Quantifying the Genetics of Disease Inheritance in Primary Immunodeficiency

Dylan Lawless*[1]

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

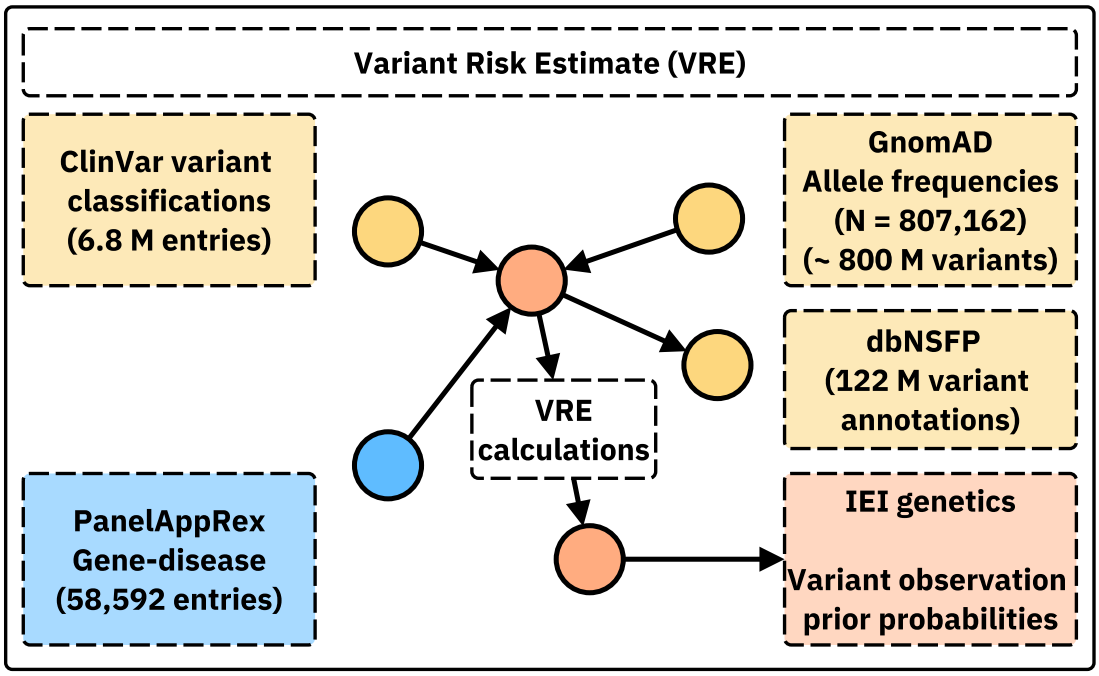
March 31, 2025

Abstract

We present an integrative framework for quantifying the prior probability of observing disease-associated variants in any gene for a given phenotype. By combining large-scale genomic annotation, —such as population allele frequencies and ClinVar variant classifications, with classical Hardy–Weinberg-based calculations, our method estimates observation probabilities for single nucleotide variants under autosomal dominant, autosomal recessive and X-linked inheritance. Applied to a panel of 557 genes implicated in primary immunodeficiency and monogenic inflammatory bowel disease, our approach generated 54,814 variant probabilities and revealed novel insights into gene constraint within functionally related pathways. Validation in autosomal dominant (*NFKB1*) and autosomal recessive (*CFTR*) disorders demonstrated close concordance between predicted and observed case counts. The resulting datasets, available in both machine- and human-readable formats via a Zenodo repository and a user-friendly web interface, provide robust priors to support Bayesian variant interpretation and improve 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: (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)
GnomAD Allele frequencies (N = 807,162) (~ 800 M variants)
dbNSFP (122 M variant annotations)
VRE calculations
PanelAppRex Gene-disease (58,592 entries)
IEI genetics
Variant observation prior probabilities

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, inheritance 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.

Quantifying the risk that a newborn inherits a disease-causing variant is a fundamental challenge in genomics. Classical statistical approaches grounded in HWE (1; 2) have long been used to calculate genetic mode of inheritance (MOI) probabilities for single nucleotide variants (SNVs). However, applying these methods becomes more complex when accounting for different modes of inheritance, 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 (3). 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.

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.

PanelAppRex is a novel tool developed to aggregate disease gene panel data from multiple sources-including GE PanelApp, ClinVar, and UniProt-and to facilitate sophisticated, natural language-based searches for clinical and research applications (4). It automatically retrieves and integrates gene panels, such as those used in the NHS National Genomic Test Directory and the 100,000 Genomes Project, into machine-readable formats that support rapid variant discovery and interpretation. By enabling queries based on gene names, phenotypes, and disease groups, PanelAppRex streamlines the process of identifying disease-associated variants and enhances the efficiency of genomic diagnostics. Critical details regarding the underlying data can be found in Martin et al. (5); Landrum et al. (6); The UniProt Consortium et al. (7).

Here, we incorporated datasets such as GnomAD, a large-scale resource that aggregates and harmonises sequencing data from diverse cohorts to provide a comprehensive view of human genetic variation (8). The v4 dataset comprises genomics data from 807,162 individuals. Beyond its extensive catalogue of variant calls, including over 786 million single nucleotide variants and 122 million InDels, the database offers detailed annotations that support analyses of population-specific allele frequencies.

dbNSFP is a comprehensive database designed for the functional prediction and annotation of all potential non-synonymous single-nucleotide variants (nsSNVs) and splicing-site SNVs in human protein-coding genes (9). It includes over 120 million variant entries and aggregates prediction scores from 33 different sources and allele frequencies from major population datasets.

ClinVar is a public archive of human genetic variation that provided detailed classifications-such as "Pathogenic," "Likely pathogenic," and "Benign" -accompanied by supporting evidence and review status (6). It aggregated submissions from multiple sources to present both consensus and conflicting interpretations, mapped variants to reference sequences according to HGVS standards, and collaborated with expert panels like ClinGen for ongoing re-evaluation.

We carry out validations for disease cohorts with *NFKB1* (10–13) and *CFTR* (14–16) to demonstrate autosomal dominant and recessive disease genes, respectively.

2 Methods

2.1 Dataset

Data from gnomAD v4 comprised 807,162 individuals, including 730,947 exomes and 76,215 genomes (8). This dataset provided 786,500,648 single nucleotide variants 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 (this version: 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 (6). For our analysis phase we also used dbNSFP which consisted of a number of annotations for 121,832,908 single nucleotide variants (9). 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. (4) A protein-protein interaction (PPI) network data was provided by STRINGdb, consisting of 19,566 proteins and 505,968 interactions (17).

2.2 Variant Class Observation Probability

As a starting point, we considered the classical Hardy–Weinberg equilibrium (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 autosomal recessive (AR) inheritance, 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 a variant 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 Inheritance. The probability that an individual was affected by a variant depended on the mode of inheritance relative to a specific phenotype. Specifically, we calculated the occurrence probability $p_{\text{disease},i}$ for each variant as follows:

- For **autosomal dominant (AD)** and **X-linked (XL)** variants, a single copy was sufficient, so

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

- For **autosomal recessive (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:

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

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 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 inheritance 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 autosomal dominant gene, we focused on *NFKB1*. Our goal was to compare the expected number of *NFKB1*-related 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. (10) to build a validation model in an autosomal dominant disease gene. A whole-genome sequencing study of 846 predominantly sporadic, unrelated primary immunodeficiency disease (PID) cases from the NIHR BioResource–Rare Diseases cohort identified *NFKB1* as one of the genes most strongly associated with PID. Sixteen novel heterozygous variants-including truncating, missense, and gene deletion variants-in *NFKB1* were found, accounting for 46% of common variable immunodeficiency (CVID) cases (n = 390) in the cohort.

Functional analyses, including structural protein evaluation, immunophenotyping, immunoblotting, and ex vivo lymphocyte stimulation, revealed that all carriers exhibited deficiencies in B-lymphocyte differentiation, particularly an increased CD21low B-cell population. These findings had established heterozygous loss-of-function variants in *NFKB1* as the most common monogenic cause of CVID, with significant prognostic implications.

2. Cohort Prevalence Calculation. Therefore, we used this UK-based cohort of 846 unrelated PID patients where 390 cases of CVID were attributed to *NFKB1*, yielding an observed cohort prevalence of

$$\text{Prevalence}_{\text{cohort}} = \frac{390}{846} \approx 0.461.$$

3. National Estimate Based on Literature. Based on literature, the prevalence of CVID in the general population was estimated at approximately 1/25 000 (10–13). For a UK population of $N_{\text{UK}} \approx 69\,433\,632$, the expected number of CVID cases was calculated as

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

Thus, the maximum expected number of *NFKB1*-related CVID cases in the entire population was estimated as

$$\text{Estimated } NFKB1 \text{ cases} \approx 2777 \times 0.461 \approx 1280,$$

with an approximate 95% confidence interval (derived from Wilson's method) of 1188 to 1374 cases.

4. Bayesian Adjustment. Given that the clinical cohort was derived from a specialized setting-likely capturing nearly all PID cases-the observed 390 cases may have better represented the true burden. To reconcile these perspectives, we performed a Bayesian adjustment by combining the known cohort data with the national estimate. Specifically, we computed a weighted average to symbolically acknowledge potential uncertainty:

$$\text{Adjusted Estimate} = w \cdot 390 + (1 - w) \cdot 1280,$$

with w set to 0.9 to reflect a strong preference for the observed data. Additionally, we modelled the uncertainty in the observed prevalence using a beta distribution:

$$p \sim \text{Beta}(390 + 1,\ 846 - 390 + 1),$$

and generated 10 000 posterior samples to obtain a density distribution for the adjusted estimate.

5. 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 validated our framework for estimating disease incidence in autosomal dominant disorders.

2.4 Validation Study for Autosomal Recessive CF Using CFTR

To validate our framework for autosomal recessive diseases, we focused on cystic fibrosis (CF). For comparability sizes between the validation studies, we analysed the most common single nucleotide variant (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 (14–16).

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 autosomal recessive trait, the expected frequencies were:

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

and

$f_{\text{comphet}} = 2p\,q$ (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. 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 autosomal dominant disorders.

2.5 Protein Network and Genetic Constraint Interpretation

A protein-protein interaction (PPI) network was constructed using protein interaction data from STRINGdb (17). 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 ANOVA was performed followed by Tukey HSD post hoc tests (and non-parametric Dunn's test for confirmation). Boxplots with jittered data points were used to depict the distribution of scores by category. GnomAD v4.1 constraint metrics data used for the PPI analysis was sourced from Karczewski et al. (8). This provided transcript-level metrics, such as observed/expected ratios, LOEUF, pLI, and Z-scores, quantifying loss-of-function and missense intolerance, along with confidence intervals and related annotations for 211,523 observations.

3 Results

3.1 Observation Probability Across Disease Genes

Our study demonstrated that by integrating large-scale annotation databases, such as gnomAD, ClinVar, (or alternatively dbNSFP), with disease gene panels from PanelAppRex, we could systematically scan every disease-gene panel according to inheritance mode. By combining reference population allele frequencies with reported ClinVar clinical significance classifications, we calculated an expected "observation probability" for each single nucleotide variant (SNV). This probability represented the likelihood of encountering a variant of a given pathogenicity class for a specific phenotype. We report the probability of disease observation for a total of 54,814 ClinVar variant classifications in 557 genes in the linked Zenodo repository (**to publish with DOI**).

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. A natural scaling system (-5 to +5) was applied to account for the frequently encountered combinations of classification labels (e.g. benign to pathogenic). The results summarized in **Table 1** illustrates that our method yielded estimations of the probability of observing a variant with a particular ClinVar classification. **Box 3.1** list the definitions for the major disease categories used throughout .

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.

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

Box 3.1 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

3.2 Validation of Dominant Disease Occurrence with *NFKB1*

To validate our genome-wide probability estimates for autosomal dominant disorders, we focused on *NFKB1*. We used a reference dataset from Tuijnenburg et al. (10), in which whole-genome sequencing of 846 PID patients identified *NFKB1* as one of the genes most strongly associated with the disease, with 390 CVID cases attributed to 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 456 *NFKB1*-related CVID cases in the UK. In the reference cohort, 390 *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 to yield an upper bound maximum of 1280 cases (95% CI: 1188–1374), while a Bayesian-adjusted mixture estimate produced a median of 835 cases (95% CI: 789–882). **Figure 2 (A)** illustrates that our predicted value of 456 lies

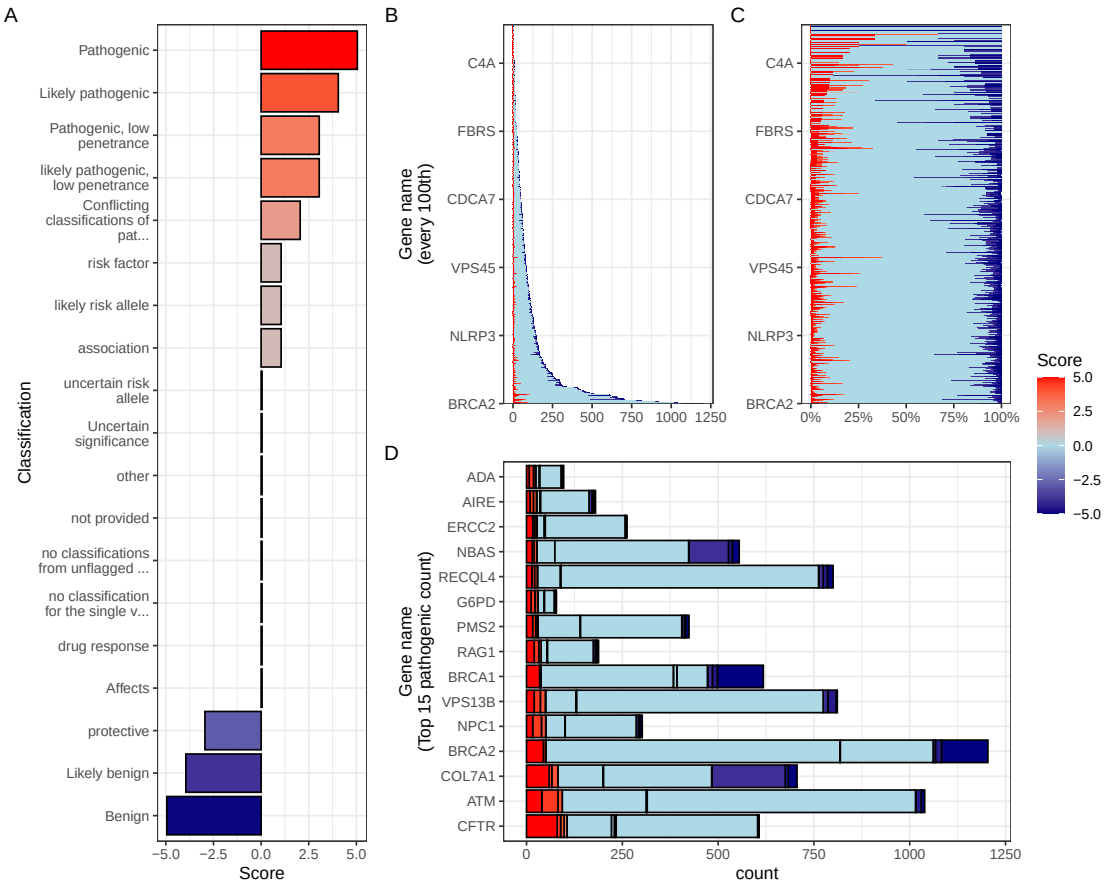


Figure 1: **Summary of ClinVar clinical significance classifications in the PID gene panel.** (A) Shows the numeric score coding for each classification. Panels (B) and (C) display the tally of classifications per gene as absolute counts and as percentages, respectively. (D) Highlights the top 15 genes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

within these ranges and is closer to the observed count, thereby supporting the validity of our integrated probability estimation framework for autosomal dominant disorders.

3.3 Validation of Recessive Disease Occurrence with *CFTR*

Our analysis predicted the number of cystic fibrosis (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 Hardy–Weinberg 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 (14). 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% CI: 696, 786) and a mixture-based estimate of 727 cases (95% CI: 705, 750).

Figure 2 (B) illustrates the close concordance between the predicted values, the Bayesian-adjusted estimates, and the national report supports the validity of our integrated approach for estimating disease incidence from population allele frequency data.

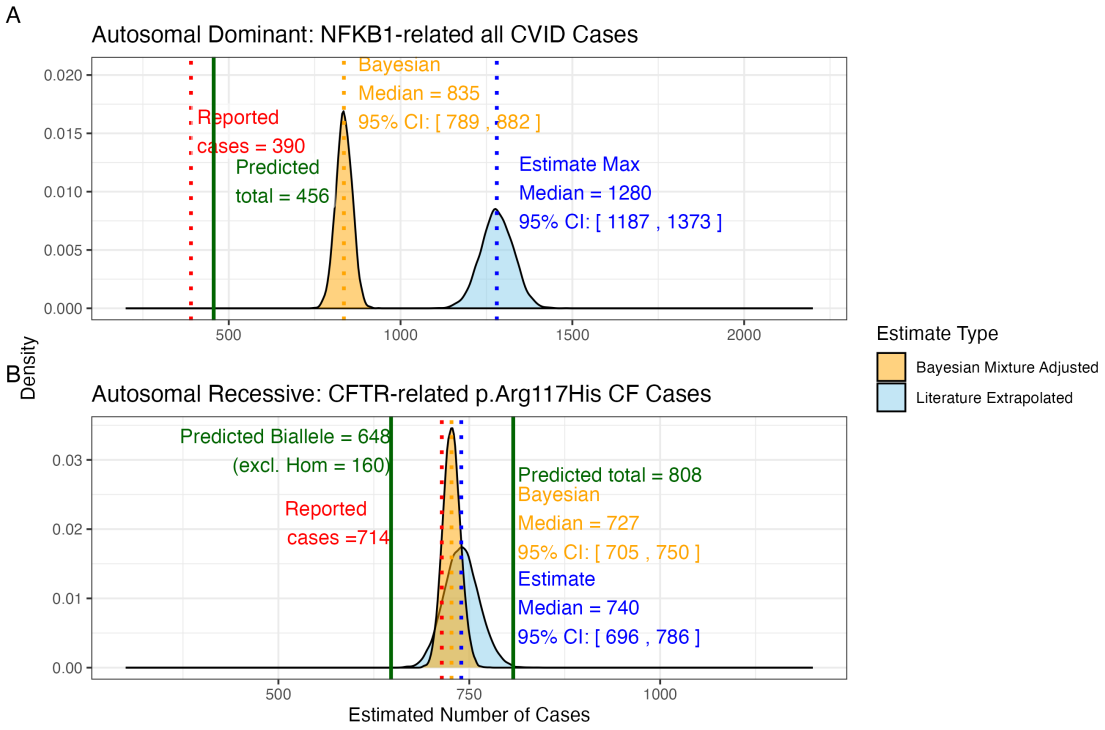


Figure 2: **Prior probabilities compared to validation disease cohort metrics.** (A) Density distributions for the number of *NFKB1*-related CVID cases in the UK. Our model (green) predicted 456 cases, which falls between the observed cohort count (red) of 390 and the upper extrapolated values. The blue 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.

The density-scatter plot (**Figure 3**) 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 autosomal recessive inheritance pattern of *CFTR*, this threshold translates into more than 100,000 heterozygous carriers, compared to only 456 carriers for the autosomal dominant 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 inheritance patterns on population carrier frequencies and the observed disease prevalence.

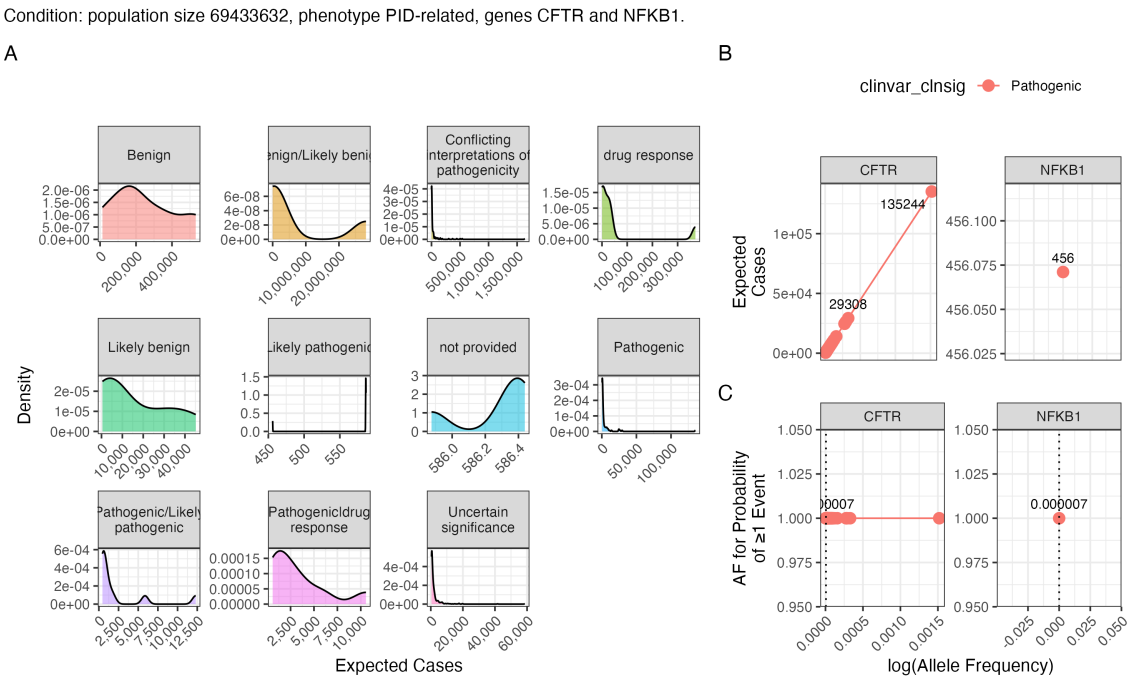


Figure 3: **Interpretation of probability of observing a variant classification.** The result from the chosen validation genes *CFTR* and *NFKB1* are shown. Case counts are dependant on the population size and phenotype. (A) The density plots of expected observations by ClinVar clinical significance. 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.

3.4 Interpretation of ClinVar Variant Observations

Figure 4 shows the two validation study PID genes, representing autosomal recessive and dominant inheritance. **Figure** 4 **(A)** illustrates the overall probability of an affected birth by ClinVar variant classification, whereas **Figure** 4 **(B)** depicts the total expected number of cases per classification for an example population, here the UK, of approximately 69.4 million.

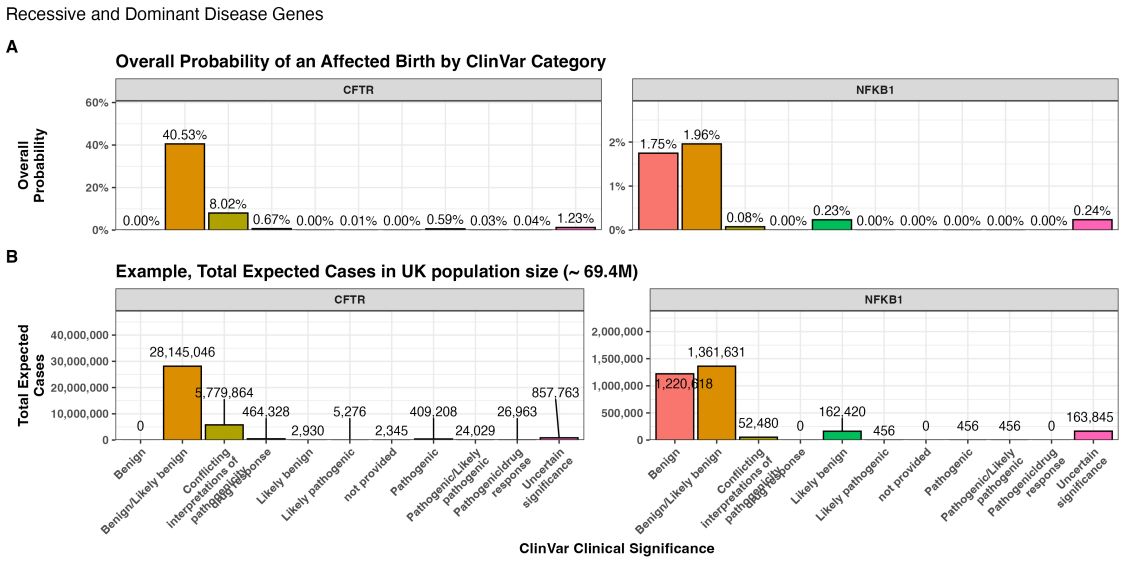


Figure 4: **Combined bar charts summarizing the genome-wide analysis of ClinVar clinical significance for the PID gene panel**. Panel (A) shows the overall probability of an affected birth by variant classification, and (B) displays the total expected number of cases per classification, both stratified by gene. These integrated results illustrate the variability in variant observations across genes and underpin our validation of the probability estimation framework.

3.5 Genetic constraint in high-impact protein networks

3.5.1 Overview of the PPI network and hub gene identification

The ClinVar classifications reported in Figure 1 are scale -5 to +5 based on their pathogenicity. We were interested in positive (damaging) but not negative (benign) scoring variants, which are incidental to subsequent analysis. We tallied gene-level positive scores to give the score positive total metric. **Figure 5 (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.

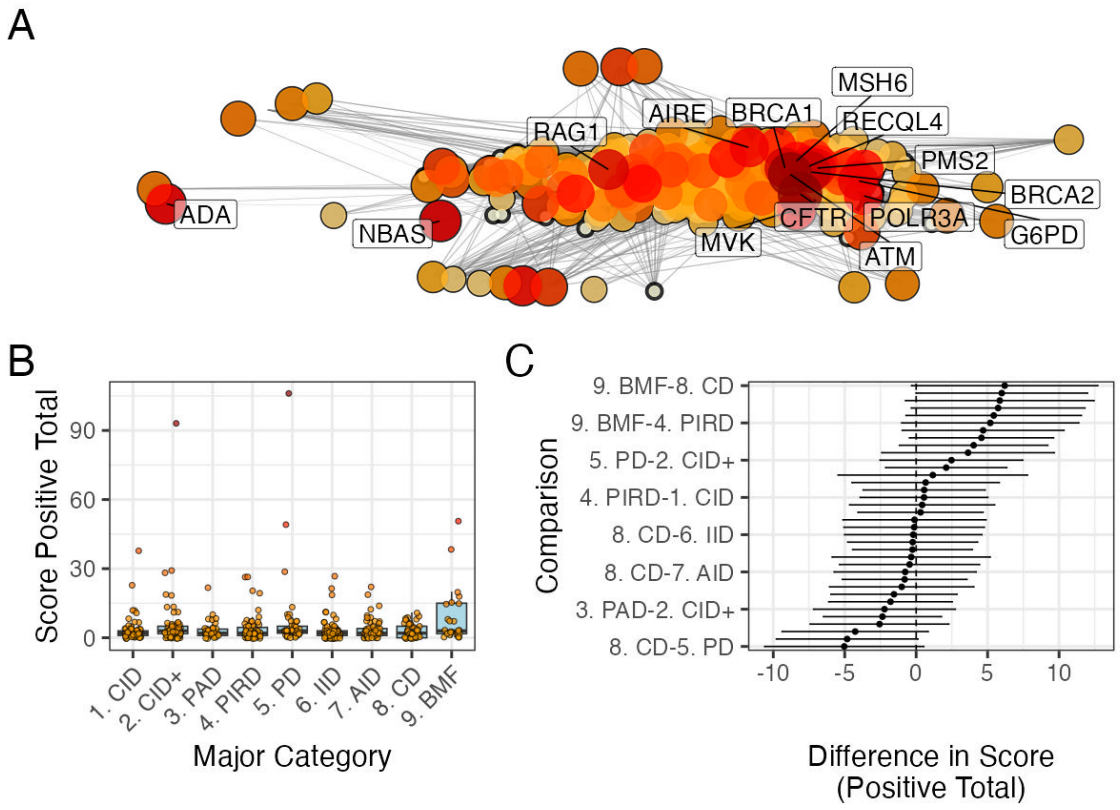


Figure 5: **PPI network and score positive total ClinVar significance variants**. (A) PPI network of disease-associated genes. Node size and colour represent the log-transformed `score_positive_total`, the top 15 genes/proteins with the highest probability of being observed in disease are labelled. (B) 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.

3.5.2 Association Analysis of Score-Positive-Total across PID 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 are not identical, which we observed in **Figure 5 (B)**. However, despite some apparent differences in median scores across categories (i.e. 9. BMF), the Tukey honest significance test (HSD) post hoc comparisons **Figure 5 (C)** showed that all pairwise differences had 95% confidence intervals overlapping zero, suggesting that individual group differences were not significant.

3.5.3 UMAP Embedding of the Protein-Protein Interaction Network

To address the density of the PPI network for the IEI gene panel, we applied UMAP (**Figure 6**). The embedding automatically defined gene clusters. Node sizes reflect interaction degree, a measure of evidence-supported connectivity. We tested for a correlation between interaction degree and `score_positive_total`. In **Figure 6**, nodes with degrees above the 95th percentile are labelled in blue, while the top 15 genes by `score_positive_total` are labelled in yellow. Notably, genes with high pathogenic variant loads segregate from highly connected nodes, suggesting that loss-of-function in hub genes is selectively constrained, whereas damaging variants in lower-degree genes yield more specific effects. This observation was subsequently tested empirically.

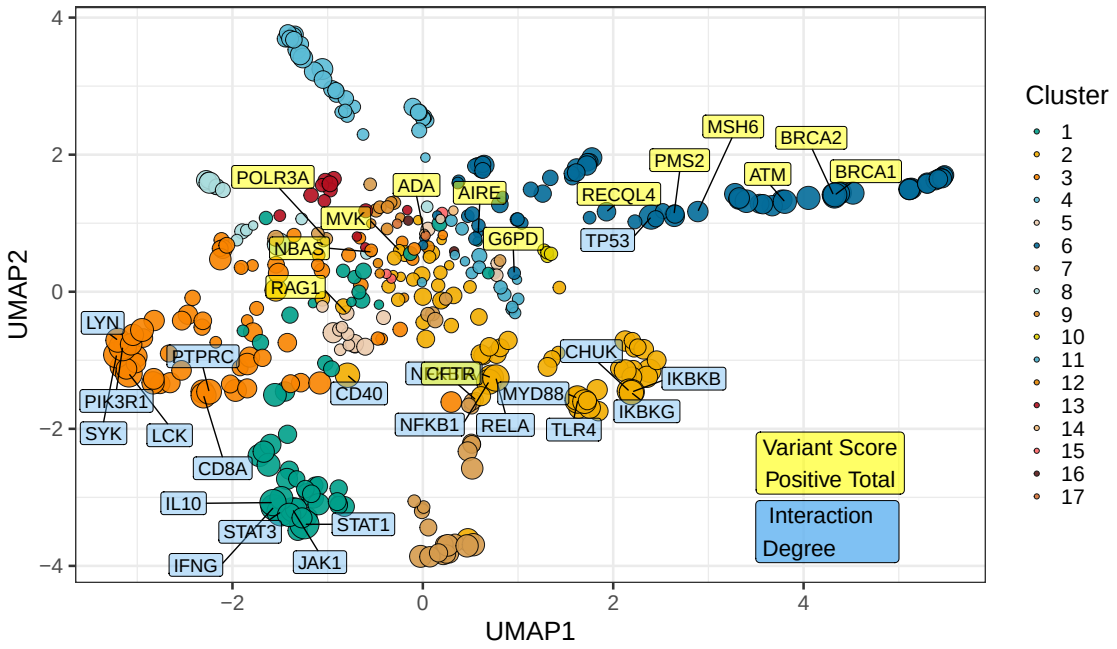


Figure 6: **UMAP embedding of the PPI network (p_umap).** The plot projects the high-dimensional protein-protein interaction network into two dimensions, with nodes coloured by cluster and sized by interaction degree. Blue labels indicate hub genes (degree above the 95th percentile) and yellow labels 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.

3.5.4 Hierarchical Clustering of Enrichment Scores for Major Disease Categories

Figure S1 presents a heatmap of standardised residuals for major disease categories across network clusters, as per **Figure 6**. A dendrogram clusters similar disease cate-

gories, 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) Bone marrow failure (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).

3.5.5 PPI Connectivity, LOEUF Constraint and Enriched Network Cluster Analysis

We evaluated the relationship between network connectivity (PPI degree) and loss-of-function constraint (LOEUF upper rank) using Spearman's rank correlation (see **Figure 7**). Overall, there is a weak but significant negative correlation (= –0.181, p = 0.00024), indicating that highly connected genes tend to be more constrained. However, when stratified by gene clusters, some clusters (e.g. clusters 2, 3, 4, and 5) show moderate to strong correlations, whereas others (clusters 1, 6, 7, and 8) do not. These findings suggest that constraint on loss-of-function variation varies among functional gene groups.

We assessed the correlation between network connectivity and LOEUF constraint. A supplementary analysis (**Figure S2**) did not reveal distinct visual associations between network clusters and constraint metrics, possibly due to the high network density. We therefore tested for correlation quantitatively. As shown in **Figure 7(A)**, there was a weak but significant negative Spearman correlation (= –0.181, p = 0.00024) between PPI degree and LOEUF upper rank, suggesting that highly connected genes are more constrained. Notably, when stratified by clusters, the correlation is stronger in specific groups, clusters 2 (= –0.375, p = 0.000994), 4 (= –0.800, p $<$ 0.000001), indicating that shared mechanisms within pathway clusters may accentuate genetic constraints.

Figure 7 (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.

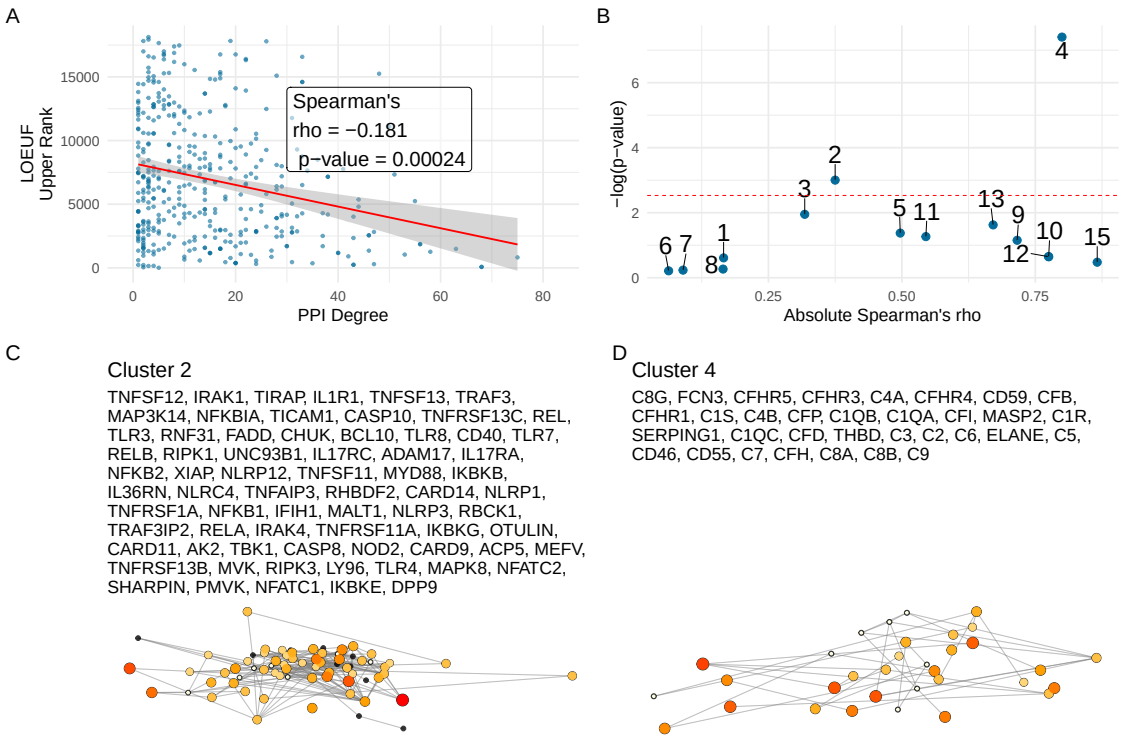


Figure 7: **Correlation between PPI degree and LOEUF upper rank. (A)** Ananlysis across all genes revealed a weak, significant negative correlation (= –0.181, p = 0.00024) between PPI degree and LOEUF upper rank. **(B)** The cluster-wise analysis showed that clusters 2 (= –0.375, p = 0.000994) and 4 (= –0.800, p < 0.000001) exhibit 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.

3.5.6 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, across autosomal dominant, autosomal recessive, and X-linked modes of inheritance, 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 Zenodo repository, and a user-friendly web interface that incorporates these new metrics.

Figure 8 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, OMIM, AlphaFold).

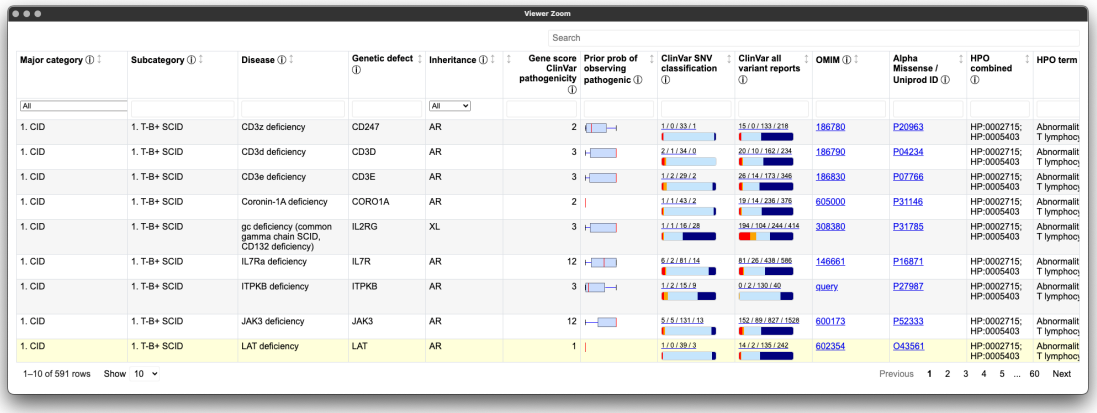


Figure 8: **Integration of variant probabilities into the IEI genetics framework.** The interface summarises the condensed variant data, with pre-calculated summary statistics and dynamic links to external databases. This integration enables immediate access to detailed variant classifications and prior probabilities for each gene.

4 Discussion

Our study presents, to our knowledge, the first comprehensive framework for calculating prior probabilities of observing disease-associated variants across all modes of inheritance for any gene–disease combination. By integrating large-scale genomic annotations, including population allele frequencies from gnomAD (8), variant classifications from ClinVar (6), and testing against 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 genes implicated in primary immunodeficiency and monogenic inflammatory bowel disease (4).

Our genome-wide approach accommodates the distinct inheritance patterns of autosomal dominant (AD), autosomal recessive (AR), and X-linked (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 inheritance complexities and enhances clinicians' understanding of expected variant occurrences, thereby improving diagnostic precision.

Moreover, our method complements existing statistical approaches for aggregating variant effects (e.g., ACAT and SKAT (18–21)) and multi-omics integration techniques (22; 23), while remaining consistent with established variant interpretation guidelines from the ACMG (24) and complementary frameworks (25; 26), as well as quality control protocols (27; 28). Standardised reporting for qualifying variant sets,

such as ACMG Secondary Findings v3.2 (29), further contextualises the integration of these probabilities into clinical decision-making.

We acknowledge that our current approach is limited to single nucleotide variants (SNVs) and does not account for other complexities of genetic disease, including complex variants, de novo events, hypomorphic phenotypes, overdominance, penetrance, tissue-specific transcription, and pleiotropy (3). 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 (6). 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 (8) 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.

Competing interest

We declare no competing interest.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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

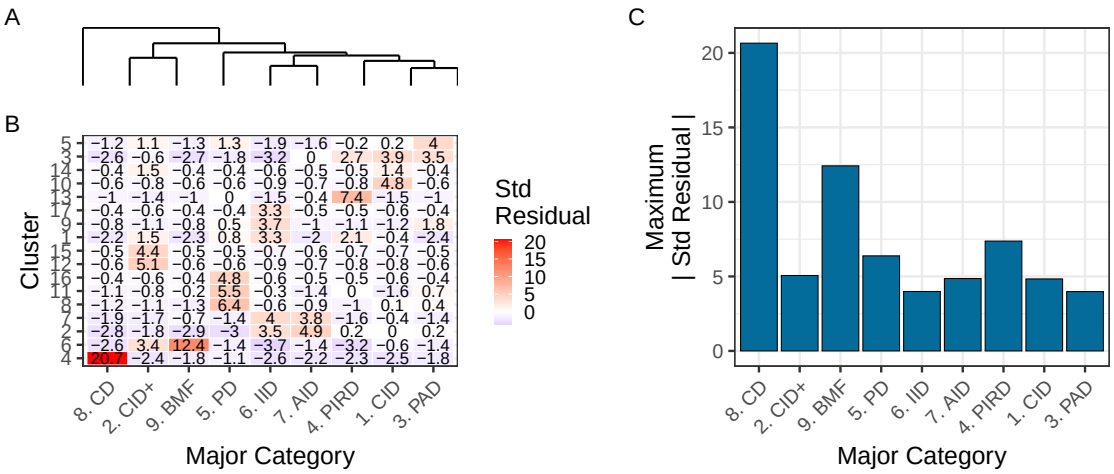


Figure S1: Hierarchical clustering of enrichment scores. The heatmap displays standardised residuals for major disease categories (x-axis) across network clusters (y-axis). A dendrogram groups similar disease categories, and the bar plot shows the maximum absolute residual per category. Cet 8.CD and 9, BMF show the highest values, indicating significant enrichment or depletion (residuals $> |2|$). Definitions in **Box 3.1**.

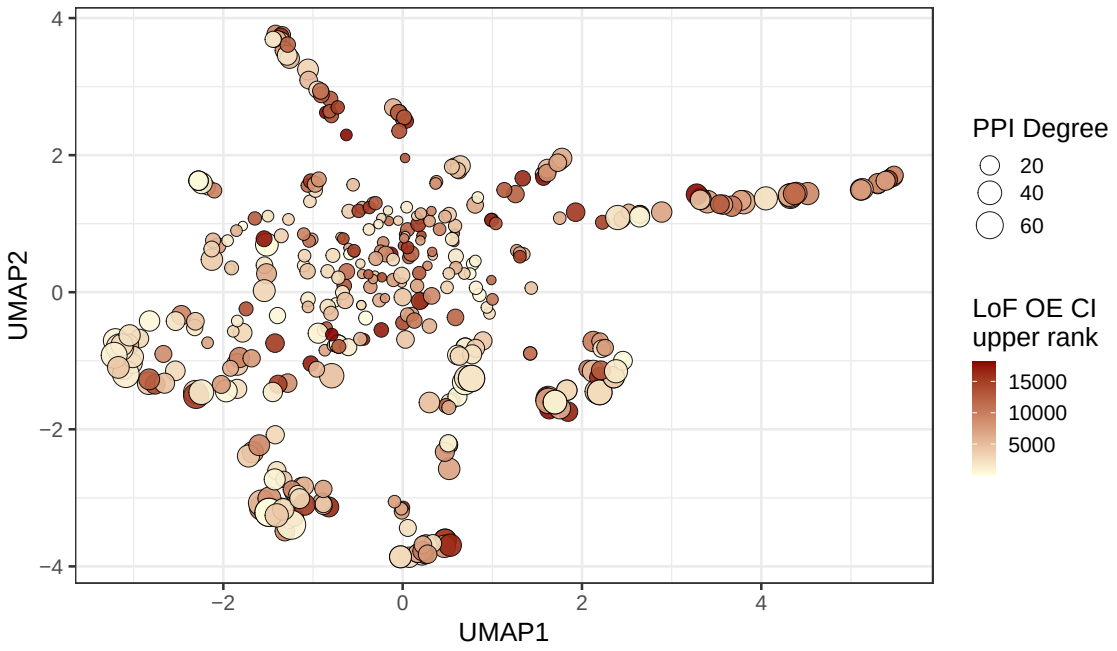


Figure S2: Supplementary analysis of PPI degree versus LOEUF upper rank with UMAP embedding of the PPI network. The relationship between PPI degree (size) and LOEUF upper rank (color) across gene clusters. No clear patterns are evident.