency and inflammatory disease, our approach generated 54,814 variant probabilities. First, these detailed, pre-calculated results provide robust priors for any gene-disease combination. Second, a score positive total metric summarises the aggregate pathogenic burden, serving as an indicator of the likelihood of observing a patient with the disease and reflecting genetic constraint. Validation in *NFKB1* (AD) and *CFTR* (AR) disorders confirmed close concordance between predicted and observed case counts. The resulting datasets, available in both machine-readable and human-friendly formats, support Bayesian variant interpretation and clinical decision-making. [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. 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.


Variant Risk Estimate (VRE)
ClinVar variant classifications (6.8 M entries)
PanelAppRex Gene-disease (58,592 entries)
AlphaMissense 71M variants 19k genes
VRE calculations
GnomAD Allele frequencies (N = 807,162) (800 M variants)
dbNSFP (122 M variant annotations)
IEI genetics
Variant observation prior probabilities

Acronyms

ACMG American College of Medical Genetics and Genomics . . . 30
ACAT Aggregated Cauchy Association Test . . . 30
AD Autosomal Dominant . . . 4
ANOVA Analysis of Variance . . . 12
AR Autosomal Recessive . . . 4
BMF Bone Marrow Failure . . . 18
CD Complement Deficiencies . . . 19
CI Confidence Interval . . . 15
CF Cystic Fibrosis . . . 10
CFTR Cystic Fibrosis Transmembrane Conductance Regulator . . . 5
CVID Common Variable Immunodeficiency . . . 8
dbNSFP database for Non-Synonymous Functional Predictions . . . 5
GE Genomics England . . . 5
gnomAD Genome Aggregation Database . . . 5
HGVS Human Genome Variation Society . . . 5
HPC High-Performance Computing . . . 8
HWE Hardy-Weinberg Equilibrium . . . 4
IEI Inborn Errors of Immunity . . . 4
Ig Immunoglobulin . . . 23
InDel Insertion/Deletion . . . 5
IUIS International Union of Immunological Societies . . . 6
LD Linkage Disequilibrium . . . 21
LOEUF Loss-Of-function Observed/Expected Upper bound Fraction . . . 12
LOF Loss-of-Function . . . 18
MOI Mode of Inheritance . . . 4
NFKB1 Nuclear Factor Kappa B Subunit 1 . . . 5
OMIM Online Mendelian Inheritance in Man . . . 27
PID Primary Immunodeficiency . . . 4
PPI Protein-Protein Interaction . . . 5
SNV Single Nucleotide Variant . . . 4
SKAT Sequence Kernel Association Test . . . 30
STRINGdb Search Tool for the Retrieval of Interacting Genes/Proteins . . . 5
HSD Honestly Significant Difference . . . 12
UMAP Uniform Manifold Approximation and Projection . . . 18
UniProt Universal Protein Resource . . . 5
VEP Variant Effect Predictor . . . 5
XL X-Linked . . . 4

1 Introduction

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

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

Quantifying the risk that a newborn inherits a disease-causing variant is a fundamental challenge in genomics. Classical statistical approaches grounded in HWE (4; 5) have long been used to calculate genetic MOI probabilities for Single Nucleotide Variant (SNV)s. However, applying these methods becomes more complex when accounting for different MOI, such as Autosomal Recessive (AR) versus Autosomal Dominant (AD) or X-Linked (XL) disorders. In AR conditions, for example, the occurrence probability must incorporate both the homozygous state and compound heterozygosity, whereas for AD and XL disorders, a single pathogenic allele is sufficient to cause disease. Advances in genetic research have revealed that MOI can be even more complex (6). Mechanisms such as dominant negative effects, haploinsufficiency, mosaicism, and digenic or epistatic interactions can further modulate disease risk and clinical presentation, underscoring the need for nuanced approaches in risk estimation. Karczewski et al. (7) made significant advances; however, the remaining challenge lay in applying the necessary statistical genomics data across all MOI for any gene-disease combination Similar approaches have been reported for disease such Wilson disease, Mucopolysaccharidoses, Primary ciliary dyskinesia, and treatable metabolic diseasse, (8; 9), as reviewed by Hannah et al. (10).

To our knowledge all approaches to date have been limited to single MOI, specific to the given disease, or restricted to a small number of genes. We argue that our integrated approach is highly powerful because the resulting probabilities can serve as informative priors in a Bayesian framework for variant and disease probability estimation; a perspective that is often overlooked in clinical and statistical genetics. Such a framework not only refines classical HWE-based risk estimates but also has the potential to enrich clinicians' understanding of what to expect in a patient and to enhance the analytical models employed by bioinformaticians. The dataset also holds

value for AI and reinforcement learning applications, providing an enriched version of the data underpinning frameworks such as AlphaFold (11) and AlphaMissense (12).

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

2 Methods

2.1 Dataset

Data from Genome Aggregation Database (gnomAD) v4 comprised 807,162 individuals, including 730,947 exomes and 76,215 genomes (7). This dataset provided 786,500,648 SNVs and 122,583,462 InDels, with variant type counts of 9,643,254 synonymous, 16,412,219 missense, 726,924 nonsense, 1,186,588 frameshift and 542,514 canonical splice site variants. ClinVar data were obtained from the variant summary dataset (as of: 16 March 2025) available from the NCBI FTP site, and included 6,845,091 entries, which were processed into 91,319 gene classification groups and a total of 38,983 gene classifications; for example, the gene *A1BG* contained four variants 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

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 extend this by incorporating the contribution from other pathogenic alleles.

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 mode of 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 manifested when two pathogenic alleles were present. In this case, we accounted for both the homozygous state and the possibility of compound heterozygosity:

  $$p_{\text{disease},i} = p_i^2 + 2\,p_i\big(P_{\text{tot}} - p_i\big).$$

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

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

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

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

4. Aggregation by Gene and ClinVar Classification. For each gene and for each ClinVar classification (e.g. "Pathogenic", "Likely pathogenic", "Uncertain significance", etc.), we aggregated the results across all variants. The total expected cases for a given group was:

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

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

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

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

For each data chunk, the relevant fields were gene name, position, allele number, allele frequency, ClinVar classification, and HGVS annotations. Missing classifications (denoted by ".") were replaced with zeros and allele frequencies were converted to numeric values. We then retained only the first transcript allele annotation for simplicity, as the analysis was based on genomic coordinates. Subsequently, the variant data were merged with gene panel data from PanelAppRex to obtain the disease-related MOI mode for each gene. For each gene, if no variant was observed for a given ClinVar classification (i.e. $p_i = 0$), a minimal risk was assigned as described above. Finally, we computed the occurrence probability, expected cases, and the probability of observing at least one case using the equations presented.

The final results were aggregated by gene and ClinVar classification and used to generate summary statistics that reviewed the predicted disease observation probabilities.

2.3 Validation of Autosomal Dominant Estimates Using *NFKB1*

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

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

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

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

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

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

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

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

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

   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

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

   which yielded a Bayesian mixture adjusted median estimate of 67 cases with a 95% credible interval of approximately 43 to 99 cases.

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

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

2.4 Validation Study for Autosomal Recessive CF Using CFTR

To validate our framework for AR diseases, we focused on Cystic Fibrosis (CF). For comparability sizes between the validation studies, we analysed the most common SNV in the *CFTR* gene, typically reported as "p.Arg117His" (GRCh38 Chr 7:117530975 G/A, MANE Select HGVSp ENST00000003084.11: p.Arg117His). Our goal was to validate our genome-wide probability estimates by comparing the expected number of CF cases attributable to the p.Arg117His variant in *CFTR* with 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

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

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

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

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:

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

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.

2.5 Validation of SCID-specific Estimates Using PID–SCID Genes

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

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

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

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

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

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

2.6 Protein Network and Genetic Constraint Interpretation

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

2.7 Gene Set Enrichment Test

To test for overrepresentation of biological functions, the prioritised genes were compared against gene sets from MsigDB (including hallmark, positional, curated, motif, computational, GO, oncogenic, and immunologic signatures) and WikiPathways using hypergeometric tests with FUMA (24; 25). The background set consisted of 24,304

genes. Multiple testing correction was applied per data source using the Benjamini-Hochberg method, and gene sets with an adjusted P-value ≤ 0.05 and more than one overlapping gene are reported.

2.8 Deriving novel PID classifications by genetic PPI and clinical features

We recategorised 315 immunophenotypic features from the original IUIS IEI annotations, reducing the original multi-level descriptors (e.g. "decreased cd8, normal or decreased cd4") first to minimal labels (e.g."low") and second to binary outcomes (normal vs. not-normal) for T cells, B cells, neutrophils, and immunoglobulins Each gene was mapped to its PPI cluster derived from STRINGdb and UMAP embeddings from previous steps. We first tested for non-random associations between these four binary immunophenotypes and PPI clusters using χ^2 tests. To generate a data-driven PID classification, we trained a decision tree (rpart) to predict PPI cluster membership from the four immunophenotypic features plus the traditional IUIS Major and Subcategory labels. Hyperparameters (complexity parameter $= 0.001$, minimum split $= 10$, minimum bucket $= 5$, maximum depth $= 30$) were optimised via five-fold cross validation using the `caret` framework. Terminal node assignments were then relabelled according to each group's predominant abnormal feature profile.

2.9 Probability of observing AlphaMissense pathogenicity

We obtained the subset pathogenicity predictions from AlphaMissense via the AlphaFold database and whole genome data from the studies data repository(12; 26). The AlphaMissense data (genome-aligned and amino acid substitutions) were merged with the panel variants based on genomic coordinate and HGVSc annotation. Occurrence probabilities were log-transformed and adjusted (y-axis displaying log10(occurrence prob + 1e-5) + 5)), to visualise the distribution of pathogenicity scores across the residue sequence. A Kruskal-Wallis test was used to compare the observed disease probability across clinical classification groups.

3 Results

3.1 Observation Probability Across Disease Genes

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

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

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 inVar 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. Additional columns, such as population allele frequency, are not shown. (27)

Gene	Panel ID	ClinVar Clinical Significance	GRCh38 Pos	HGVSc (VEP)	HGVSp (VEP)	Inheritance	Occurrence Probability
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

3.2 Validation studies

3.2.1 Validation of Dominant Disease Occurrence with *NFKB1*

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

Our model calculated 0 known pathogenic variant *NFKB1*-related CVID cases in the UK with a minimal risk of 456 unknown de novo variants. In the reference cohort, 16 *NFKB1* CVID cases were reported. We additionally wanted to account for potential under-reporting in the reference study. We used an extrapolated national CVID prevalence which yielded a median estimate of 118 cases (95% CI: 70–181), while a Bayesian-adjusted mixture estimate produced a median of 67 cases (95% CI: 43–99). **Figure S1 (A)** illustrates that our predicted values reflect these ranges and are closer to the observed count. This case supports the validity of our integrated

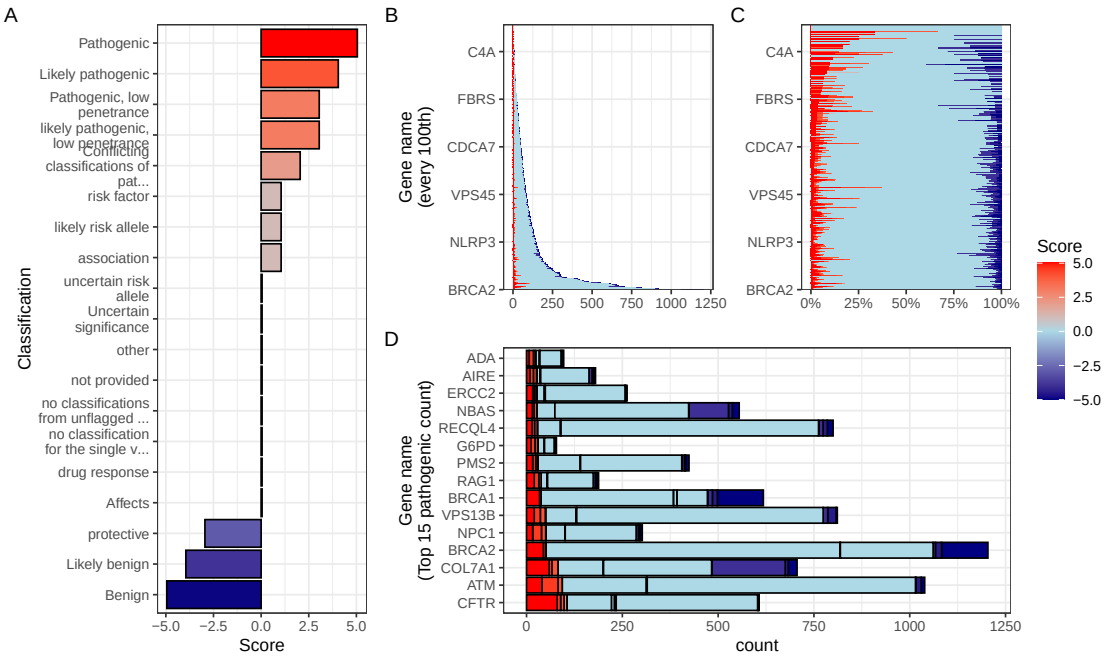


Figure 1: **Summary of ClinVar clinical significance classifications in the PID gene panel.** (A) Shows the numeric score coding for each classification. Panels (B) and (C) display the tally of classifications per gene as absolute counts and as percentages, respectively. (D) Highlights the top 15 genes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

probability estimation framework for AD disorders, and represents a challenging example where pathogenic SNV are not reported in the reference population of gnomAD. Our min-max values successfully contained the true reported values.

3.2.2 Validation of Recessive Disease Occurrence with *CFTR*

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

In contrast, the nationally reported number of CF cases was 714, as recorded in the UK Cystic Fibrosis Registry 2023 Annual Data Report (21). To account for factors such as reduced penetrance and the mortality-adjusted expected genotype, we derived a Bayesian-adjusted estimate via posterior simulation. Our Bayesian approach yielded a median estimate of 740 cases (95% Confidence Interval (CI): 696, 786) and a mixture-based estimate of 727 cases (95% CI: 705, 750). **Figure S1 (B)** illustrates the close concordance between the predicted values, the Bayesian-adjusted estimates, and the national report supports the validity of our approach for estimating disease.

Figure S2 shows the final values for these genes of interest in a given population size and phenotype. It reveals that an allele frequency threshold of approximately 0.000007 is required to observe a single heterozygous disease-causing variant carrier in the UK population for both genes. However, owing to the AR MOI pattern of *CFTR*, this threshold translates into more than 100,000 heterozygous carriers, compared to only 456 carriers for the AD gene *NFKB1*. Note that this allele frequency threshold, being derived from the current reference population, represents a lower bound that can become more precise as public datasets continue to grow. This marked difference underscores the significant impact of MOI patterns on population carrier frequencies and the observed disease prevalence.

3.2.3 Interpretation of ClinVar Variant Observations

Figure S9 shows the two validation study PID genes, representing AR and dominant MOI. **Figure S9 (A)** illustrates the overall probability of an affected birth by ClinVar variant classification, whereas **Figure S9 (B)** depicts the total expected number of cases per classification for an example population, here the UK, of approximately 69.4 million.

3.2.4 Validation of SCID-specific Disease Occurrence

Given that SCID is a subset of PID, our probability estimates reflect the likelihood of observing a genetic variant as a diagnosis when the phenotype is PID. However, we additionally tested our results against SCID cohorts in **Figure S4**. The summarised raw cohort data for SCID-specific gene counts are summarised and compared across countries in **Figure S3**. True counts for *IL2RG* and *DCLRE1C* from ten distinct locations yielded 95% confidence intervals surrounding our predicted values. For *IL2RG*, the prediction was low (approximately 1 case per 1,000,000 PID), as expected since loss-of-function variants in this X-linked gene are highly deleterious and rarely observed in gnomAD. In contrast, the predicted value for *RAG1* was substantially higher (553 cases per 1,000,000 PID) than the observed counts (ranging from 0 to 200). We attributed this discrepancy to the lower penetrance and higher background frequency of *RAG1* variants in recessive inheritance, whereby reference studies may underreport the true national incidence. Overall, we argued that agreement within an order of magnitude was tolerable given the inherent uncertainties from reference studies arising from variable penetrance and allele frequencies.

3.3 Genetic constraint in high-impact protein networks

We next examined genetic constraint in high-impact protein networks across the whole IEI gene set of over 500 known disease-gene phenotypes (1). By integrating ClinVar variant classification scores with PPI data, we quantified the pathogenic burden per

gene and assessed its relationship with network connectivity and genetic constraint (7; 16).

3.3.1 Score-Positive-Total within IEI PPI network

The ClinVar classifications reported in **Figure 1** were scaled -5 to +5 based on their pathogenicity. We were interested in positive (potentially damaging) but not negative (benign) scoring variants, which are statistically incidental in this analysis. We tallied gene-level positive scores to give the score positive total metric. **Figure 2 (A)** shows the PPI network of disease-associated genes, where node size and colour encode the score positive total (log-transformed). The top 15 genes with the highest total prior probabilities of being observed with disease are labelled (as per **Figure 1**).

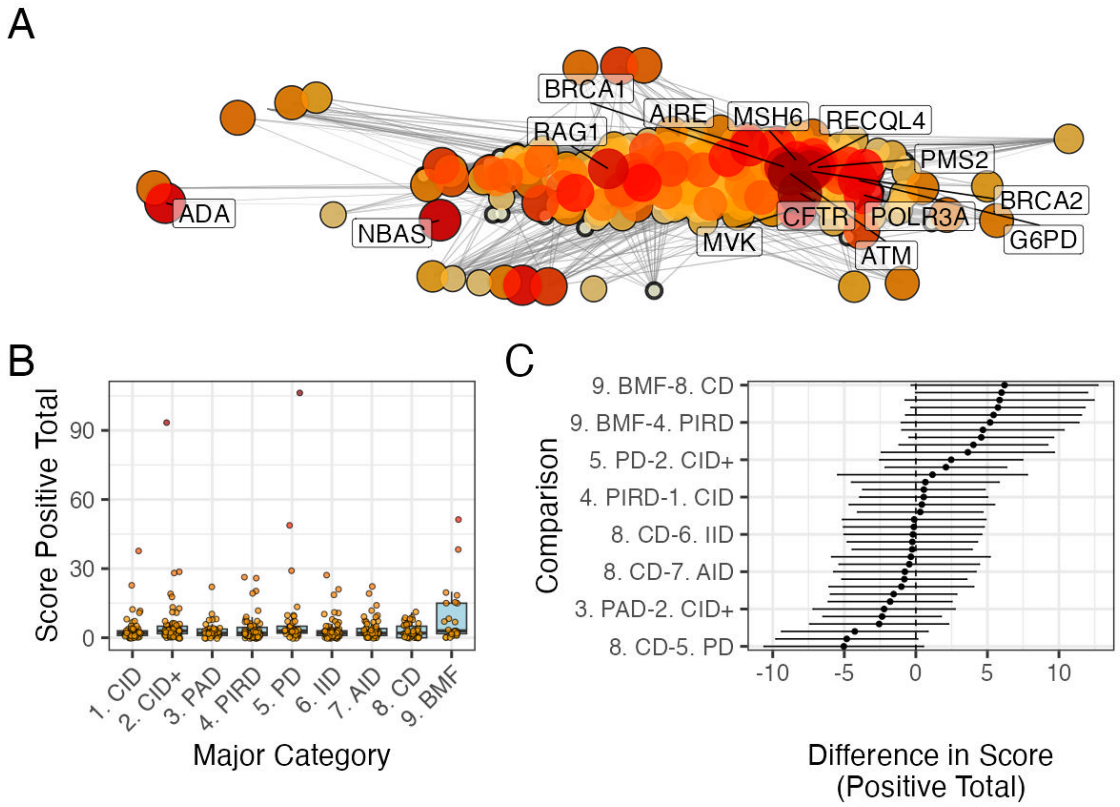


Figure 2: **PPI network and score positive total ClinVar significance variants**. (A) PPI network of disease-associated genes. Node size and colour represent the log-transformed score positive total, the top 15 genes/proteins with the highest probability of being observed in disease are labelled. (B) 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.

3.3.2 Association Analysis of Score-Positive-Total across IEI Categories

We checked for any statistical enrichment in score positive totals, which represents the expected observation of pathogenicity, between the IEI categories. The one-way ANOVA revealed an effect of major disease category on score positive total ($F(8, 500) = 2.82$, $p = 0.0046$), indicating that group means were not identical, which we observed in **Figure 2 (B)**. However, despite some apparent differences in median scores across categories (i.e. 9. Bone Marrow Failure (BMF)), the Tukey HSD post hoc comparisons **Figure 2 (C)** showed that all pairwise differences had 95% confidence intervals overlapping zero, suggesting that individual group differences were not significant.

3.3.3 UMAP Embedding of the PPI Network

To address the density of the PPI network for the IEI gene panel, we applied Uniform Manifold Approximation and Projection (UMAP) (**Figure 3**). Node sizes reflect interaction degree, a measure of evidence-supported connectivity (16). We tested for a correlation between interaction degree and score positive total. In **Figure 3**, gene names with degrees above the 95th percentile are labelled in blue, while the top 15 genes by score positive total are labelled in yellow (as per **Figure 1**). Notably, genes with high pathogenic variant loads segregated from highly connected nodes, suggesting that Loss-of-Function (LOF) in hub genes is selectively constrained, whereas damaging variants in lower-degree genes yield more specific effects. This observation was subsequently tested empirically.

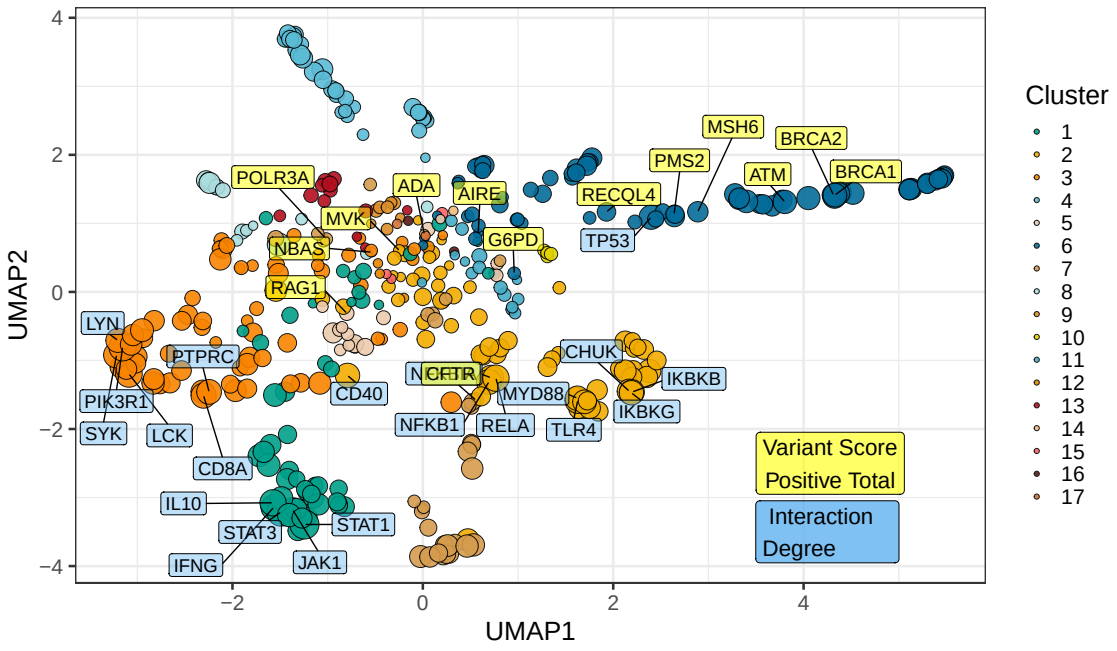


Figure 3: **UMAP embedding of the PPI network (p_umap).** The plot projects the high-dimensional protein-protein interaction network into two dimensions, with nodes coloured by cluster and sized by interaction degree. Blue labels indicate hub genes (degree above the 95th percentile) and yellow labels 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.

3.3.4 Hierarchical Clustering of Enrichment Scores for Major Disease Categories

Figure S5 presents a heatmap of standardised residuals for major disease categories across network clusters, as per **Figure 3**. A dendrogram clusters similar disease categories, while the accompanying bar plot displays the maximum absolute standardised residual for each category. Notably, (8) Complement Deficiencies (CD) shows the highest maximum enrichment, followed by (9) BMF. While all maximum values exceed 2, the threshold for significance, this likely reflects the presence of protein clusters with strong damaging variant scores rather than uniform significance across all categories (i.e. genes from cluster 4 in 8 CD).

3.3.5 PPI Connectivity, LOEUF Constraint and Enriched Network Cluster Analysis

Based on the preliminary insight from **Figure S5** , we evaluated the relationship between network connectivity (PPI degree) and LOF constraint (LOEUF upper rank) Karczewski et al. (7) using Spearman's rank correlation. Overall, there was a weak

but significant negative correlation ($\rho = -0.181$, $p = 0.00024$) at the global scale, indicating that highly connected genes tend to be more constrained. A supplementary analysis (**Figure S6**) did not reveal distinct visual associations between network clusters and constraint metrics, likely due to the high network density. However once stratified by gene clusters, the natural biological scenario based on quantitative PPI evidence (16), some groups showed strong correlations; for instance, cluster 2 ($\rho = -0.375$, $p = 0.000994$) and cluster 4 ($\rho = -0.800$, $p < 0.000001$), while others did not. This indicated that shared mechanisms within pathway clusters may underpin genetic constraints, particularity for LOF intolerance. We observe that the score positive total metric effectively summarises the aggregate pathogenic burden across IEI genes, serving as a robust indicator of genetic constraint and highlighting those with elevated disease relevance.

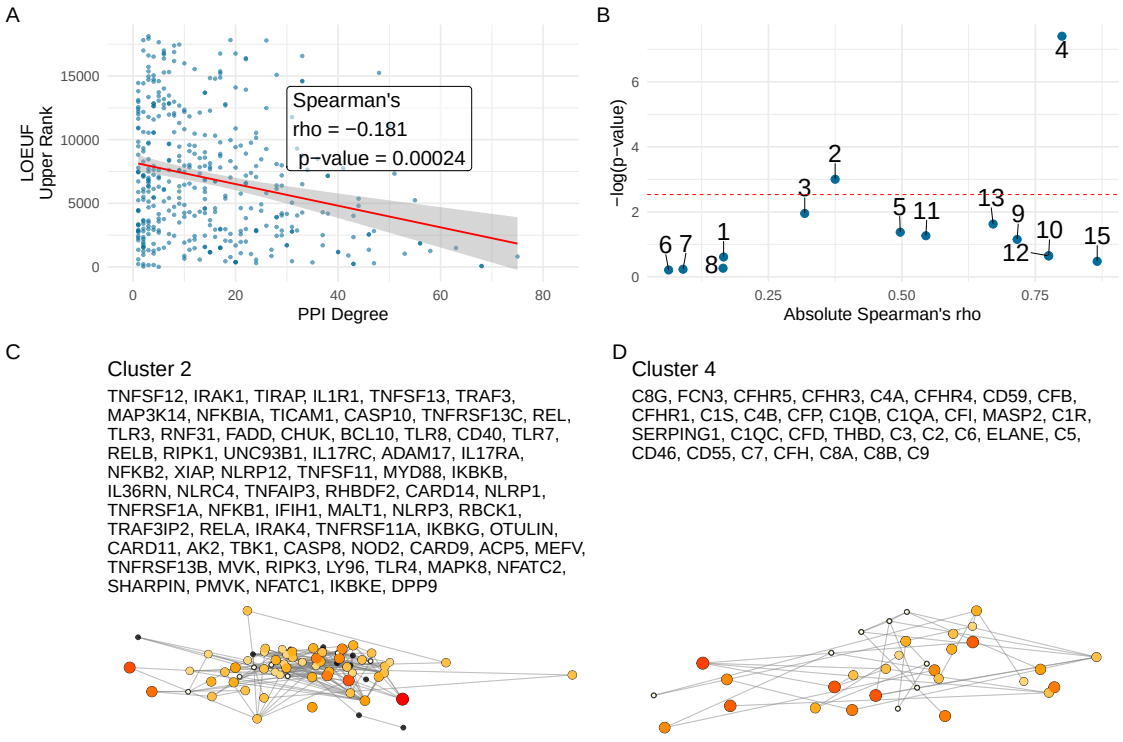


Figure 4: **Correlation between PPI degree and LOEUF upper rank. (A)** Ananlysis across all genes revealed a weak, significant negative correlation between PPI degree and LOEUF upper rank. **(B)** The cluster-wise analysis showed that clusters 2 and 4 exhibited moderate to strong correlations, while 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.

Figure 4 (C, D) shows the re-plotted PPI networks for clusters with significant correlations between PPI degree and LOEUF upper rank. In these networks, node size is scaled by a normalised variant score, while node colour reflects the variant score according to a predefined palette.

3.4 New Insight from Functional Enrichment

To interpret the functional relevance of our prioritised IEI gene sets with the highest load of damaging variants (i.e. clusters 2 and 4 in **Figure 4**), we performed functional enrichment analysis for known disease associations using MsigDB with FUMA (i.e. GWAScatalog and Immunologic Signatures) (24). Composite enrichment profiles (**Figure S7**) reveal that our enriched PPI clusters were associated with distinct disease-related phenotypes, providing functional insights beyond traditional IUIS IEI groupings (1). The gene expression profiles shown in **Figure S8** (GTEx v8 54 tissue types) offer the tissue-specific context for these associations. Together, these results enable the annotation of IEI gene sets with established disease phenotypes, supporting a data-driven classification of IEI.

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

3.5 Genome-wide Gene Distribution and Locus-specific Variant Occurrence

Figure 5 (A) shows a genome-wide karyoplot of all IEI panel genes across GRCh38, with colour-coding based on MOI. Figures **(B)** and **(C)** display zoomed-in locus plots for *NFKB1* and *CFTR*, respectively. In **Figure 5 (B)**, the probability of observing variants with known classifications is high only for variants such as p.Ala475Gly, which are considered benign in the AD *NFKB1* gene that is intolerant to LOF. In **Figure 5 (C)**, high probabilities of observing patients with pathogenic variants in *CFTR* are evident, reproducing this well-established phenomenon. Furthermore, the analysis of Linkage Disequilibrium (LD) using R^2 shows that high LD regions can be modelled effectively, allowing independent variant signals to be distinguished.

3.6 Novel PID classifications derived from genetic PPI and clinical features

We recategorised 315 immunophenotypic features from the original IUIS IEI annotations, reducing detailed descriptions (e.g. “decreased cd8, normal or decreased cd4”), first to minimal labels (e.g.“low”), and second to binary outcomes (normal vs. not-normal) for T cells, B cells, neutrophils, and immunoglobulins (**Figure 6**). These

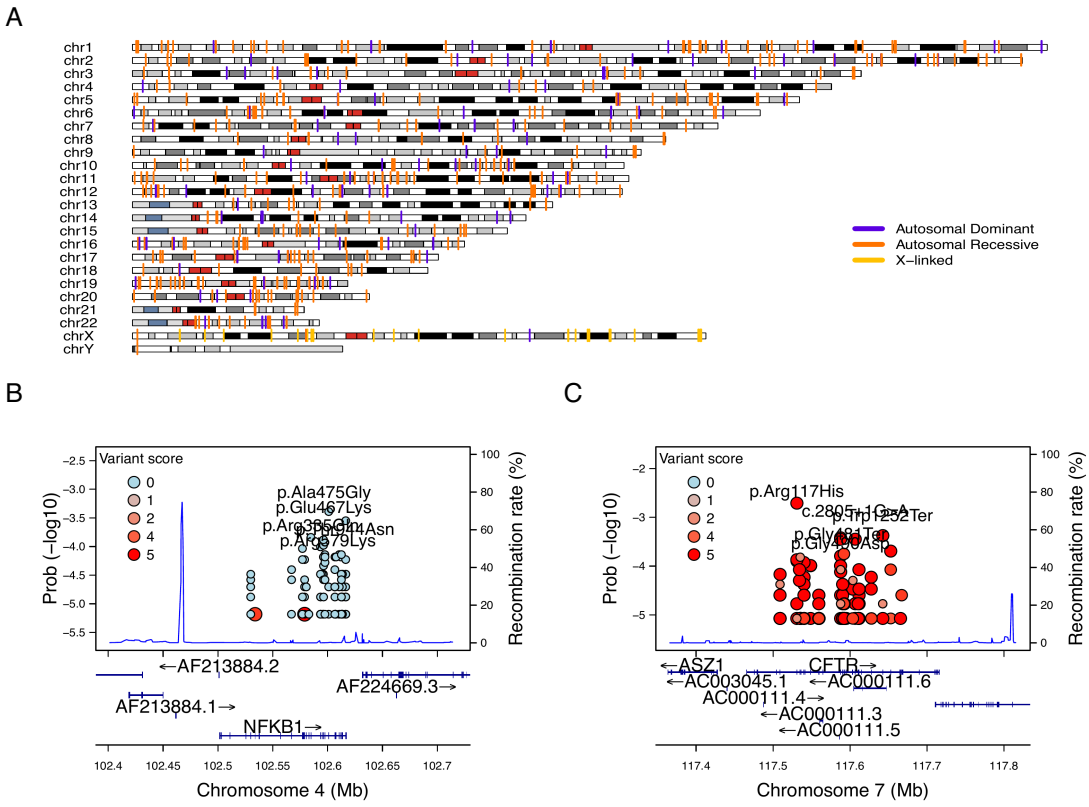


Figure 5: **Genome-wide IEI, variant occurance probability and LD by R^2.** (A) Genome-wide karyoplot of all IEI panel genes mapped to GRCh38, with colours indicating MOI. (B) Zoomed-in locus plot for *NFKB1* showing variant observation probabilities; only benign variants such exhibit high probabilities 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 S10**). The resulting novel PID classifications, illustrated by the decision tree and gene group distributions (**Figure 9**), provide a more coherent and data-driven framework for categorising PID genes.

3.7 Novel PID classifications derived from genetic PPI and clinical features

We recategorised 315 immunophenotypic features from the original IUIS IEI annotations, 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 cells, B cells, Immunoglobulin (Ig) and neutrophils (**Figure 6**). These simplified profiles were mapped onto STRINGdb PPI clusters, revealing non-random distributions ($\chi^2 < 1e$-13; **Figure 7**), indicating that network context captures key immunophenotypic variation.

We next compared four classifiers under 5-fold cross-validation to determine which features predicted PPI clustering. As shown in **Figure 8**, the fully combined model achieved the highest accuracy among the four: (i) phenotypes only (33 %) (i.e. T cell, B cell, Ig, Neutrophil); (ii) phenotypes + IUIS major category (50 %) (e.g. CID. See **Box 2.1** for more); (iii) IUIS major + subcategory only (59 %) (e.g. CID, T-B+ SCID); and (iv) phenotypes + IUIS major + subcategory (61 %). This demonstrated that incorporating both traditional IUIS classifications and core immunophenotypic markers into the PPI-based framework yielded the most robust discrimination of PID gene clusters. Variable importance analysis highlighted abnormality status for Ig and T cells were among the top ten features in addition to the other IUIS major and sub categories. Per-class specificity remained uniform across the classes while sensitivity dropped.

The PPI and immunophenotype model yielded 17 data-driven PID groups, whereas incorporating the full complement of IUIS categories expanded this to 33 groups. For clarity, we only demonstrate the decision tree from the smaller 17-group model in **Figure 9**. Each terminal node is annotated by its predominant immunophenotypic signature (for example, "group 65 with abnormal T cell and B cell features"), and the full resulting gene counts per 33 class are plotted in **Figure 9**. Although, less user-friendly, this data-driven taxonomy both aligns with and refines traditional IUIS IEI classifications to provide a scaffold for large-scale computational analyses. Because this framework is fully reproducible, alternative PPI embeddings or incorporate additional molecular annotations can readily swapped to continue building on these PID classification schemes.

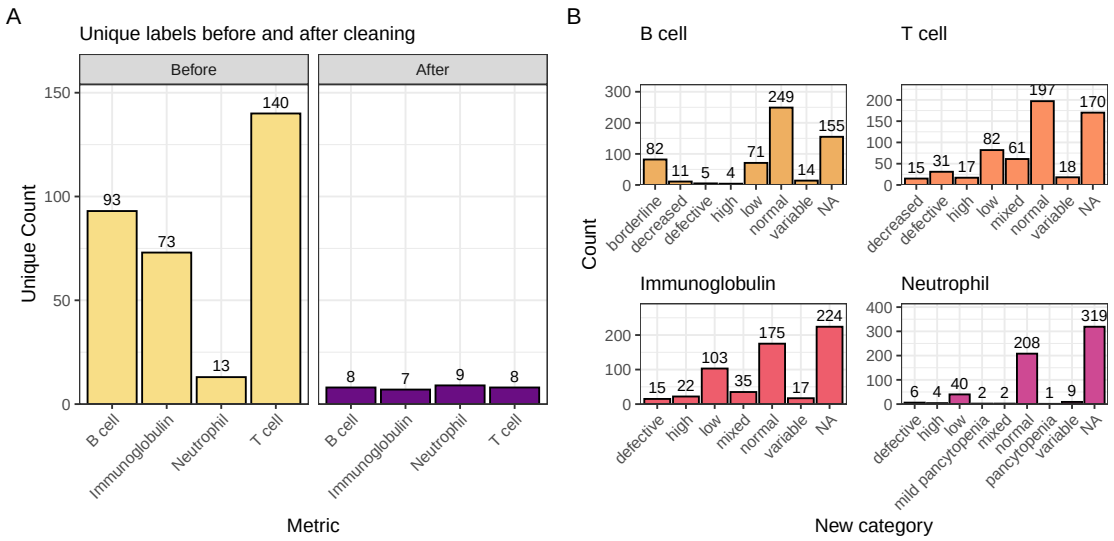


Figure 6: **Distribution of immunophenotypic features before and after recategorisation.** The original IUIS IEI descriptions contain information such as T cell-related "decreased cd8, normal or decreased cd4 cells" which we recategorise as "low". The bar plot shows the count of unique labels for each status (normal, not_normal) across the T cell, B cell, Ig, and Neutrophil features.

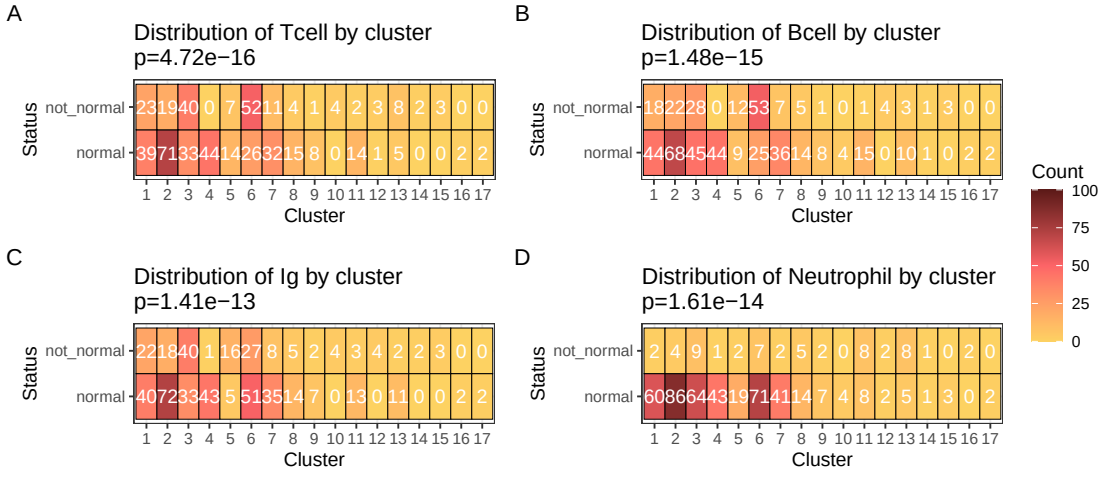


Figure 7: **Heatmaps of clinical feature distributions by PPI cluster.** The heatmaps display the count of observations for abnormality of each clinical feature (A) T cell, (B) B cell, (C) Immunoglobulin, (D) Neutrophil, in relation to the PPI clusters, with p-values from chi-square tests annotated in the titles.

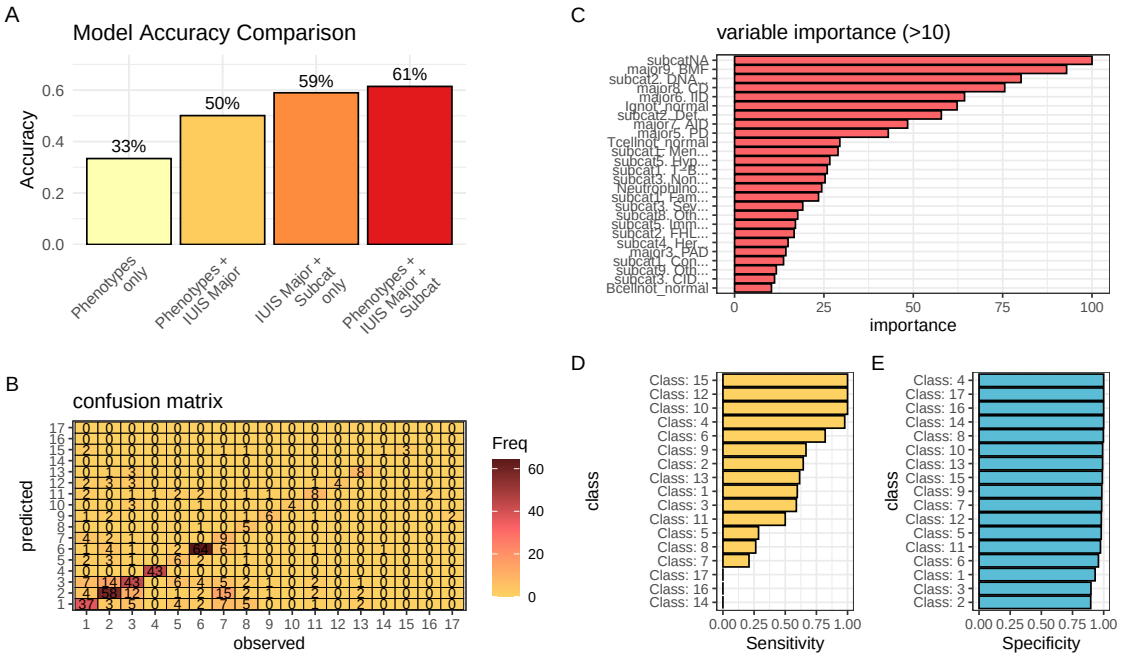


Figure 8: **Performance comparison of PID classifiers.** Classification predicting PPI cluster membership from IUIS major category, subcategory, and immunological features. (A) Overall accuracy for four rpart models used to predict PPI clustering. The combined model achieves 61.4 % accuracy, exceeding 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.

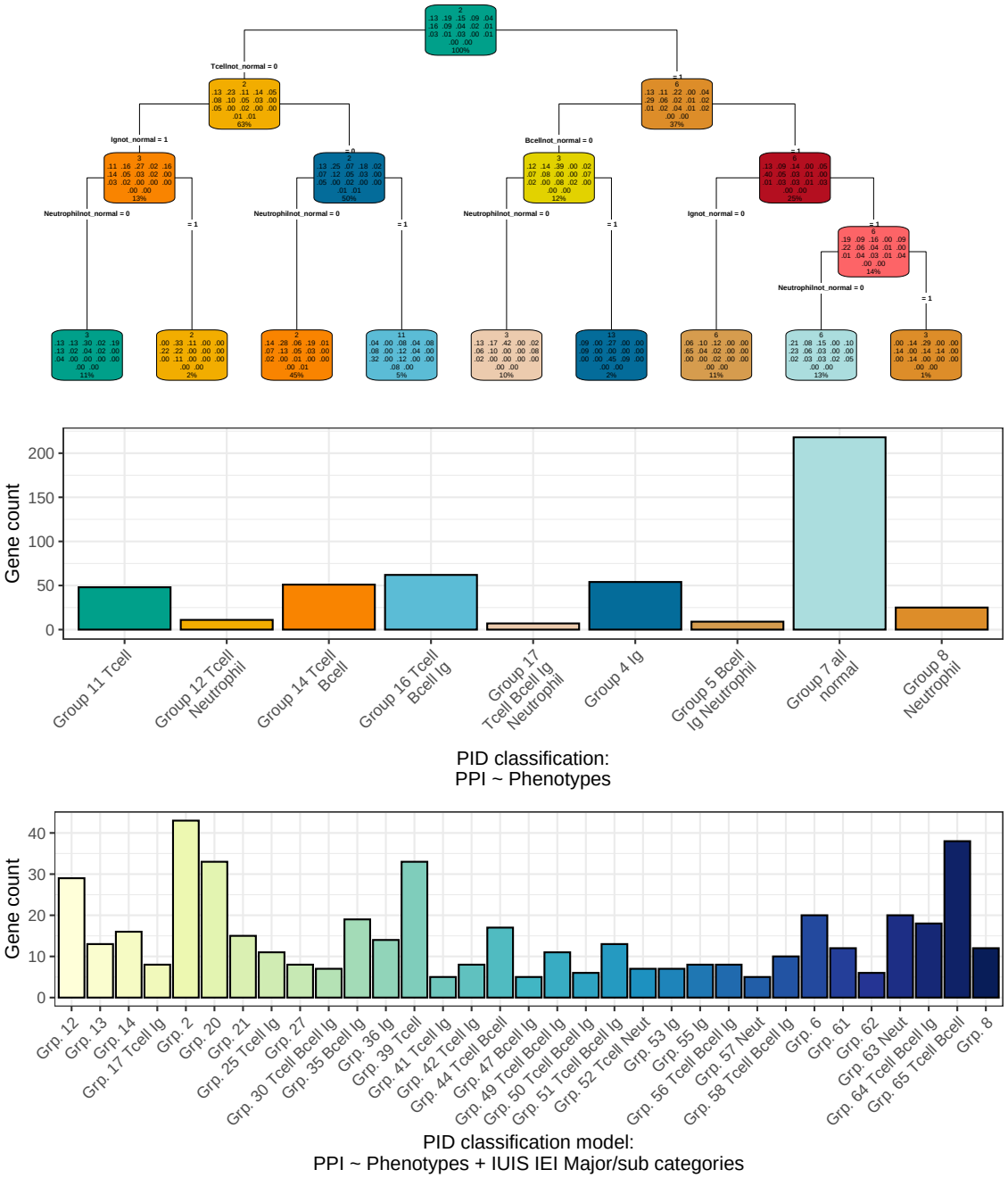


Figure 9: **Fine-tuned model for PID classification.** (Top) In each terminal node, the top block indicates the number of genes in the node; the middle block shows the fitted class probabilities (which sum to 1); and the bottom block displays the percentage of the total sample in that node. These metrics 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.

3.7.1 Integration of Variant Probabilities into IEI Genetics Data

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

Figure 10 shows the interface summarising integrated variant data. Server-side pre-calculation of summary statistics minimises browser load, while clinical significance is converted to numerical metrics. Key quantiles (min, Q1, median, Q3, max) for each gene are rendered as sparkline box plots, and dynamic URLs link table entries to external databases (e.g. ClinVar, Online Mendelian Inheritance in Man (OMIM), AlphaFold).

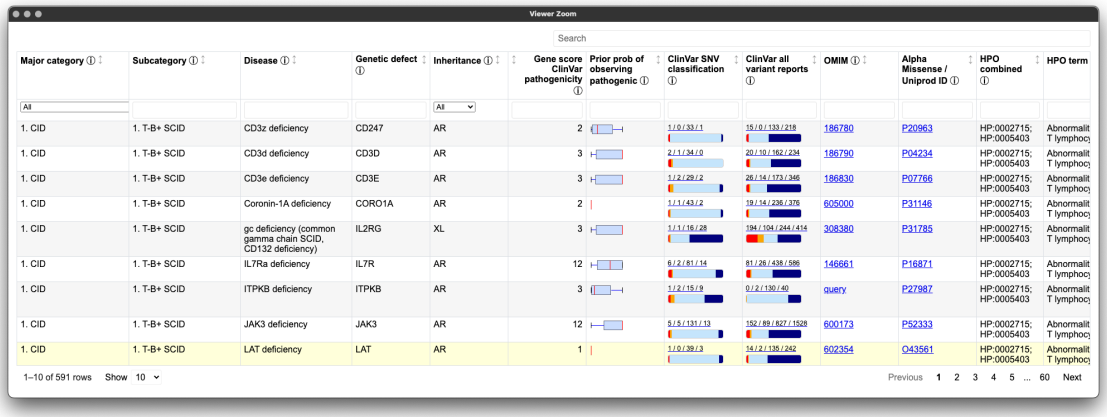


Figure 10: **Integration of variant probabilities into the IEI genetics framework.** The interface summarises the condensed variant data, with pre-calculated summary statistics and dynamic links to external databases. This integration enables immediate access to detailed variant classifications and prior probabilities for each gene.

3.8 Probability of observing AlphaMissense pathogenicity

AlphaMissense provides pathogenicity scores for all possible amino acid substitutions; however, our results in **Figure 11** show that the most probable observations in patients occur predominantly for benign or unknown variants. This finding places the likelihood of disease-associated substitutions into perspective and offers a data-driven foundation for future improvements in variant prediction. The values in **Figure 11 (A)** can be directly compared to **Figure 1 (D)** to view the distribution of classifications. A Kruskal-Wallis test was used to compare the observed disease probability

across clinical classification groups and no significant differences were detected. In general, most variants in patients are classified as benign or unknown, indicating limited discriminative power in the current classification, such that pathogenicity prediction does not infer observation prediction (**Figure S11**).

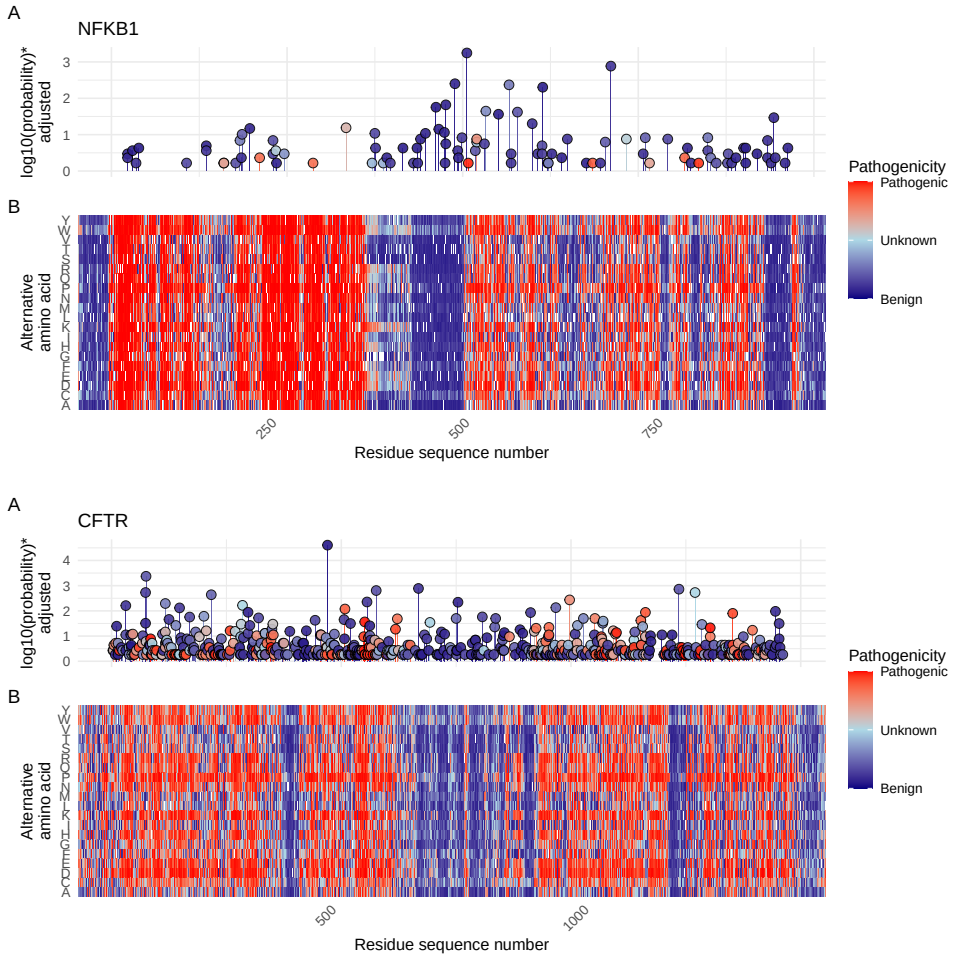


Figure 11: **(A) Probabilities of observing a patient with (B) AlphaMissense-derived pathogenicity scores.** Although AlphaMissense provides scores for every possible amino acid substitution, the most frequently observed variants in patients tend to be classified as benign or of unknown significance. 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.

4 Discussion

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

Our approach yielded two key results. First, our detailed, per-variant pre-calculated results provide prior probabilities of observing disease-associated variants across all MOI for any gene-disease combination. Second, the score positive total metric effectively summarises the aggregate pathogenic burden across genes, serving as a robust 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:

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

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

$$P(D) = P(G)\, P(D \mid G).$$

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

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

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

Our framework addresses this challenge by combining variant classifications, population allele frequencies, and curated gene-disease associations. While imperfect on an individual level, these sources exhibit predictable aggregate behaviour, supported by James-Stein estimation principles (28). Curated gene-disease associations help identify genes that explainable for most disease cases, allowing us to approximate $P(G \mid D)$ close to one. In this way, we obtain robust estimates of $P(G)$ (the frequency of disease-associated genotypes), even when exact values of penetrance and 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 $P(G_i \mid D)$ is the probability that a case of disease D is attributable to variant i, we assume that, in aggregate,

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

Even if the cumulative contribution is slightly less than one, the resultant risk estimates remain robust within the broad confidence intervals 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.

5 Conclusion

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

Acknowledgements

We acknowledge Genomics England for providing public access to the PanelApp data. The use of data from Genomics England panelapp was licensed under the Apache License 2.0. The use of data from UniProt was licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). ClinVar asks its users who distribute or copy data to provide attribution to them as a data source in publications and websites (13). dbNSFP version 4.4a is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0); while we cite this dataset as used our research publication, it is not used for the final version which instead used ClinVar and gnomAD directly. GnomAD is licensed under Creative Commons Zero Public Domain Dedication (CC0 1.0 Universal). GnomAD request that usages cites the gnomAD flagship paper (7) and any online resources that include the data set provide a link to the browser, and note that tool includes data from the gnomAD v4.1 release. AlphaMissense asks to cite Cheng et al. (12) for usage in research, with data available from Cheng et al. (26).

Competing interest

We declare no competing interest.

References

[1] Stuart G. Tangye, Waleed Al-Herz, Aziz Bousfiha, Charlotte Cunningham-Rundles, Jose Luis Franco, Steven M. Holland, Christoph Klein, Tomohiro Morio, Eric Oksenhendler, Capucine Picard, Anne Puel, Jennifer Puck, Mikko R. J. Seppänen, Raz Somech, Helen C. Su, Kathleen E. Sullivan, Troy R. Torgerson, and Isabelle Meyts. Human Inborn Errors of Immunity: 2022 Update on the Classification from the International Union of Immunological Societies Expert Committee. *Journal of Clinical Immunology*, 42(7):1473–1507, October 2022. ISSN 0271-9142, 1573-2592. doi: 10.1007/s10875-022-01289-3. URL https://link.springer.com/10.1007/s10875-022-01289-3.

[2] Dylan Lawless. PanelAppRex aggregates disease gene panels and facilitates sophisticated search. March 2025. doi: 10.1101/2025.03.20.25324319. URL http://medrxiv.org/lookup/doi/10.1101/2025.03.20.25324319.

[3] Antonio Rueda Martin, Eleanor Williams, Rebecca E. Foulger, Sarah Leigh, Louise C. Daugherty, Olivia Niblock, Ivone U. S. Leong, Katherine R. Smith, Oleg Gerasimenko, Eik Haraldsdottir, Ellen Thomas, Richard H. Scott, Emma Baple, Arianna Tucci, Helen Brittain, Anna De Burca, Kristina Ibañez, Dalia Kasperaviciute, Damian Smedley, Mark Caulfield, Augusto Rendon, and Ellen M. McDonagh. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nature Genetics*, 51(11):1560–1565, November 2019. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-019-0528-2. URL https://www.nature.com/articles/s41588-019-0528-2.

[4] Oliver Mayo. A Century of Hardy–Weinberg Equilibrium. *Twin Research and Human Genetics*, 11(3):249–256, June 2008. ISSN 1832-4274, 1839-2628. doi: 10.1375/twin.11.3.249. URL https://www.cambridge.org/core/product/identifier/S1832427400009051/type/journal_article.

[5] Nikita Abramovs, Andrew Brass, and May Tassabehji. Hardy-Weinberg Equilibrium in the Large Scale Genomic Sequencing Era. *Frontiers in Genetics*, 11:210, March 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.00210. URL https://www.frontiersin.org/article/10.3389/fgene.2020.00210/full.

[6] Johannes Zschocke, Peter H. Byers, and Andrew O. M. Wilkie. Mendelian inheritance revisited: dominance and recessiveness in medical genetics. *Nature Reviews Genetics*, 24(7):442–463, July 2023. ISSN 1471-0056, 1471-0064. doi: 10.1038/s41576-023-00574-0. URL https://www.nature.com/articles/s41576-023-00574-0.

[7] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

[8] Sarah L. Bick, Aparna Nathan, Hannah Park, Robert C. Green, Monica H. Wojcik, and Nina B. Gold. Estimating the sensitivity of genomic newborn screening for treatable inherited metabolic disorders. *Genetics in Medicine*, 27(1):101284, January 2025. ISSN 10983600. doi: 10.1016/j.gim.2024.101284. URL https://linkinghub.elsevier.com/retrieve/pii/S1098360024002181.

[9] Benjamin D. Evans, Piotr Słowiński, Andrew T. Hattersley, Samuel E. Jones, Seth Sharp, Robert A. Kimmitt, Michael N. Weedon, Richard A. Oram, Krasimira Tsaneva-Atanasova, and Nicholas J. Thomas. Estimating disease prevalence in large datasets using genetic risk scores. *Nature Communications*, 12(1):6441, November 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-26501-7. URL https://www.nature.com/articles/s41467-021-26501-7.

[10] William B. Hannah, Mitchell L. Drumm, Keith Nykamp, Tiziano Pramparo, Robert D. Steiner, and Steven J. Schrodi. Using genomic databases to determine the frequency and population-based heterogeneity of autosomal recessive conditions. *Genetics in Medicine Open*, 2:101881, 2024. ISSN 29497744.

doi: 10.1016/j.gimo.2024.101881. URL https://linkinghub.elsevier.com/retrieve/pii/S2949774424010276.

[11] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2.

[12] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, and Žiga Avsec. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664):eadg7492, September 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adg7492. URL https://www.science.org/doi/10.1126/science.adg7492.

[13] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J Bradley Holmes, Brandi L Kattman, and Donna R Maglott. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, January 2018. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkx1153. URL http://academic.oup.com/nar/article/46/D1/D1062/4641904.

[14] The UniProt Consortium, Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Aduragbemi Adesina, Shadab Ahmad, Bowler-Barnett, and Others. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, January 2025. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkae1010. URL https://academic.oup.com/nar/article/53/D1/D609/7902999.

[15] Xiaoming Liu, Chang Li, Chengcheng Mou, Yibo Dong, and Yicheng Tu. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine*, 12(1):103, December 2020. ISSN 1756-994X. doi: 10.1186/s13073-020-00803-9. URL https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00803-9.

[16] Damian Szklarczyk, Katerina Nastou, Mikaela Koutrouli, Rebecca Kirsch, Farrokh Mehryary, Radja Hachilif, Dewei Hu, Matteo E Peluso, Qingyao Huang, Tao Fang, et al. The string database in 2025: protein networks with directionality of regulation. *Nucleic Acids Research*, 53(D1):D730–D737, 2025.

[17] Paul Tuijnenburg, Hana Lango Allen, Siobhan O. Burns, Daniel Greene, Machiel H. Jansen, and Others. Loss-of-function nuclear factor B subunit 1 (NFKB1) variants are the most common monogenic cause of common variable immunodeficiency in Europeans. *Journal of Allergy and Clinical Immunology*, 142(4):1285–1296, October 2018. ISSN 00916749. doi: 10.1016/j.jaci.2018.01.039. URL https://linkinghub.elsevier.com/retrieve/pii/S0091674918302860.

[18] WHO Scientific Group et al. Primary immunodeficiency diseases: report of a who scientific group. *Clin. Exp. Immunol.*, 109(1):1–28, 1997.

[19] Charlotte Cunningham-Rundles and Carol Bodian. Common variable immunodeficiency: clinical and immunological features of 248 patients. *Clinical immunology*, 92(1):34–48, 1999.

[20] Eric Oksenhendler, Laurence Gérard, Claire Fieschi, Marion Malphettes, Gael Mouillot, Roland Jaussaud, Jean-François Viallard, Martine Gardembas, Lionel Galicier, Nicolas Schleinitz, et al. Infections in 252 patients with common variable immunodeficiency. *Clinical Infectious Diseases*, 46(10):1547–1554, 2008.

[21] Y Naito, F Adams, S Charman, J Duckers, G Davies, and S Clarke. Uk cystic fibrosis registry 2023 annual data report. *London: Cystric Fibrosis Trust*, 2023.

[22] Carlo Castellani, CFTR2 team, et al. Cftr2: how will it help care? *Paediatric respiratory reviews*, 14:2–5, 2013.

[23] Hartmut Grasemann and Felix Ratjen. Cystic fibrosis. *New England Journal of Medicine*, 389(18):1693–1707, 2023. doi: 10.1056/NEJMra2216474. URL https://www.nejm.org/doi/full/10.1056/NEJMra2216474.

[24] Kyoko Watanabe, Erdogan Taskesen, Arjen Van Bochoven, and Danielle Posthuma. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*, 8(1):1826, November 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-01261-5. URL https://www.nature.com/articles/s41467-017-01261-5.

[25] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, June 2011. ISSN 1367-4811, 1367-4803. doi: 10.1093/bioinformatics/btr260. URL https://academic.oup.com/bioinformatics/article/27/12/1739/257711.

[26] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, and Žiga Avsec. Predictions for alphamissense, September 2023. URL https://doi.org/10.5281/zenodo.8208688.

[27] Dylan Lawless. Variant risk estimate probabilities for iei genes. March 2025. doi: 10.5281/zenodo.15111584. URL https://doi.org/10.5281/zenodo.15111584.

[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.

[29] Yaowu Liu, Sixing Chen, Zilin Li, Alanna C Morrison, Eric Boerwinkle, and Xihong Lin. Acat: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*, 104(3):410–421, 2019.

[30] Xihao Li, Zilin Li, Hufeng Zhou, Sheila M Gaynor, Yaowu Liu, Han Chen, Ryan Sun, Rounak Dey, Donna K Arnett, Stella Aslibekyan, et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature genetics*, 52 (9):969–983, 2020.

[31] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

[32] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah A Nickerson, David C Christiani, Mark M Wurfel, and Xihong Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237, 2012.

[33] Augustine Kong, Gudmar Thorleifsson, Michael L Frigge, Bjarni J Vilhjalmsson, Alexander I Young, Thorgeir E Thorgeirsson, Stefania Benonisdottir, Asmundur Oddsson, Bjarni V Halldorsson, Gisli Masson, et al. The nature of nurture: Effects of parental genotypes. *Science*, 359(6374):424–428, 2018.

[34] Laurence J Howe, Michel G Nivard, Tim T Morris, Ailin F Hansen, Humaira Rasheed, Yoonsu Cho, Geetha Chittoor, Penelope A Lind, Teemu Palviainen, Matthijs D van der Zee, et al. Within-sibship gwas improve estimates of direct genetic effects. *BioRxiv*, pages 2021–03, 2021.

[35] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al.

Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in medicine*, 17 (5):405–423, 2015.

[36] Sean V Tavtigian, Steven M Harrison, Kenneth M Boucher, and Leslie G Biesecker. Fitting a naturally scaled point system to the acmg/amp variant classification guidelines. *Human mutation*, 41(10):1734–1737, 2020.

[37] Quan Li and Kai Wang. Intervar: clinical interpretation of genetic variants by the 2015 acmg-amp guidelines. *The American Journal of Human Genetics*, 100 (2):267–280, 2017.

[38] Brent S Pedersen, Joe M Brown, Harriet Dashnow, Amelia D Wallace, Matt Velinder, Martin Tristani-Firouzi, Joshua D Schiffman, Tatiana Tvrdik, Rong Mao, D Hunter Best, et al. Effective variant filtering and expected candidate variant yield in studies of rare human disease. *NPJ Genomic Medicine*, 6(1):1–8, 2021.

[39] Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564–1573, 2010. URL https://doi.org/10.1038/nprot.2010.116.

[40] David T Miller, Kristy Lee, Noura S Abul-Husn, Laura M Amendola, Kyle Brothers, Wendy K Chung, Michael H Gollob, Adam S Gordon, Steven M Harrison, Ray E Hershberger, et al. Acmg sf v3. 2 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the american college of medical genetics and genomics (acmg). *Genetics in Medicine*, 25(8): 100866, 2023.

6 Supplemental

6.1 Validation studies

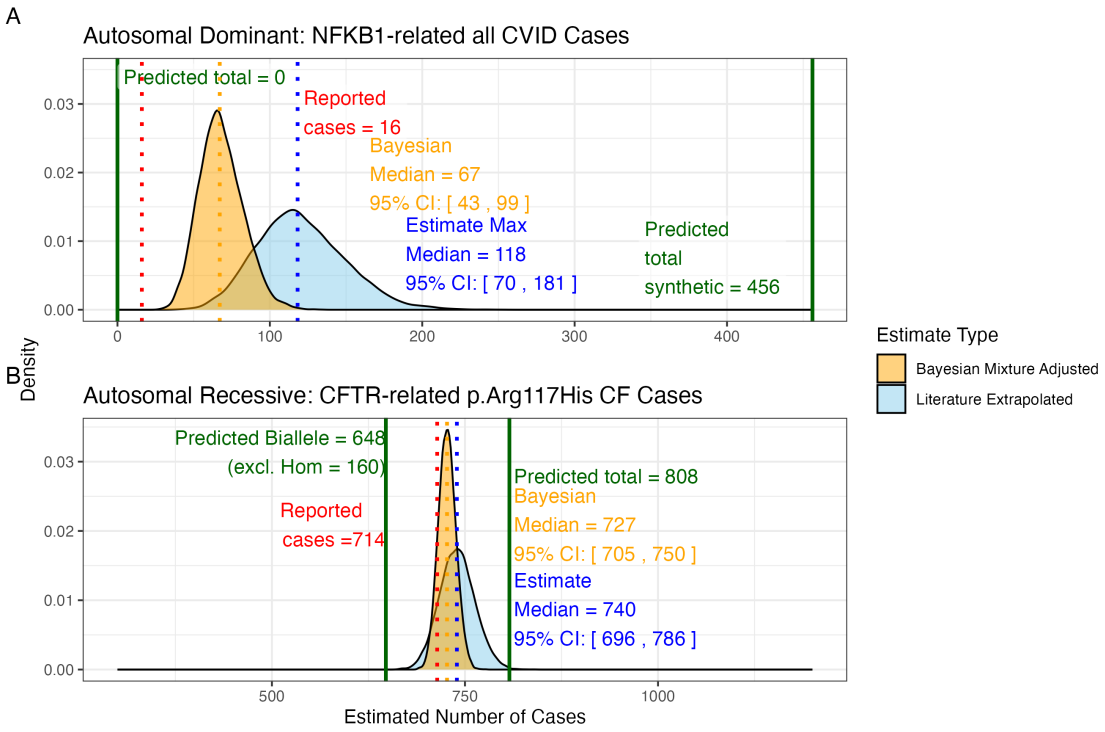


Figure S1: **Prior probabilities compared to validation disease cohort metrics.** (A) Density distributions for the number of *NFKB1*-related CVID cases in the UK. Our model (green) predicted 456 cases, which falls between the observed cohort count (red) of 390 and the upper extrapolated values. The 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.

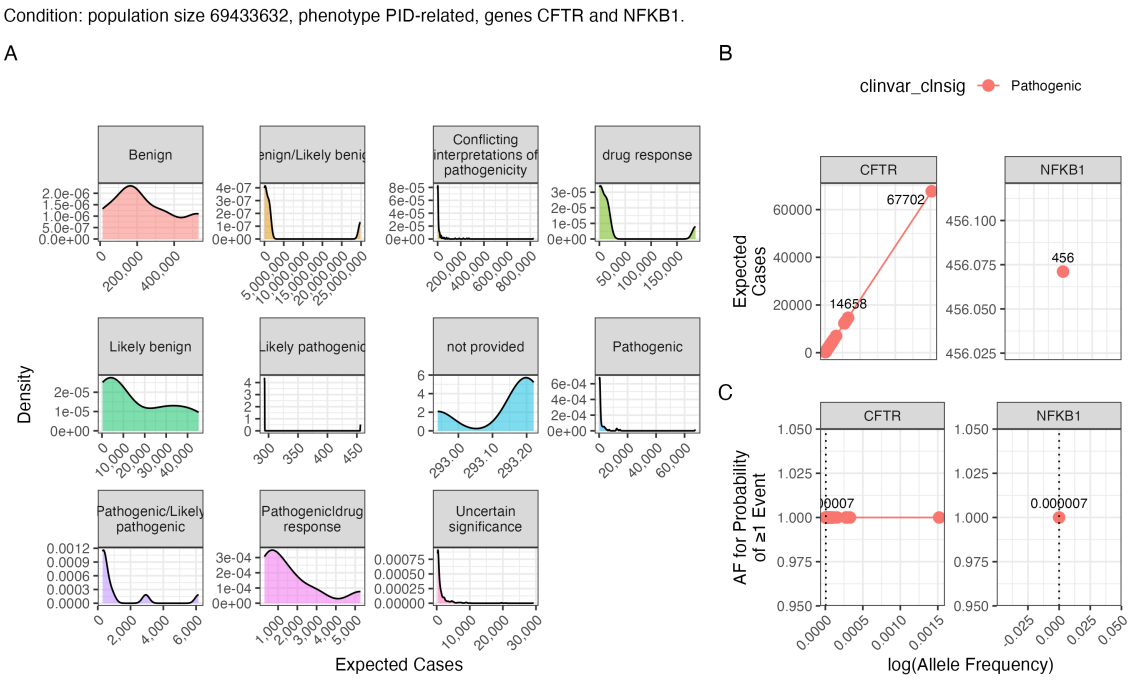


Figure S2: **Interpretation of probability of observing a variant classification.** The result from the chosen validation genes *CFTR* and *NFKB1* are shown. Case counts are dependant on the population size and phenotype. (A) The density plots of expected observations by ClinVar clinical significance. 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.

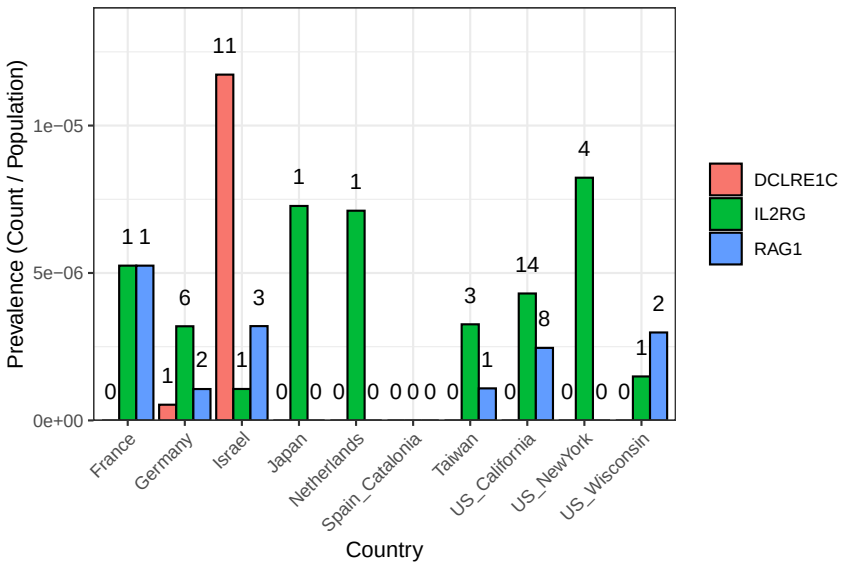


Figure S3: **SCID-specific gene comparison across regions.** The bar plot shows the prevalence of SCID-related cases (count divided by population) for each gene and country (or region), with numbers printed above the bars representing the actual counts in the original cohort (ranging from 0 to 11 per region and gene).

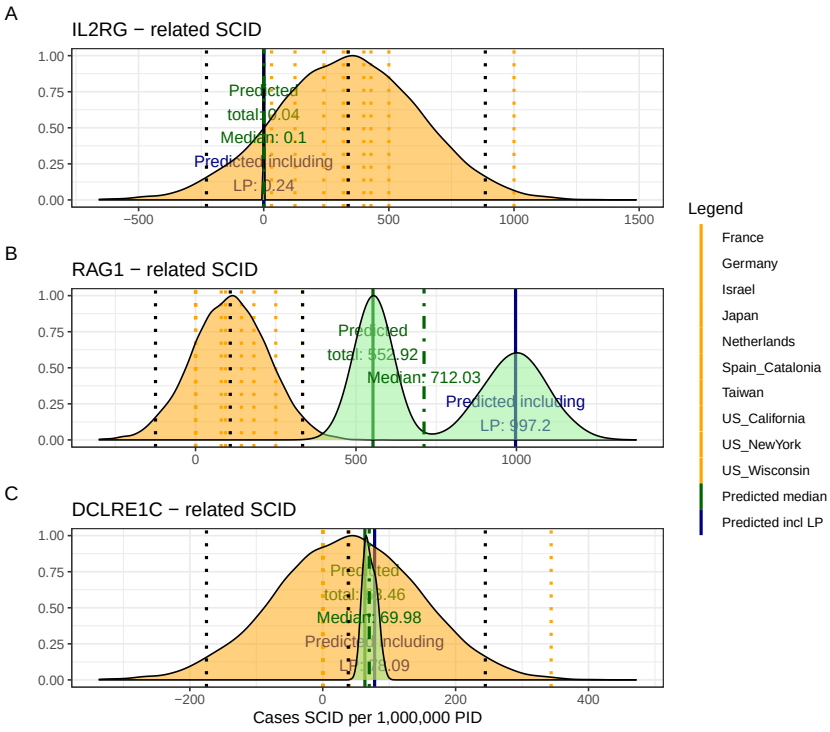


Figure S4: **Combined SCID-specific Predictions and Observed Rates per 1,000,000 PID.** The figure presents density distributions for the predicted SCID case counts (per 1,000,000 PID) for three genes: *IL2RG*, *RAG1*, and *DCLRE1C*. Country-specific rates (displayed as dotted vertical lines) are overlaid 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.

6.2 Hierarchical Clustering of Enrichment Scores for Major Disease Categories

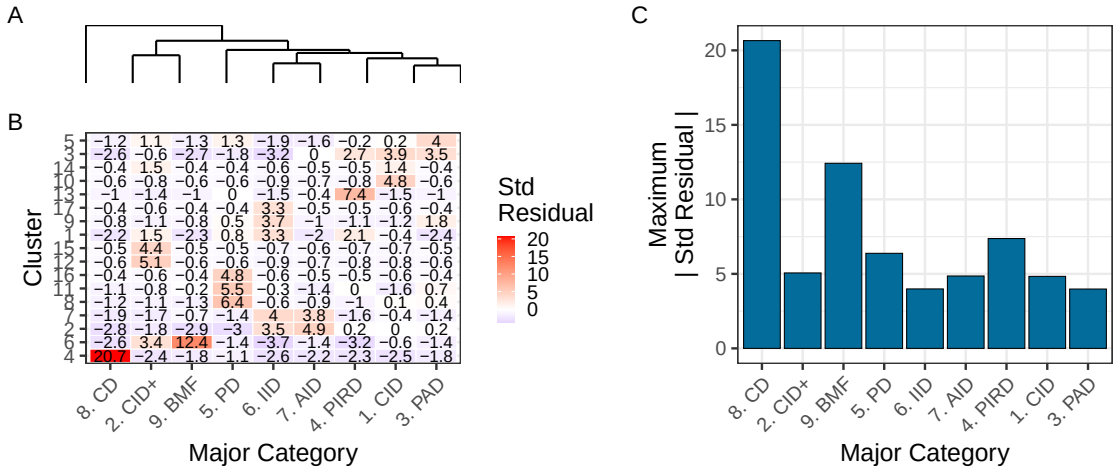


Figure S5: **Hierarchical clustering of enrichment scores.** The heatmap displays standardised residuals for major disease categories (x-axis) across network clusters (y-axis). A dendrogram groups similar disease categories, and the bar plot shows the maximum absolute residual per category. (8) CD and (9)BMF show the highest vales, indicating significant enrichment or depletion (residuals $> |2|$). Definitions in **Box 2.1**.

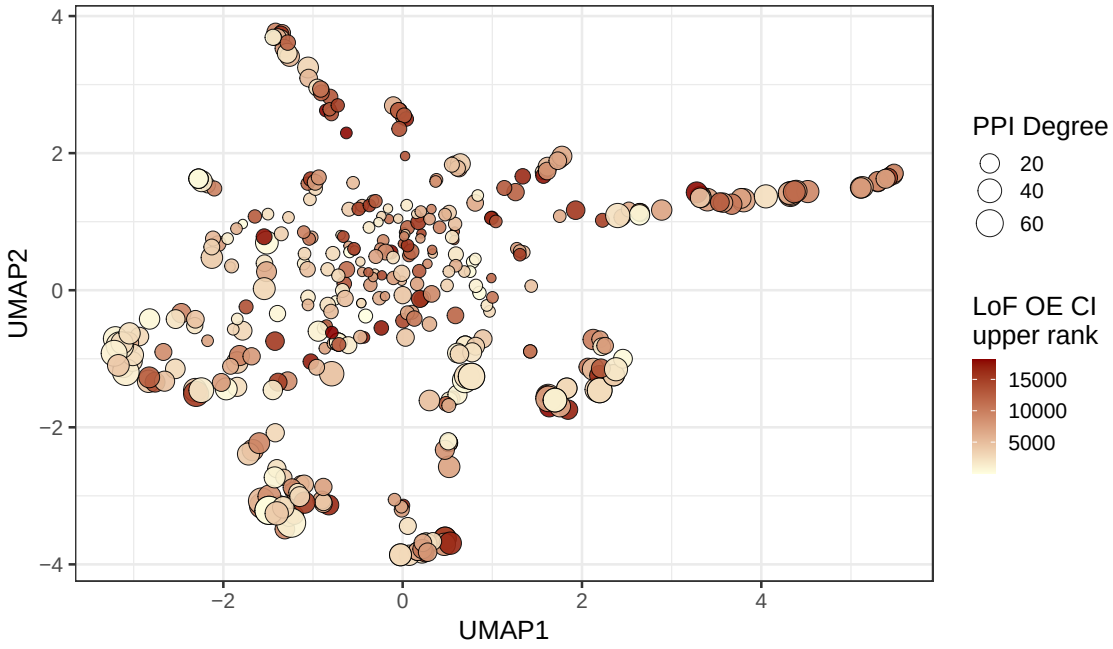


Figure S6: **Analysis of PPI degree versus LOEUF upper rank with UMAP embedding of the PPI network.** The relationship between PPI degree (size) and LOEUF upper rank (color) across gene clusters. No clear patterns are evident.

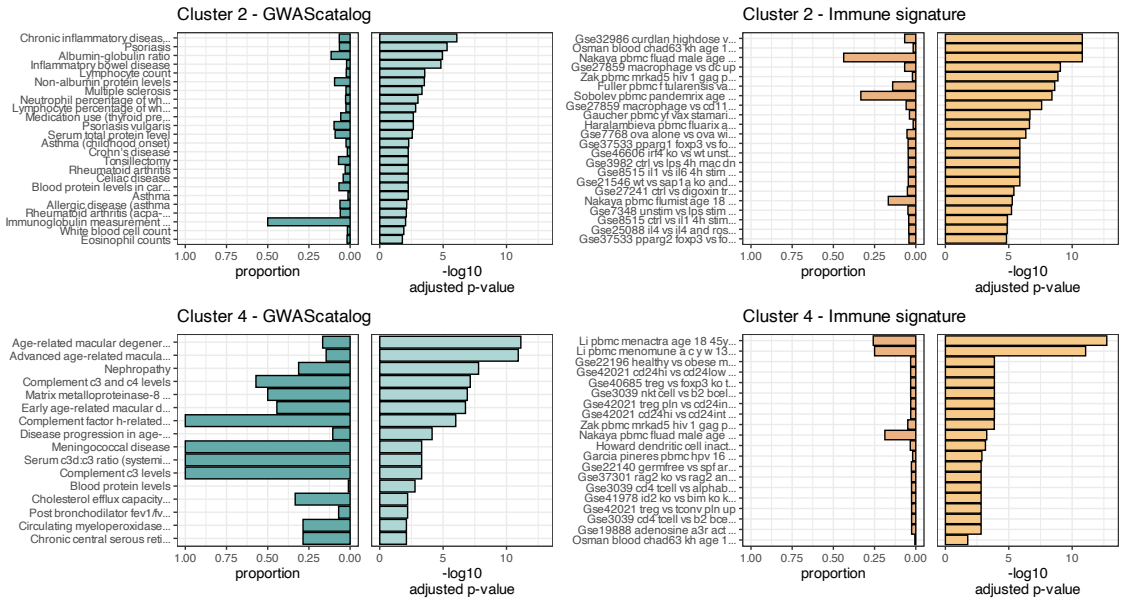


Figure S7: **Composite Enrichment Profiles for IEI Gene Sets.** We selected the top two enriched clusters (as per **Figure** 4) and performed functional enrichment analysis derived from known disease associations. For each gene set, the left panel displays the proportion of input genes overlapping with 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.

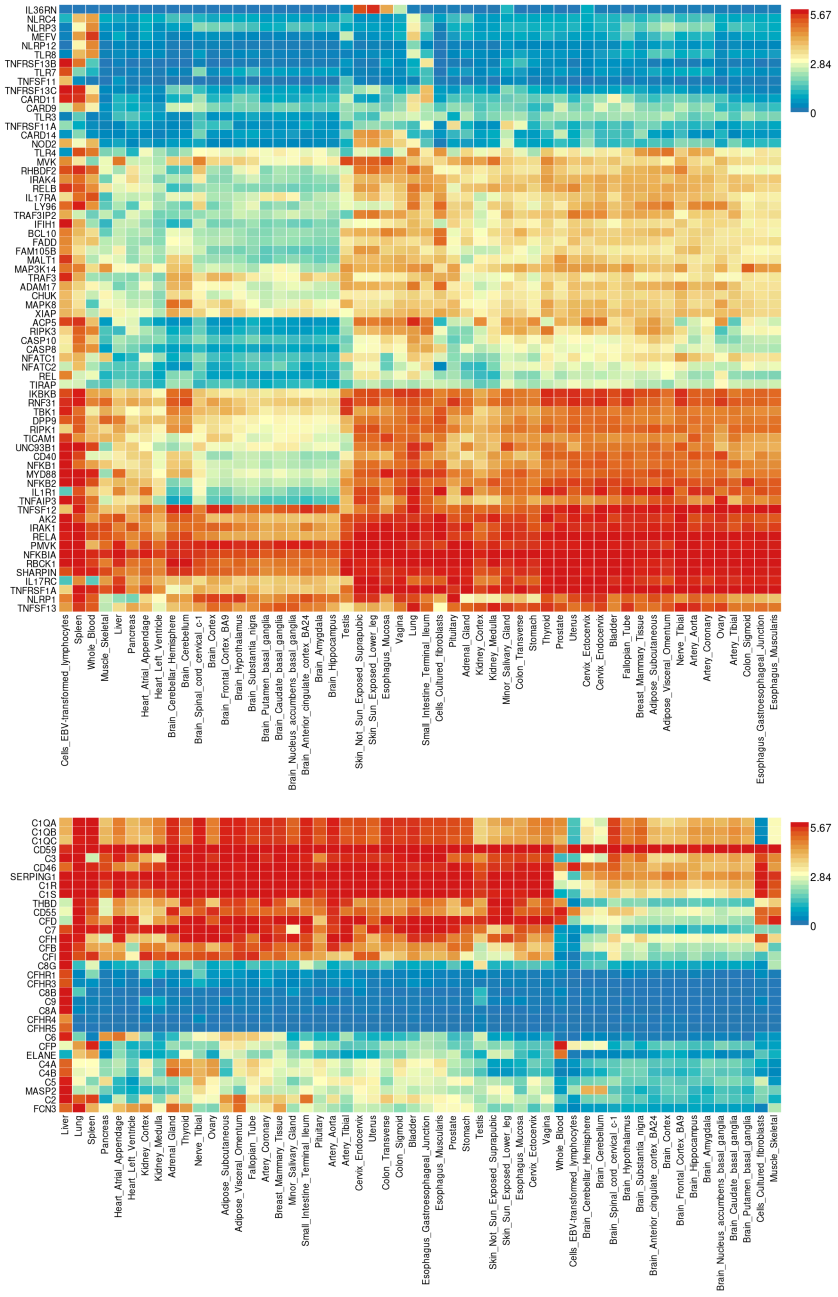


Figure S8: **Gene Expression Heatmaps for IEI Genes.** GTEx v8 data from 54 tissue types display the average expression per tissue label ($\log_2$ transformed) for the IEI gene panels. Top: Cluster 2; Bottom: Cluster 4.

6.3 Interpretation of ClinVar Variant Observations

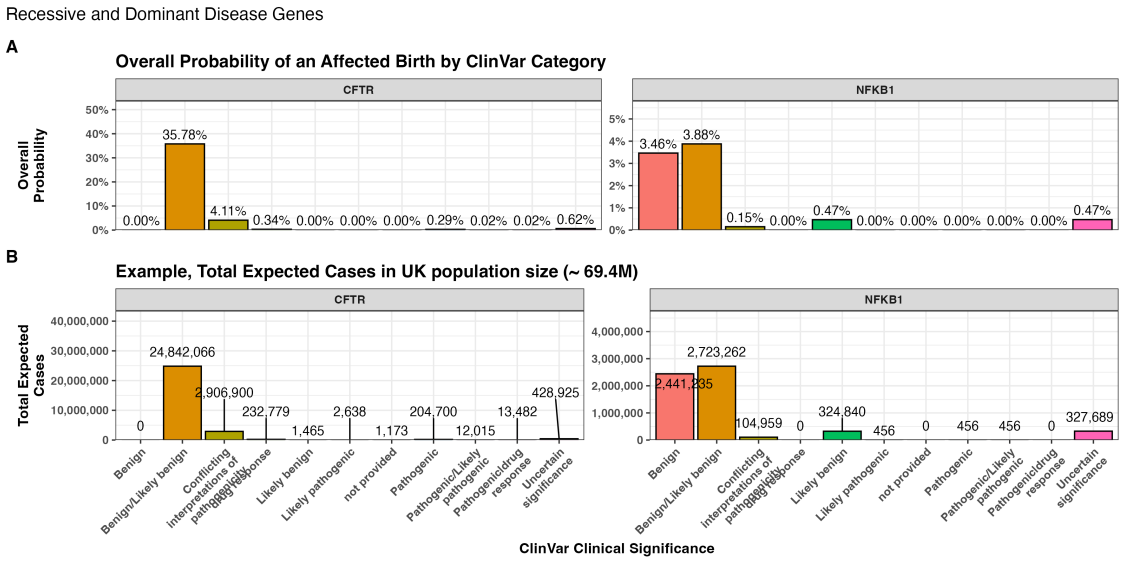


Figure S9: **Combined bar charts summarizing the genome-wide analysis of ClinVar clinical significance for the PID gene panel**. Panel (A) shows the overall probability of an affected birth by variant classification, and (B) displays the total expected number of cases per classification, both stratified by gene. These integrated results illustrate the variability in variant observations across genes and underpin our validation of the probability estimation framework.

6.4 Novel PID classifications

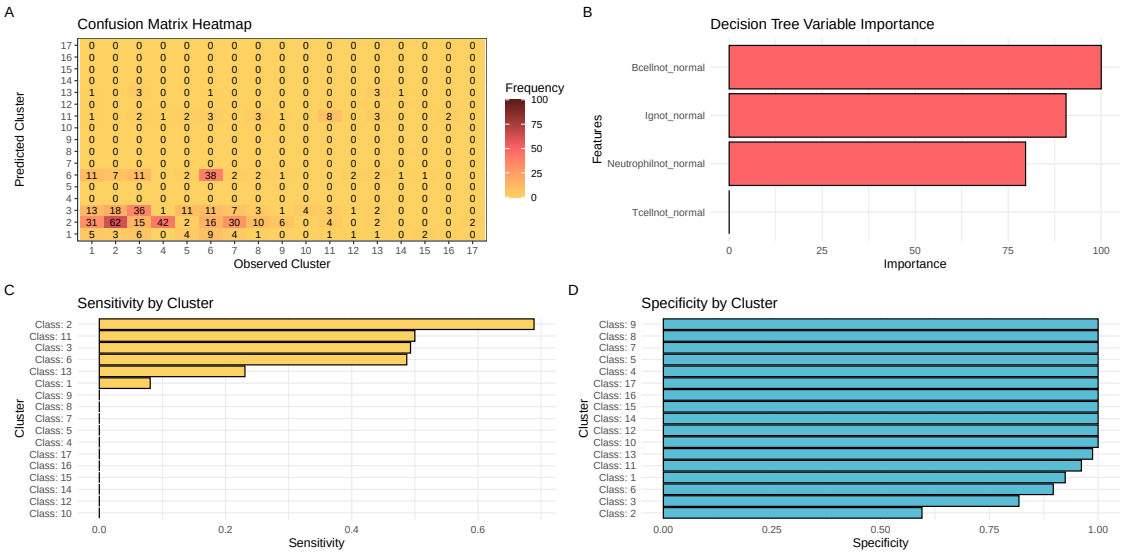


Figure S10: **Model performance for fine-tuned PID classification.** (A) Confusion matrix heatmap comparing observed and predicted PPI clusters. (B) Variable importance plot ranking immunophenotypic features contributing to the classifier. (C) Per-class sensitivity and (D) per-class specificity bar plots. These panels collectively demonstrate the performance of the decision tree classifier in stratifying PID genes based on immunophenotypic and PPI features.

6.5 Probability of observing AlphaMissense pathogenicity

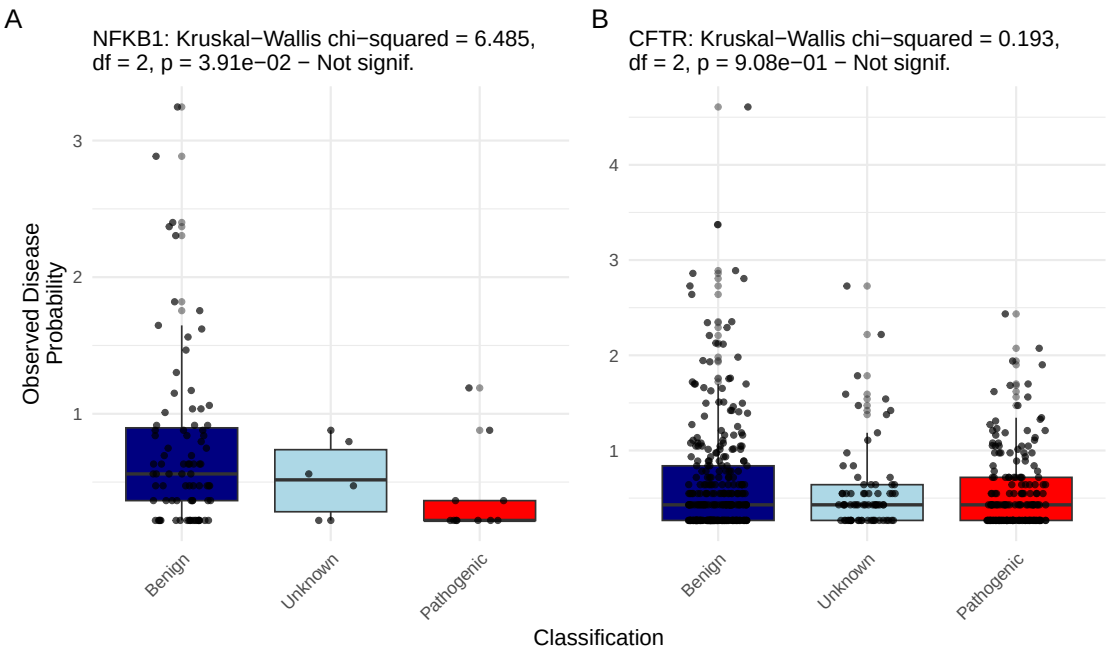


Figure S11: **Observed Disease Probability by Clinical Classification with AlphaMissense.** The figure displays the Kruskal-Wallis test results for NFKB1 and CFTR, showing no significant differences.

7 Clinical Genetics Application

In this section, we detail our approach to integrating sequencing data with prior pathogenicity evidence. Our method is designed to account for all possible outcomes of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), by first ensuring that every nucleotide corresponding to known pathogenic variants in a gene has been accurately sequenced. Only after confirming that these positions match the reference alleles (i.e. no unaccounted variant is present) do we calculate the probability that additional, alternative pathogenic variants (those not observed in the sequencing data) could be present. Our confidence interval (CI) for pathogenicity thus incorporates uncertainty from the entire process, including the tally of TP, FP, TN, and FN outcomes.

7.1 Methods

7.1.1 Quality Control:

Before performing any probability calculations, we inspect the gVCF to confirm that all known pathogenic variant positions in the gene are adequately covered and appear as reference alleles. This step not only verifies true negatives (TN) but also flags instances where sequencing quality is insufficient, leading to missing sequence information, and prevents false confidence. For example, if a nucleotide position corresponding to a known pathogenic variant has low quality reads and fails QC, it is flagged as missing, thereby affecting the overall probability estimate for unobserved variants.

7.1.2 Prior Probability Calculation:

For variants with an established ClinVar classification, the occurrence probability is derived directly from the allele frequency. For variants lacking a ClinVar label (i.e. variants of uncertain significance, VUS), we utilise an ACMG evidence score (0–100) to compute a prior probability as follows:

1. **Convert the ACMG Score:** The evidence score S is normalised to a fractional support level:
$$S_{\text{adj}} = \frac{S}{100}$$
This value reflects the strength of the pathogenic support.

2. **Assign a Minimal Risk (ϵ):** In the absence of a ClinVar classification, we assign a minimal risk based on the maximum observed allele number, $\max(AN)$, scaled by the evidence support:
$$\epsilon = \frac{1}{\max(AN) + 1} \times S_{\text{adj}}$$

This step ensures that even low-frequency variants receive a baseline risk proportional to the qualitative evidence.

3. **Adjust the Allele Frequency:** The observed allele frequency p_i is then increased by ϵ to yield an adjusted frequency:

$$p_i^{\text{adj}} = p_i + \epsilon$$

This adjusted frequency reflects both the empirical observation and the ACMG evidence.

4. **Calculate the Prior Probability of Disease:**

- For **Autosomal Dominant (AD)** or **X-Linked (XL)** inheritance, the prior probability is:

$$p_{\text{disease}} = p_i^{\text{adj}}$$

- For **Autosomal Recessive (AR)** inheritance—which considers both homozygosity and compound heterozygosity—the probability is calculated as:

$$p_{\text{disease}} = \left(p_i^{\text{adj}}\right)^2 + 2\,p_i^{\text{adj}}\left(P_{\text{tot}} - p_i^{\text{adj}}\right)$$

where

$$P_{\text{tot}} = \sum_{j \in \text{gene}} p_j^{\text{adj}}$$

7.1.3 Deriving the Confidence Interval (CI)

To capture uncertainty from all possible outcomes (TP, FP, TN, FN) in our sequencing and variant classification process, we propagate the variance arising from:

- The observed allele frequency and its adjustment via ϵ.
- The potential misclassification of variants (e.g. a VUS might be miscalled, contributing to FP or FN counts).
- Missing sequence data at known pathogenic sites.

We demonstrate two methods for deriving the 95% CI of the final occurrence probability: (1) the Wilson score interval and (2) a Bayesian credible interval using a Beta distribution.

1. Wilson Score Interval Assume the adjusted occurrence probability is estimated as $\hat{p} = p_i^{\text{adj}}$ based on an effective sample size N (which reflects the number of informative reads or quality-controlled observations). The Wilson score interval is computed as:

$$\hat{p}_W = \frac{\hat{p} + \frac{z^2}{2N}}{1 + \frac{z^2}{N}}$$

$$\text{Margin} = \frac{z}{1 + \frac{z^2}{N}} \sqrt{\frac{\hat{p}(1-\hat{p})}{N} + \frac{z^2}{4N^2}}$$

$$\text{CI}_{\text{Wilson}} = [\hat{p}_W - \text{Margin},\ \hat{p}_W + \text{Margin}]$$

where $z = 1.96$ for a 95% confidence level. This interval integrates uncertainty from the adjusted allele frequency and any variability in the count data.

2. Bayesian Credible Interval Alternatively, we can model the uncertainty using a Bayesian framework. Suppose that, after accounting for TP, FP, TN, and FN outcomes, the posterior distribution of the pathogenic probability is approximated by a Beta distribution, $\text{Beta}(\alpha, \beta)$. Here, the parameters α and β are chosen based on the effective counts of "successes" (e.g. detection or strong evidence of pathogenicity) and "failures" (e.g. absence or refutation), respectively. For example, if k is the effective number of positive events and $N - k$ the negatives, then:

$$\alpha = k + 1, \quad \beta = N - k + 1.$$

The 95% credible interval is then given by the 2.5th and 97.5th percentiles of the Beta distribution:

$$\text{CI}_{\text{Bayesian}} = \left[\text{BetaInv}(0.025; \alpha, \beta),\ \text{BetaInv}(0.975; \alpha, \beta)\right],$$

where $\text{BetaInv}(q; \alpha, \beta)$ denotes the quantile function of the Beta distribution at probability q.

Both methods integrate the uncertainty from the observed data, the adjustment via ϵ from the ACMG evidence score, and the potential misclassification or missing sequence data. In our analysis, the resulting 95% CI for pathogenicity is derived from such propagation of uncertainty, ensuring that all outcomes (TP, FP, TN, FN) are reflected in the final confidence bounds.

7.2 Results

We illustrate our method with two examples:

Example 1: Missing Sequence Information In one case, a known pathogenic nucleotide position in *GENE_XYZ* exhibited low quality reads and did not pass QC. This missing information prevents confirmation of the absence of the known variant (a potential false negative), thereby widening the uncertainty in our probability estimate. In such cases, the adjusted allele frequency is calculated with additional variance, leading to a broader CI. For instance, if the observed allele frequency is 1.0×10^{-5} and after adjusting with the ACMG score the estimated occurrence probability is 1.0×10^{-5}, the propagated uncertainty might yield a 95% CI of [0.70, 0.85]. This broader interval reflects the impact of missing sequence data on our confidence.

Example 2: Heterozygous Variant in an Autosomal Recessive Gene In another case, a patient carries a heterozygous variant in an autosomal recessive (AR) gene. In this scenario, there is also a second VUS in the same gene. Both variants are assessed using the ACMG evidence score adjustment. Their adjusted allele frequencies are used to compute the overall prior probability of disease, accounting for the possibility of compound heterozygosity. The two VUS are then ranked based on their evidence and the resulting 95% CIs. For instance, one variant may yield an occurrence probability of 2.5×10^{-4} with a 95% CI of [0.80, 0.88], while the other might have a lower probability of 1.8×10^{-4} with a CI of [0.75, 0.83]. The variant with the higher occurrence probability and narrower CI would be ranked as the more likely causal variant in the context of AR inheritance.

Table S1 shows the final variant results for a male patient carrying an X-linked loss-of-function (LOF) variant in *GENE_XYZ* where all known pathogenic positions were confirmed as reference alleles. For the variant c. 1234del (p.Glu412Argfs*5), the observed allele frequency is 1.2×10^{-5}. After applying the ACMG evidence score adjustment (for a VUS lacking a ClinVar classification), the adjusted allele frequency remains consistent with the observed data. The resulting occurrence probability is 1.2×10^{-5}, and by propagating the uncertainty from the allele frequency, evidence score adjustment, and the full range of possible outcomes (TP, FP, TN, FN), we derive a 95% CI for causality of [0.92, 0.97]. This confirms the variant as the top causal variant in this patient, with no evidence of additional alternative pathogenic variants.

Table S1: Final Variant Results for Patient (XL LOF)

Parameter	**Value**
Gene	*GENE_XYZ*
Variant	c. 1234del (p.Glu412Argfs*5)
Variant Type	Loss-of-Function (LOF)
Inheritance	X-Linked (XL)
Patient Sex	Male (hemizygous)
Allele Frequency	1.2×10^{-5}
Occurrence Probability	1.2×10^{-5}
95% CI for Causality	[0.92, 0.97]
Clinical Interpretation	Top causal variant confirmed; no evidence of additional alternative pathogenic variants