

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

Dylan Lawless<sup>\*1</sup>

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

April 18, 2025

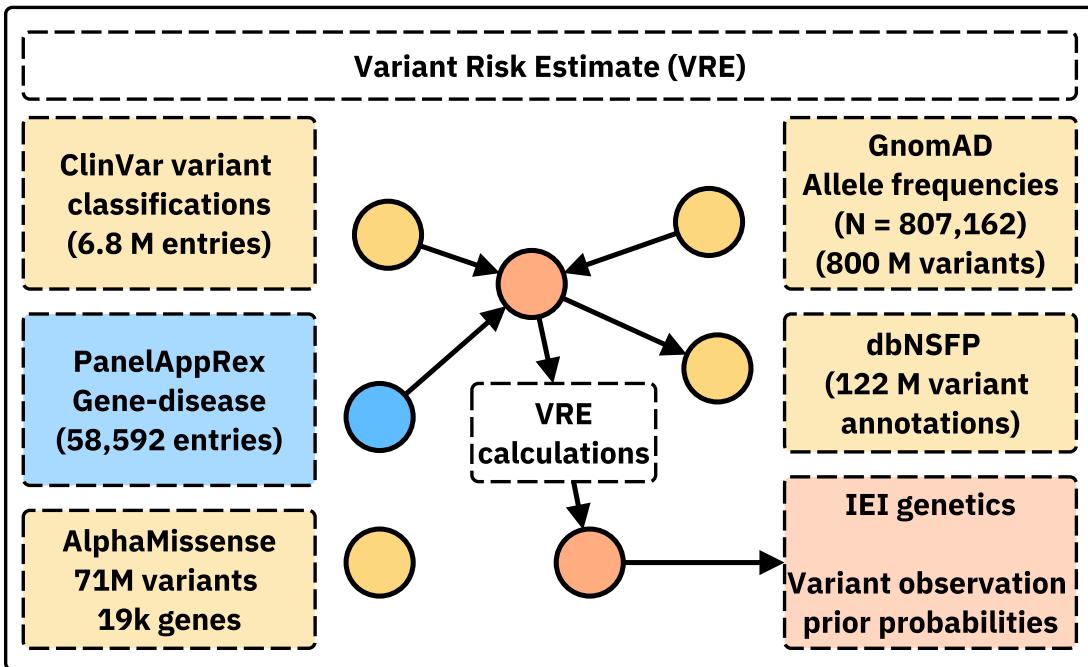
## Abstract

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.<sup>1</sup>

---

<sup>\*</sup>Addresses for correspondence: [Dylan.Lawless@kispi.uzh.ch](mailto:Dylan.Lawless@kispi.uzh.ch)

<sup>1</sup> **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](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.



18

## <sup>19</sup> Acronyms

<sup>20</sup> <b>ACMG</b> American College of Medical Genetics and Genomics.....	<sup>30</sup>
<sup>21</sup> <b>ACAT</b> Aggregated Cauchy Association Test .....	<sup>30</sup>
<sup>22</sup> <b>AD</b> Autosomal Dominant.....	<sup>4</sup>
<sup>23</sup> <b>ANOVA</b> Analysis of Variance .....	<sup>13</sup>
<sup>24</sup> <b>AR</b> Autosomal Recessive .....	<sup>4</sup>
<sup>25</sup> <b>BMF</b> Bone Marrow Failure.....	<sup>18</sup>
<sup>26</sup> <b>CD</b> Complement Deficiencies .....	<sup>20</sup>
<sup>27</sup> <b>CI</b> Confidence Interval.....	<sup>16</sup>
<sup>28</sup> <b>CF</b> Cystic Fibrosis .....	<sup>10</sup>
<sup>29</sup> <b>CFTR</b> Cystic Fibrosis Transmembrane Conductance Regulator.....	<sup>5</sup>
<sup>30</sup> <b>CVID</b> Common Variable Immunodeficiency .....	<sup>8</sup>
<sup>31</sup> <b>dbNSFP</b> database for Non-Synonymous Functional Predictions .....	<sup>5</sup>
<sup>32</sup> <b>GE</b> Genomics England .....	<sup>5</sup>
<sup>33</sup> <b>gnomAD</b> Genome Aggregation Database .....	<sup>5</sup>
<sup>34</sup> <b>HGVS</b> Human Genome Variation Society.....	<sup>5</sup>
<sup>35</sup> <b>HPC</b> High-Performance Computing.....	<sup>8</sup>
<sup>36</sup> <b>HWE</b> Hardy-Weinberg Equilibrium .....	<sup>4</sup>
<sup>37</sup> <b>IEI</b> Inborn Errors of Immunity.....	<sup>4</sup>
<sup>38</sup> <b>Ig</b> Immunoglobulin .....	<sup>23</sup>
<sup>39</sup> <b>InDel</b> Insertion/Deletion .....	<sup>5</sup>
<sup>40</sup> <b>IUIS</b> International Union of Immunological Societies .....	<sup>6</sup>
<sup>41</sup> <b>LD</b> Linkage Disequilibrium .....	<sup>22</sup>
<sup>42</sup> <b>LOEUF</b> Loss-Of-function Observed/Expected Upper bound Fraction .....	<sup>13</sup>
<sup>43</sup> <b>LOF</b> Loss-of-Function .....	<sup>19</sup>
<sup>44</sup> <b>MOI</b> Mode of Inheritance .....	<sup>4</sup>
<sup>45</sup> <b>NFKB1</b> Nuclear Factor Kappa B Subunit 1 .....	<sup>5</sup>
<sup>46</sup> <b>OMIM</b> Online Mendelian Inheritance in Man .....	<sup>27</sup>
<sup>47</sup> <b>PID</b> Primary Immunodeficiency .....	<sup>4</sup>
<sup>48</sup> <b>PPI</b> Protein-Protein Interaction .....	<sup>5</sup>
<sup>49</sup> <b>SNV</b> Single Nucleotide Variant .....	<sup>4</sup>
<sup>50</sup> <b>SKAT</b> Sequence Kernel Association Test.....	<sup>30</sup>
<sup>51</sup> <b>STRINGdb</b> Search Tool for the Retrieval of Interacting Genes/Proteins.....	<sup>5</sup>
<sup>52</sup> <b>HSD</b> Honestly Significant Difference .....	<sup>13</sup>
<sup>53</sup> <b>UMAP</b> Uniform Manifold Approximation and Projection .....	<sup>19</sup>
<sup>54</sup> <b>UniProt</b> Universal Protein Resource.....	<sup>5</sup>
<sup>55</sup> <b>VEP</b> Variant Effect Predictor.....	<sup>5</sup>
<sup>56</sup> <b>XL</b> X-Linked .....	<sup>4</sup>

## 94 1 Introduction

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

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

108 Quantifying the risk that a newborn inherits a disease-causing variant is a fun-  
 109 damental challenge in genomics. Classical statistical approaches grounded in HWE  
 110 (4; 5) have long been used to calculate genetic MOI probabilities for Single Nucleotide  
 111 Variant (SNV)s. However, applying these methods becomes more complex when ac-  
 112 counting for different MOI, such as Autosomal Recessive (AR) versus Autosomal  
 113 Dominant (AD) or X-Linked (XL) disorders. In AR conditions, for example, the  
 114 occurrence probability must incorporate both the homozygous state and compound  
 115 heterozygosity, whereas for AD and XL disorders, a single pathogenic allele is suffi-  
 116 cient to cause disease. Advances in genetic research have revealed that MOI can be  
 117 even more complex (6). Mechanisms such as dominant negative effects, haploinsuffi-  
 118 ciency, mosaicism, and digenic or epistatic interactions can further modulate disease  
 119 risk and clinical presentation, underscoring the need for nuanced approaches in risk  
 120 estimation. Karczewski et al. (7) made significant advances; however, the remain-  
 121 ing challenge lay in applying the necessary statistical genomics data across all MOI  
 122 for any gene-disease combination. Similar approaches have been reported for disease  
 123 such Wilson disease, Mucopolysaccharidoses, Primary ciliary dyskinesia, and treat-  
 124 able metabolic diseases, (8; 9), as reviewed by Hannah et al. (10).

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

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

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

## 151 2 Methods

### 152 2.1 Dataset

153 Data from Genome Aggregation Database (gnomAD) v4 comprised 807,162 indi-  
154 viduals, including 730,947 exomes and 76,215 genomes (7). This dataset provided  
155 786,500,648 SNVs and 122,583,462 InDels, with variant type counts of 9,643,254 syn-  
156 onymous, 16,412,219 missense, 726,924 nonsense, 1,186,588 frameshift and 542,514  
157 canonical splice site variants. ClinVar data were obtained from the variant summary  
158 dataset (as of: 16 March 2025) available from the NCBI FTP site, and included  
159 6,845,091 entries, which were processed into 91,319 gene classification groups and a  
160 total of 38,983 gene classifications; for example, the gene *A1BG* contained four vari-  
161 ants classified as likely benign and 102 total entries (13). For our analysis phase  
162 we also used dbNSFP which consisted of a number of annotations for 121,832,908  
163 SNVs (15). The PanelAppRex core model contained 58,592 entries consisting of  
164 52 sets of annotations, including the gene name, disease-gene panel ID, diseases-  
165 related features, confidence measurements. (2) A Protein-Protein Interaction (PPI)  
166 network data was provided by Search Tool for the Retrieval of Interacting Genes/Pro-  
167 teins (STRINGdb), consisting of 19,566 proteins and 505,968 interactions (16). The  
168 Human Genome Variation Society (HGVS) nomenclature is used with Variant Effect  
169 Predictor (VEP)-based codes for variant IDs. We carried out validations for disease  
170 cohorts with Nuclear Factor Kappa B Subunit 1 (*NFKB1*) (17–20) and Cystic Fibrosis  
171 Transmembrane Conductance Regulator (*CFTR*) (21–23) to demonstrate applications

<sup>172</sup> in AD and AR disease genes, respectively. AlphaMissense includes pathogenicity pre-  
<sup>173</sup> diction classifications for 71 million variants in 19 thousand human genes (12; 26).  
<sup>174</sup> We used these scores to compare against the probability of observing the same given  
<sup>175</sup> variants. **Box 2.1** list the definitions from the International Union of Immunological  
<sup>176</sup> 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

<sup>177</sup>

## <sup>178</sup> 2.2 Variant Class Observation Probability

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

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

<sup>179</sup> where  $p$  is the allele frequency,  $q = 1 - p$ ,  $p^2$  represents the homozygous dominant,  
<sup>180</sup>  $2pq$  the heterozygous, and  $q^2$  the homozygous recessive genotype frequencies. For  
<sup>181</sup> disease phenotypes, particularly under AR MOI, the risk is traditionally linked to  
<sup>182</sup> the homozygous state ( $p^2$ ); however, to account for compound heterozygosity across  
<sup>183</sup> multiple variants, we allocated the overall gene-level risk proportionally among vari-  
<sup>184</sup> ants.

<sup>185</sup> Our computational pipeline estimated the probability of observing a disease-associated  
<sup>186</sup> genotype for each variant and aggregated these probabilities by gene and ClinVar  
<sup>187</sup> classification. This approach included all variant classifications, not limited solely to  
<sup>188</sup> those deemed “pathogenic”, and explicitly conditioned the classification on the given  
<sup>189</sup> phenotype, recognising that a variant could only be considered pathogenic relative to  
<sup>190</sup> a defined clinical context. The core calculations proceeded as follows:

<sup>191</sup> **1. Allele Frequency and Total Variant Frequency.** For each variant  $i$  in a  
<sup>192</sup> gene, the allele frequency was denoted as  $p_i$ . For each gene, we defined the total  
<sup>193</sup> variant frequency (summing across all reported variants in that gene) as:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## 232 **2.3 Validation of Autosomal Dominant Estimates Using *NFKB1***

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

237 **1. Reference Dataset.** We used a reference dataset reported by Tuijnenburg  
238 et al. (17) to build a validation model in an AD disease gene. This study performed  
239 whole-genome sequencing of 846 predominantly sporadic, unrelated PID cases from  
240 the NIHR BioResource-Rare Diseases cohort. There were 390 CVID cases in the  
241 cohort. The study identified *NFKB1* as one of the genes most strongly associated  
242 with PID. Sixteen novel heterozygous variants including truncating, missense, and  
243 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

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

244 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

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

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

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

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

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

250 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 HGVS 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

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

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

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

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

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

Thus, the mortality-adjusted expected genotype count became:

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

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

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

271 using a large effective allele count ( $\text{AN}_{\text{eff}}$ ) for illustration. By generating 10,000 poste-  
272 rior samples of  $p$ , we obtained a distribution of the literature-based adjusted expected  
273 counts,  $E_{\text{adj}}$ .

**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:

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

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

278 **2.5 Validation of SCID-specific Estimates Using PID–SCID  
279 Genes**

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

285 **1. Incidence Conversion.** Based on literature, PID occurs in approximately 1 in  
286 1,000 births, whereas SCID occurs in approximately 1 in 100,000 births. Consequently,  
287 in a population of 1,000,000 births there are about 1,000 PID cases and 10 SCID cases.  
288 To express SCID-related variant counts on a per-million PID scale, the observed SCID  
289 counts were multiplied by 100. For example, if a gene is expected to cause SCID in  
290 10 cases within the total PID population, then on a per-million PID basis the count  
291 is  $10 \times 100 = 1,000$  cases (across all relevant genes).

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

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

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

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

## 314 **2.6 Protein Network and Genetic Constraint Interpretation**

315 A PPI network was constructed using protein interaction data from STRINGdb (16).  
316 We previously prepared and reported on this dataset consisting of 19,566 proteins and  
317 505,968 interactions (<https://github.com/DylanLawless/ProteoMCLustR>). Node  
318 attributes were derived from log-transformed score-positive-total values, which in-  
319 formed both node size and colour. Top-scoring nodes (top 15 based on score) were

320 labelled to highlight prominent interactions. To evaluate group differences in score-  
321 positive-total across major disease categories, one-way Analysis of Variance (ANOVA)  
322 was performed followed by Tukey Honestly Significant Difference (HSD) post hoc tests  
323 (and non-parametric Dunn's test for confirmation). GnomAD v4.1 constraint metrics  
324 data was used for the PPI analysis and was sourced from Karczewski et al. (7). This  
325 provided transcript-level metrics, such as observed/expected ratios, Loss-Of-function  
326 Observed/Expected Upper bound Fraction (LOEUF), pLI, and Z-scores, quantifying  
327 loss-of-function and missense intolerance, along with confidence intervals and related  
328 annotations for 211,523 observations.

## 329 **2.7 Gene Set Enrichment Test**

330 To test for overrepresentation of biological functions, the prioritised genes were com-  
331 pared against gene sets from MsigDB (including hallmark, positional, curated, motif,  
332 computational, GO, oncogenic, and immunologic signatures) and WikiPathways using  
333 hypergeometric tests with FUMA (24; 25). The background set consisted of 24,304  
334 genes. Multiple testing correction was applied per data source using the Benjamini-  
335 Hochberg method, and gene sets with an adjusted P-value  $\leq 0.05$  and more than one  
336 overlapping gene are reported.

## 337 **2.8 Deriving novel PID classifications by genetic PPI and 338 clinical features**

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

## 352 **2.9 Probability of observing AlphaMissense pathogenicity**

353 We obtained the subset pathogenicity predictions from AlphaMissense via the Al-  
354 phaFold database and whole genome data from the studies data repository(12; 26).  
355 The AlphaMissense data (genome-aligned and amino acid substitutions) were merged

356 with the panel variants based on genomic coordinate and HGVSc annotation. Occur-  
357 rence probabilities were log-transformed and adjusted (y-axis displaying  $\log_{10}(\text{occurrence}$   
358  $\text{prob} + 1e-5) + 5$ ), to visualise the distribution of pathogenicity scores across the  
359 residue sequence. A Kruskal-Wallis test was used to compare the observed disease  
360 probability across clinical classification groups.

361 

### 3 Results

362 

#### 3.1 Observation Probability Across Disease Genes

363 Our study integrated large-scale annotation databases with gene panels from Pan-  
364 elAppRex to systematically assess disease genes by MOI. By combining population  
365 allele frequencies with ClinVar clinical classifications, we computed an expected obser-  
366 vation probability for each SNV, representing the likelihood of encountering a variant  
367 of a specific pathogenicity for a given phenotype. We report these probabilities for  
368 54,814 ClinVar variant classifications across 557 genes (linked dataset (27)).

369 In practice, our approach computed a simple observation probability for every  
370 SNV across the genome and was applicable to any disease-gene panel. Here, we fo-  
371 cused on panels related to Primary Immunodeficiency or Monogenic Inflammatory  
372 Bowel Disease, using PanelAppRex panel ID 398 as a case study. **Figure 1** dis-  
373 plays all reported ClinVar variant classifications for this panel. The resulting natural  
374 scaling system (-5 to +5) accounts for the frequently encountered combinations of  
375 classification labels (e.g. benign to pathogenic). The resulting data set (27) is briefly  
376 shown in **Table 1** to illustrate that our method yielded estimations of the probability  
377 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

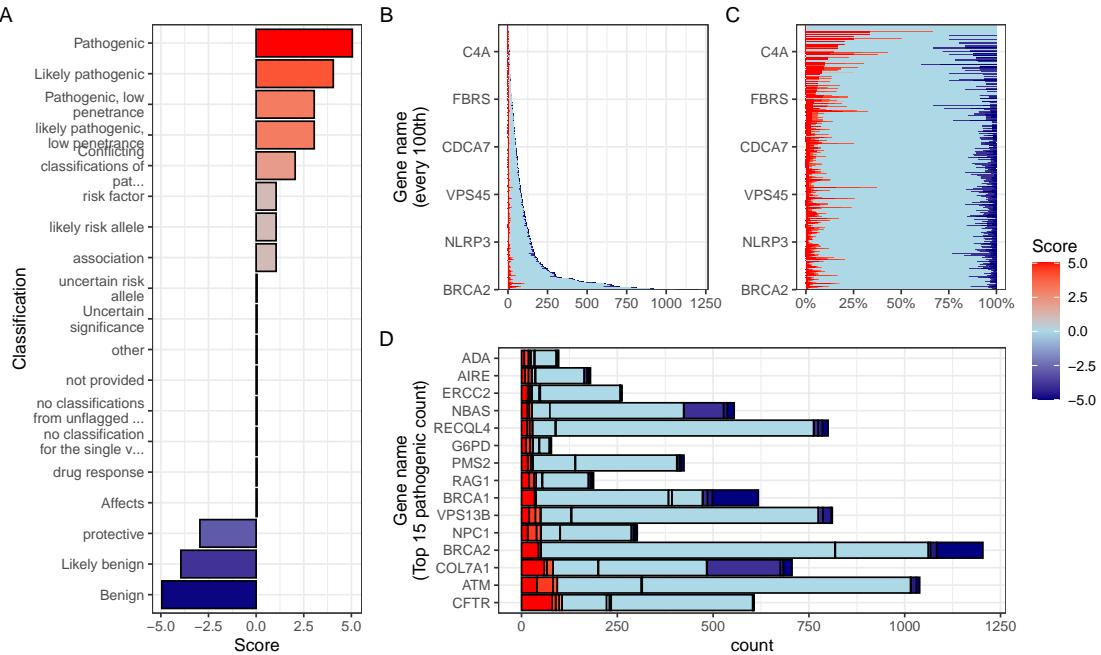


Figure 1: **Summary of ClinVar clinical significance classifications in the PID gene panel.** (A) Shows the numeric score coding for each classification. Panels (B) and (C) display the tally of classifications per gene as absolute counts and as percentages, respectively. (D) Highlights the top 15 genes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

## 378 3.2 Validation studies

### 379 3.2.1 Validation of Dominant Disease Occurrence with *NFKB1*

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

387 Our model calculated 0 known pathogenic variant *NFKB1*-related CVID cases  
 388 in the UK with a minimal risk of 456 unknown de novo variants. In the reference  
 389 cohort, 16 *NFKB1* CVID cases were reported. We additionally wanted to account for  
 390 potential under-reporting in the reference study. We used an extrapolated national  
 391 CVID prevalence which yielded a median estimate of 118 cases (95% CI: 70–181),  
 392 while a Bayesian-adjusted mixture estimate produced a median of 67 cases (95% CI:  
 393 43–99). **Figure S1 (A)** illustrates that our predicted values reflect these ranges and  
 394 are closer to the observed count. This case supports the validity of our integrated

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

### 398 3.2.2 Validation of Recessive Disease Occurrence with *CFTR*

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

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

412 **Figure S2** shows the final values for these genes of interest in a given population  
413 size and phenotype. It reveals that an allele frequency threshold of approximately  
414 0.000007 is required to observe a single heterozygous disease-causing variant carrier in  
415 the UK population for both genes. However, owing to the AR MOI pattern of *CFTR*,  
416 this threshold translates into more than 100,000 heterozygous carriers, compared to  
417 only 456 carriers for the AD gene *NFKB1*. Note that this allele frequency threshold,  
418 being derived from the current reference population, represents a lower bound that  
419 can become more precise as public datasets continue to grow. This marked difference  
420 underscores the significant impact of MOI patterns on population carrier frequencies  
421 and the observed disease prevalence.

### 422 3.2.3 Interpretation of ClinVar Variant Observations

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

### 428 3.2.4 Validation of SCID-specific Disease Occurrence

429 Given that SCID is a subset of PID, our probability estimates reflect the likelihood of  
430 observing a genetic variant as a diagnosis when the phenotype is PID. However, we

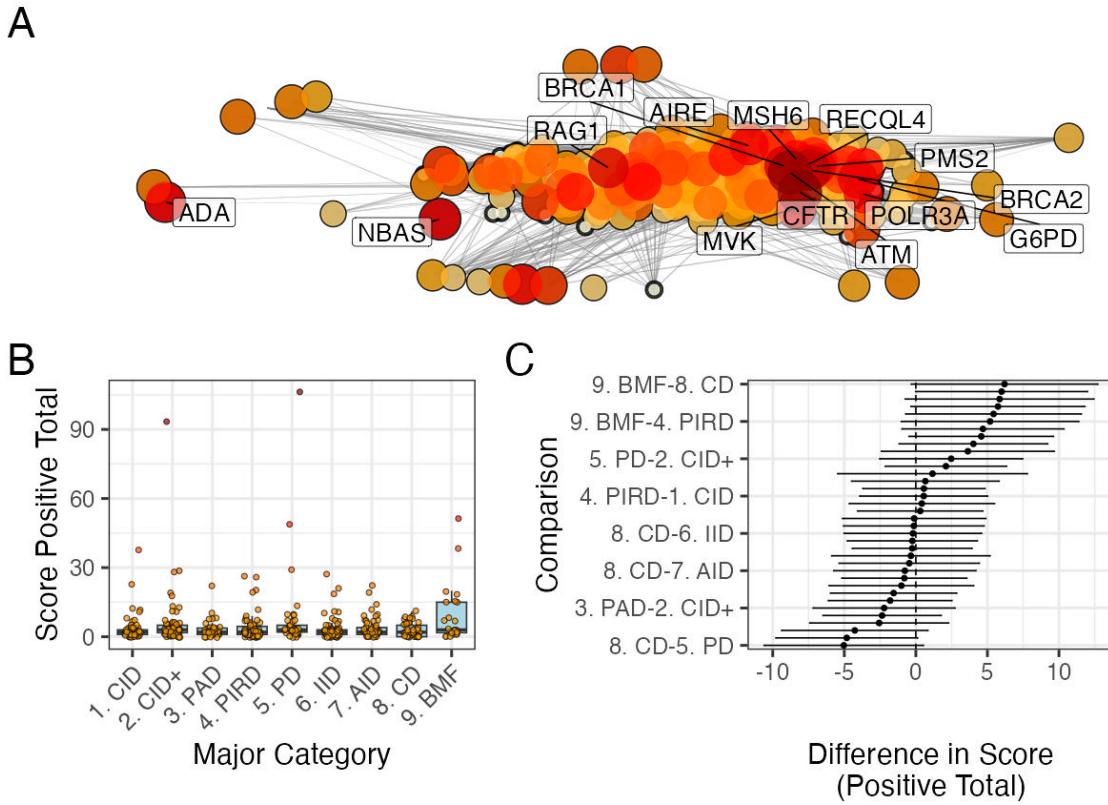
431 additionally tested our results against SCID cohorts in **Figure S4**. The summarised  
432 raw cohort data for SCID-specific gene counts are summarised and compared across  
433 countries in **Figure S3**. True counts for *IL2RG* and *DCLRE1C* from ten distinct  
434 locations yielded 95% confidence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

### 444 3.3 Genetic constraint in high-impact protein networks

445 We next examined genetic constraint in high-impact protein networks across the whole  
446 IEI gene set of over 500 known disease-gene phenotypes (1). By integrating ClinVar  
447 variant classification scores with PPI data, we quantified the pathogenic burden per  
448 gene and assessed its relationship with network connectivity and genetic constraint  
449 (7; 16).

#### 450 3.3.1 Score-Positive-Total within IEI PPI network

451 The ClinVar classifications reported in **Figure 1** were scaled -5 to +5 based on their  
452 pathogenicity. We were interested in positive (potentially damaging) but not negative  
453 (benign) scoring variants, which are statistically incidental in this analysis. We tallied  
454 gene-level positive scores to give the score positive total metric. **Figure 2 (A)** shows  
455 the PPI network of disease-associated genes, where node size and colour encode the  
456 score positive total (log-transformed). The top 15 genes with the highest total prior  
457 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.

#### 458 3.3.2 Association Analysis of Score-Positive-Total across IEI Categories

459 We checked for any statistical enrichment in score positive totals, which represents  
 460 the expected observation of pathogenicity, between the IEI categories. The one-way  
 461 ANOVA revealed an effect of major disease category on score positive total ( $F(8, 500) =$   
 462 2.82,  $p = 0.0046$ ), indicating that group means were not identical, which we observed  
 463 in **Figure 2 (B)**. However, despite some apparent differences in median scores across  
 464 categories (i.e. 9. Bone Marrow Failure (BMF)), the Tukey HSD post hoc compar-  
 465 isons **Figure 2 (C)** showed that all pairwise differences had 95% confidence intervals  
 466 overlapping zero, suggesting that individual group differences were not significant.

467 **3.3.3 UMAP Embedding of the PPI Network**

468 To address the density of the PPI network for the IEI gene panel, we applied Uniform  
 469 Manifold Approximation and Projection (UMAP) (**Figure 3**). Node sizes reflect  
 470 interaction degree, a measure of evidence-supported connectivity (16). We tested  
 471 for a correlation between interaction degree and score positive total. In **Figure**  
 472 **3**, gene names with degrees above the 95th percentile are labelled in blue, while  
 473 the top 15 genes by score positive total are labelled in yellow (as per **Figure 1**).  
 474 Notably, genes with high pathogenic variant loads segregated from highly connected  
 475 nodes, suggesting that Loss-of-Function (LOF) in hub genes is selectively constrained,  
 476 whereas damaging variants in lower-degree genes yield more specific effects. This  
 477 observation was subsequently tested empirically.

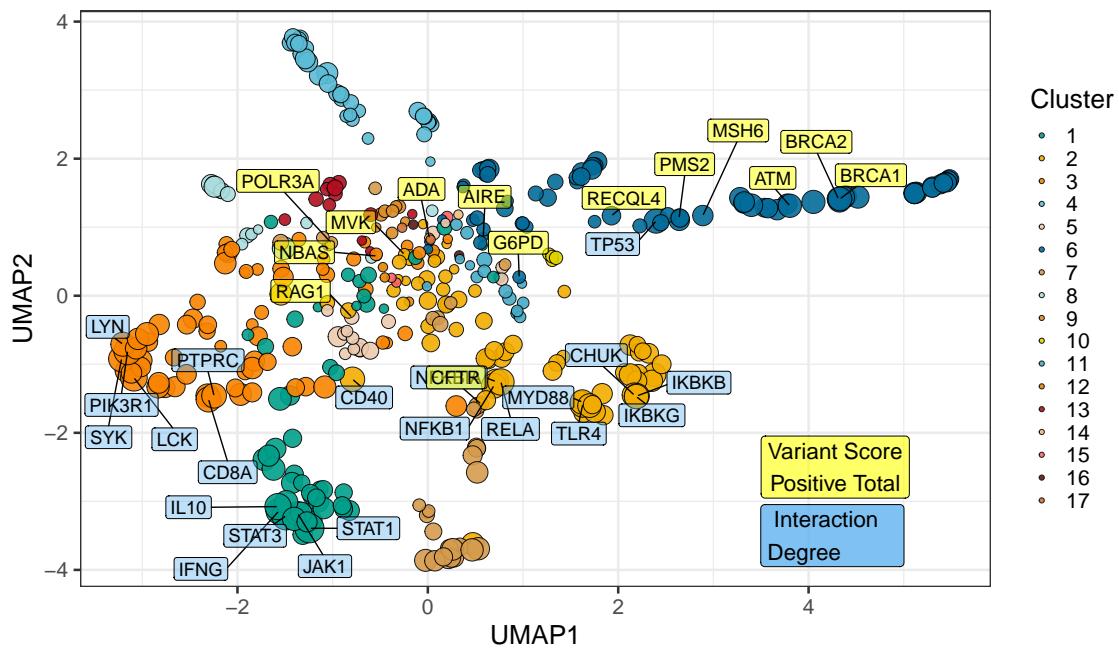


Figure 3: **UMAP embedding of the PPI network (p\_umap).** The plot projects the high-dimensional protein-protein interaction network into two dimensions, with nodes coloured by cluster and sized by interaction degree. Blue labels indicate hub genes (degree above the 95th percentile) and yellow labels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

478 **3.3.4 Hierarchical Clustering of Enrichment Scores for Major Disease Cat-**  
 479 **egories**

480 **Figure S5** presents a heatmap of standardised residuals for major disease categories  
 481 across network clusters, as per **Figure 3**. A dendrogram clusters similar disease cate-

482 gories, while the accompanying bar plot displays the maximum absolute standardised  
483 residual for each category. Notably, (8) Complement Deficiencies (CD) shows the  
484 highest maximum enrichment, followed by (9) BMF. While all maximum values  
485 exceed 2, the threshold for significance, this likely reflects the presence of protein  
486 clusters with strong damaging variant scores rather than uniform significance across  
487 all categories (i.e. genes from cluster 4 in 8 CD).

488 **3.3.5 PPI Connectivity, LOEUF Constraint and Enriched Network Clus-  
489 ter Analysis**

490 Based on the preliminary insight from **Figure S5**, we evaluated the relationship  
491 between network connectivity (PPI degree) and LOEUF constraint (LOEUF upper rank)  
492 Karczewski et al. (7) using Spearman's rank correlation. Overall, there was a weak  
493 but significant negative correlation ( $\rho = -0.181, p = 0.00024$ ) at the global scale,  
494 indicating that highly connected genes tend to be more constrained. A supplementary  
495 analysis (**Figure S6**) did not reveal distinct visual associations between network  
496 clusters and constraint metrics, likely due to the high network density. However  
497 once stratified by gene clusters, the natural biological scenario based on quantitative  
498 PPI evidence (16), some groups showed strong correlations; for instance, cluster 2  
499 ( $\rho = -0.375, p = 0.000994$ ) and cluster 4 ( $\rho = -0.800, p < 0.000001$ ), while others did  
500 not. This indicated that shared mechanisms within pathway clusters may underpin  
501 genetic constraints, particularly for LOF intolerance. We observe that the score  
502 positive total metric effectively summarises the aggregate pathogenic burden across  
503 IEI genes, serving as a robust indicator of genetic constraint and highlighting those  
504 with elevated disease relevance.

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

509 **3.4 New Insight from Functional Enrichment**

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

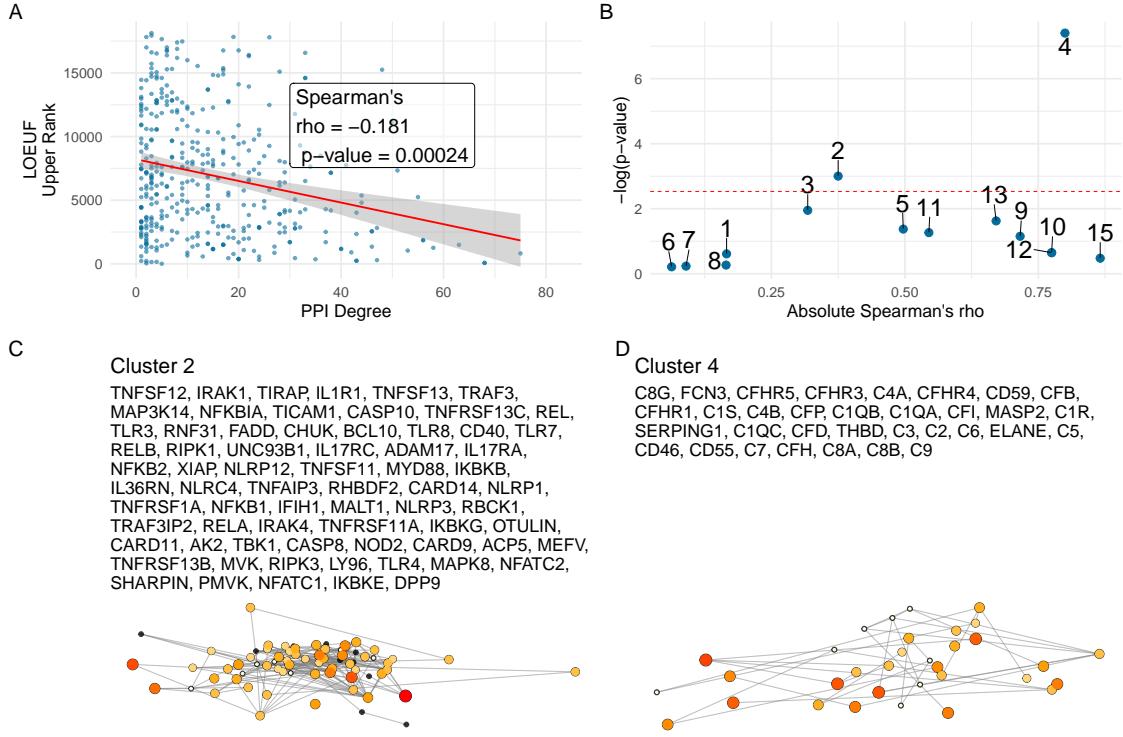


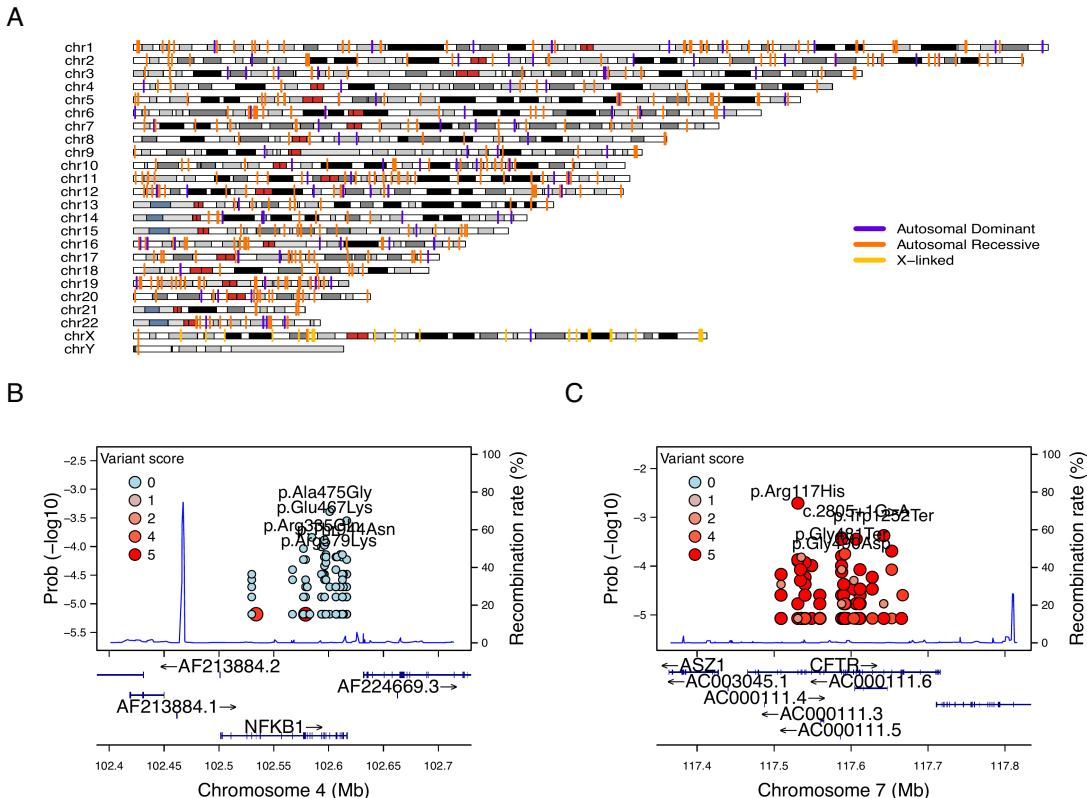
Figure 4: **Correlation between PPI degree and LOEUF upper rank.** **(A)** Ananlysis across all genes revealed a weak, significant negative correlation between PPI degree and LOEUF upper rank. **(B)** The cluster-wise analysis showed that clusters 2 and 4 exhibited moderate to strong correlations, while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

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

### 529 3.5 Genome-wide Gene Distribution and Locus-specific Vari- 530 ant Occurrence

531 **Figure 5 (A)** shows a genome-wide karyoplot of all IEI panel genes across GRCh38,  
 532 with colour-coding based on MOI. Figures (B) and (C) display zoomed-in locus plots  
 533 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.



**Figure 5: Genome-wide IEI, variant occurrence probability and LD by  $R^2$ .**  
 (A) Genome-wide karyoplot of all IEI panel genes mapped to GRCh38, with colours indicating MOI. (B) Zoomed-in locus plot for *NFKB1* showing variant observation probabilities; only benign variants such exhibit high probabilities in this AD gene intolerant to LOF. (C) Locus plot for *CFTR* displaying high probabilities for pathogenic variants; due to the dense clustering of pathogenic variants, score filter  $>0$  was applied. Top five variant are labelled per gene.

### 540 3.6 Novel PID classifications derived from genetic PPI and 541 clinical features

542 We recategorised 315 immunophenotypic features from the original IUIS IEI annotations,  
543 reducing detailed descriptions (e.g. “decreased cd8, normal or decreased cd4”),

544 first to minimal labels (e.g.“low”), and second to binary outcomes (normal vs. not-  
545 normal) for T cells, B cells, neutrophils, and immunoglobulins (**Figure 6**). These  
546 simplified profiles were integrated with PPI network clustering from STRINGdb to  
547 refine PID gene groupings. Chi-square analyses confirmed significant associations be-  
548 tween specific clinical abnormalities and PPI clusters (**Figure ??**). A decision tree  
549 classifier, with hyperparameters optimised via 5-fold cross validation, demonstrated  
550 high sensitivity and specificity, as shown in the confusion matrices and variable impor-  
551 tance metrics (**Figure S10**). The resulting novel PID classifications, illustrated by  
552 the decision tree and gene group distributions (**Figure 9**), provide a more coherent  
553 and data-driven framework for categorising PID genes.

554 **3.7 Novel PID classifications derived from genetic PPI and**  
555 **clinical features**

556 We recategorised 315 immunophenotypic features from the original IUIS IEI annota-  
557 tions, reducing detailed descriptions (e.g. “decreased CD8, normal or decreased CD4”) to  
558 minimal labels (e.g. “low”) and then binarising them (normal vs. not-normal) for T  
559 cells, B cells, Immunoglobulin (Ig) and neutrophils (**Figure 6**). These simplified pro-  
560 files were mapped onto STRINGdb PPI clusters, revealing non-random distributions  
561 ( $\chi^2 < 1e-13$ ; **Figure 7**), indicating that network context captures key immunophe-  
562 notypic variation.

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

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

585 classification schemes.

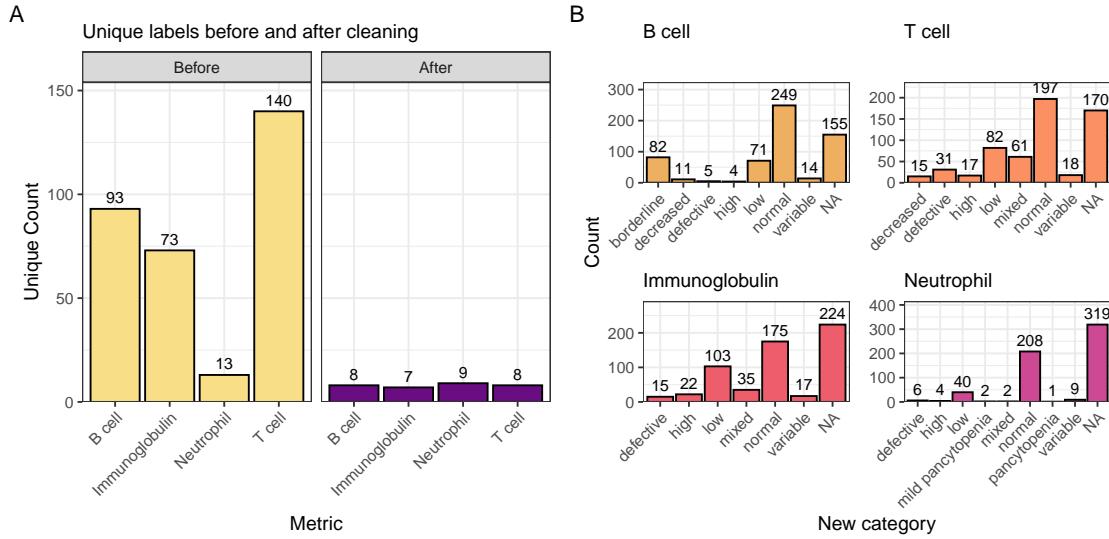


Figure 6: **Distribution of immunophenotypic features before and after recategorisation.** The original IUIS IEI descriptions contain information such as T cell-related “decreased cd8, normal or decreased cd4 cells” which we recategorise as “low”. The bar plot shows the count of unique labels for each status (normal, not\_normal) across the T cell, B cell, Ig, and Neutrophil features.

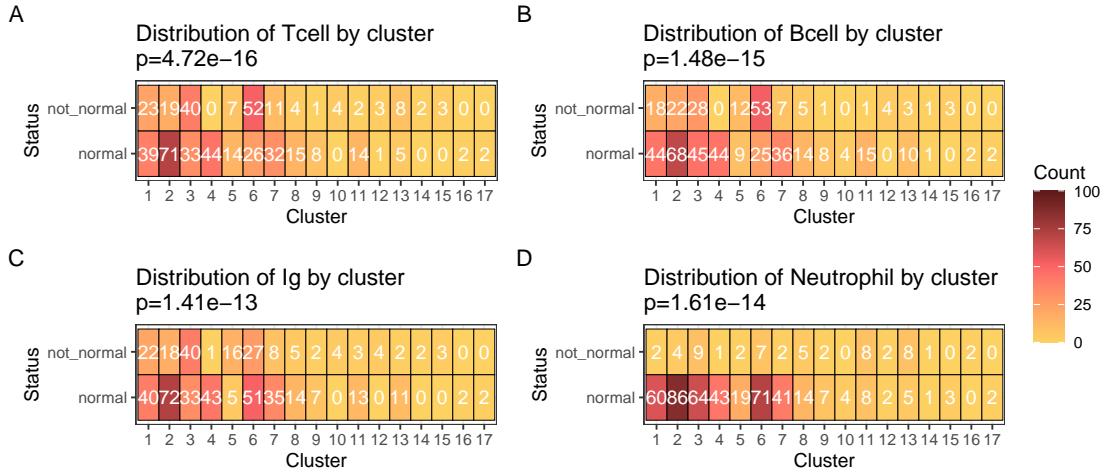


Figure 7: **Heatmaps of clinical feature distributions by PPI cluster.** The heatmaps display the count of observations for abnormality of each clinical feature (A) T cell, (B) B cell, (C) Immunoglobulin, (D) Neutrophil, in relation to the PPI clusters, with p-values from chi-square tests annotated in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

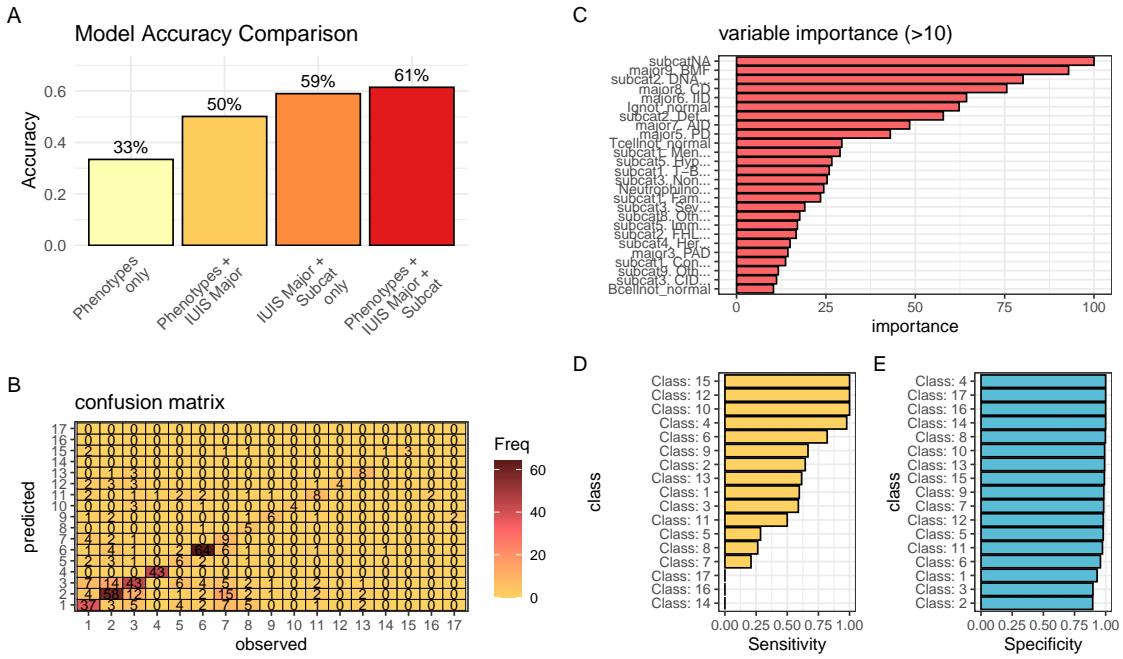


Figure 8: **Performance comparison of PID classifiers.** Classification predicting PPI cluster membership from IUIS major category, subcategory, and immunological features. (A) Overall accuracy for four rpart models used to predict PPI clustering. The combined model achieves 61.4 % accuracy, exceeding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

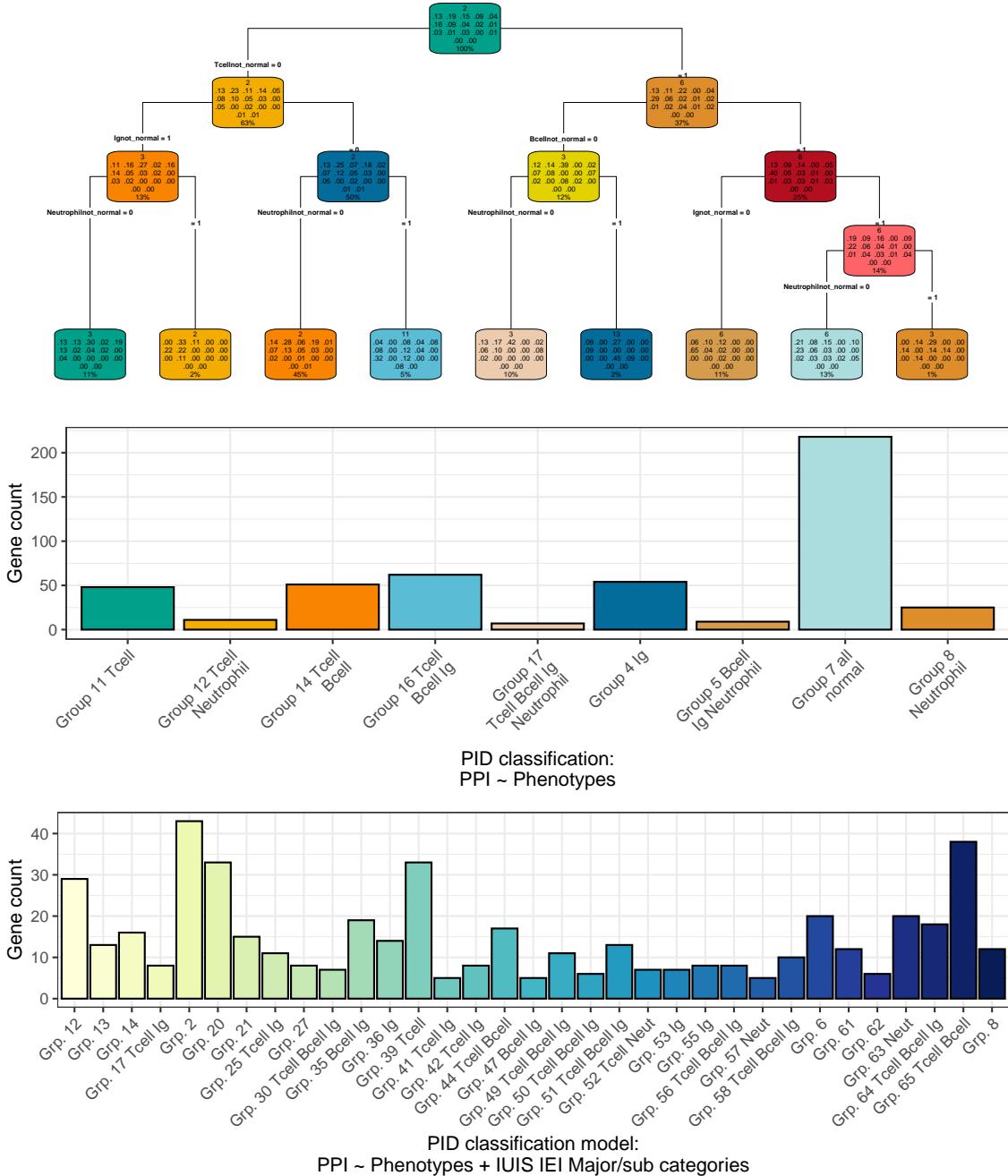


Figure 9: **Fine-tuned model for PID classification.** (Top) In each terminal node, the top block indicates the number of genes in the node; the middle block shows the fitted class probabilities (which sum to 1); and the bottom block displays the percentage of the total sample in that node. These metrics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

586 **3.7.1 Integration of Variant Probabilities into IEI Genetics Data**

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

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

The screenshot displays a table titled "Viewer Zoom" with a search bar at the top. The table has 13 columns: Major category, Subcategory, Disease, Genetic defect, Inheritance, Gene score, Prior prob of pathogenicity, ClinVar SNV classification, ClinVar all variant reports, OMIM, Alpha Missense / Uniprot ID, HPO combined, and HPO term. The rows represent different genes and their variants, such as CD3e deficiency, IL2RG, IL7Ra, ITPKB, JAK3, and LAT. Each row contains numerical values for the first four columns and sparkline box plots for the next three columns. The last four columns contain hyperlinks to external databases. A footer at the bottom shows "1-10 of 591 rows" and "Show 10".

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.

600 **3.8 Probability of observing AlphaMissense pathogenicity**

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

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

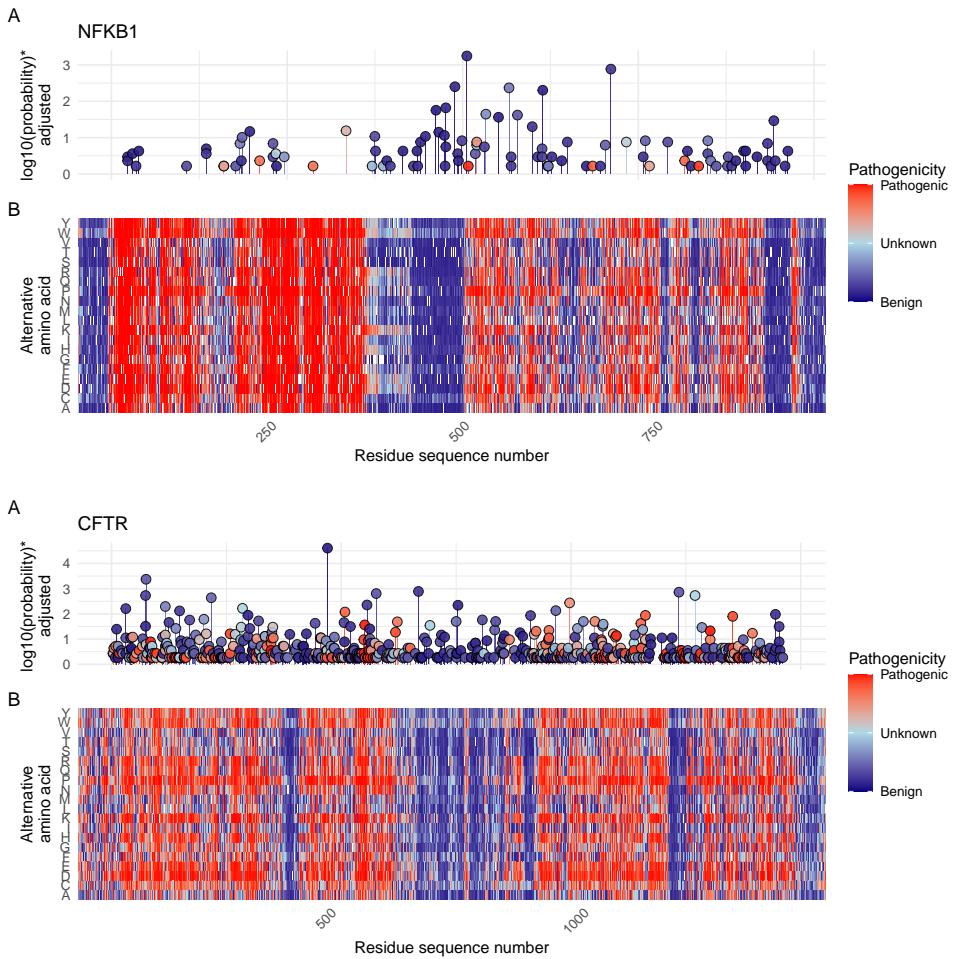


Figure 11: **(A) Probabilities of observing a patient with (B) AlphaMissense-derived pathogenicity scores.** Although AlphaMissense provides scores for every possible amino acid substitution, the most frequently observed variants in patients tend to be classified as benign or of unknown significance.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

## 612 4 Discussion

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

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

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

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

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

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

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

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

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

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

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

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

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

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

671 **5 Conclusion**

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

675 **Acknowledgements**

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

690 **Competing interest**

691 We declare no competing interest.

692 **References**

- 693 [1] Stuart G. Tangye, Waleed Al-Herz, Aziz Bousfiha, Charlotte Cunningham-  
694 Rundles, Jose Luis Franco, Steven M. Holland, Christoph Klein, Tomohiro Morio,  
695 Eric Oksenhendler, Capucine Picard, Anne Puel, Jennifer Puck, Mikko R. J.  
696 Seppänen, Raz Somech, Helen C. Su, Kathleen E. Sullivan, Troy R. Torgerson,  
697 and Isabelle Meyts. Human Inborn Errors of Immunity: 2022 Update  
698 on the Classification from the International Union of Immunological Societies  
699 Expert Committee. *Journal of Clinical Immunology*, 42(7):1473–1507, October  
700 2022. ISSN 0271-9142, 1573-2592. doi: 10.1007/s10875-022-01289-3. URL  
701 <https://link.springer.com/10.1007/s10875-022-01289-3>.
- 702 [2] Dylan Lawless. PanelAppRex aggregates disease gene panels and facilitates  
703 sophisticated search. March 2025. doi: 10.1101/2025.03.20.25324319. URL  
704 <http://medrxiv.org/lookup/doi/10.1101/2025.03.20.25324319>.

- 705 [3] Antonio Rueda Martin, Eleanor Williams, Rebecca E. Foulger, Sarah Leigh,  
706 Louise C. Daugherty, Olivia Niblock, Ivone U. S. Leong, Katherine R. Smith,  
707 Oleg Gerasimenko, Eik Haraldsdottir, Ellen Thomas, Richard H. Scott, Emma  
708 Baple, Arianna Tucci, Helen Brittain, Anna De Burca, Kristina Ibañez, Dalia  
709 Kasperaviciute, Damian Smedley, Mark Caulfield, Augusto Rendon, and Ellen M.  
710 McDonagh. PanelApp crowdsources expert knowledge to establish consensus  
711 diagnostic gene panels. *Nature Genetics*, 51(11):1560–1565, November 2019.  
712 ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-019-0528-2. URL <https://www.nature.com/articles/s41588-019-0528-2>.
- 713
- 714 [4] Oliver Mayo. A Century of Hardy–Weinberg Equilibrium. *Twin Research  
and Human Genetics*, 11(3):249–256, June 2008. ISSN 1832-4274, 1839-  
715 2628. doi: 10.1375/twin.11.3.249. URL [https://www.cambridge.org/core/product/identifier/S1832427400009051/type/journal\\_article](https://www.cambridge.org/core/product/identifier/S1832427400009051/type/journal_article).
- 716
- 717 [5] Nikita Abramovs, Andrew Brass, and May Tassabehji. Hardy-Weinberg Equi-  
718 librium in the Large Scale Genomic Sequencing Era. *Frontiers in Genetics*,  
719 11:210, March 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.00210. URL  
720 <https://www.frontiersin.org/article/10.3389/fgene.2020.00210/full>.
- 721
- 722 [6] Johannes Zschocke, Peter H. Byers, and Andrew O. M. Wilkie. Mendelian  
723 inheritance revisited: dominance and recessiveness in medical genetics. *Nature  
Reviews Genetics*, 24(7):442–463, July 2023. ISSN 1471-0056, 1471-0064.  
724 doi: 10.1038/s41576-023-00574-0. URL <https://www.nature.com/articles/s41576-023-00574-0>.
- 725
- 726
- 727 [7] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings,  
728 Jessica Alfoldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea  
729 Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified  
730 from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.
- 731
- 732 [8] Sarah L. Bick, Aparna Nathan, Hannah Park, Robert C. Green, Monica H. Wo-  
733 jcik, and Nina B. Gold. Estimating the sensitivity of genomic newborn screen-  
734 ing for treatable inherited metabolic disorders. *Genetics in Medicine*, 27(1):  
735 101284, January 2025. ISSN 10983600. doi: 10.1016/j.gim.2024.101284. URL  
<https://linkinghub.elsevier.com/retrieve/pii/S1098360024002181>.
- 736
- 737 [9] Benjamin D. Evans, Piotr Słowiński, Andrew T. Hattersley, Samuel E. Jones,  
738 Seth Sharp, Robert A. Kimmitt, Michael N. Weedon, Richard A. Oram,  
739 Krasimira Tsaneva-Atanasova, and Nicholas J. Thomas. Estimating disease  
740 prevalence in large datasets using genetic risk scores. *Nature Communications*,  
741 12(1):6441, November 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-26501-7.  
742 URL <https://www.nature.com/articles/s41467-021-26501-7>.
- 743
- 744 [10] William B. Hannah, Mitchell L. Drumm, Keith Nykamp, Tiziano Prampano,  
745 Robert D. Steiner, and Steven J. Schrodi. Using genomic databases to de-  
termine the frequency and population-based heterogeneity of autosomal reces-  
sive conditions. *Genetics in Medicine Open*, 2:101881, 2024. ISSN 29497744.

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

- 747
- 748 [11] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov,  
749 Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek,  
750 Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J.  
751 Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishabh  
752 Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy,  
753 Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer,  
754 Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray  
755 Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein  
756 structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August  
757 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03819-2. URL  
758 <https://www.nature.com/articles/s41586-021-03819-2>.
- 759 [12] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor  
760 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  
761 <https://www.science.org/doi/10.1126/science.adg7492>.
- 762 [13] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao,  
763 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:  
764 10.1093/nar/gkx1153. URL <http://academic.oup.com/nar/article/46/D1/D1062/4641904>.
- 765
- 766 [14] The UniProt Consortium, Alex Bateman, Maria-Jesus Martin, Sandra Orchard,  
767 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:  
768 10.1093/nar/gkae1010. URL <https://academic.oup.com/nar/article/53/D1/D609/7902999>.
- 769
- 770 [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>.
- 771
- 772
- 773
- 774
- 775
- 776
- 777
- 778
- 779
- 780
- 781
- 782
- 783
- 784
- 785
- 786

- 787 [16] Damian Szklarczyk, Katerina Nastou, Mikaela Koutrouli, Rebecca Kirsch, Far-  
788 rokh Mehryary, Radja Hachilif, Dewei Hu, Matteo E Peluso, Qingyao Huang,  
789 Tao Fang, et al. The string database in 2025: protein networks with directional-  
790 ity of regulation. *Nucleic Acids Research*, 53(D1):D730–D737, 2025.
- 791 [17] Paul Tuijnenburg, Hana Lango Allen, Siobhan O. Burns, Daniel Greene,  
792 Machiel H. Jansen, and Others. Loss-of-function nuclear factor B subunit  
793 1 (NFKB1) variants are the most common monogenic cause of common vari-  
794 able immunodeficiency in Europeans. *Journal of Allergy and Clinical Im-*  
795 *munology*, 142(4):1285–1296, October 2018. ISSN 00916749. doi: 10.1016/  
796 j.jaci.2018.01.039. URL <https://linkinghub.elsevier.com/retrieve/pii/S0091674918302860>.
- 797 [18] WHO Scientific Group et al. Primary immunodeficiency diseases: report of a  
798 who scientific group. *Clin. Exp. Immunol.*, 109(1):1–28, 1997.
- 800 [19] Charlotte Cunningham-Rundles and Carol Bodian. Common variable immunod-  
801 eficiency: clinical and immunological features of 248 patients. *Clinical immunol-*  
802 *ogy*, 92(1):34–48, 1999.
- 803 [20] Eric Oksenhendler, Laurence Gérard, Claire Fieschi, Marion Malphettes, Gael  
804 Mouillot, Roland Jaussaud, Jean-François Viallard, Martine Gardembas, Lionel  
805 Galicier, Nicolas Schleinitz, et al. Infections in 252 patients with common variable  
806 immunodeficiency. *Clinical Infectious Diseases*, 46(10):1547–1554, 2008.
- 807 [21] Y Naito, F Adams, S Charman, J Duckers, G Davies, and S Clarke. Uk cystic  
808 fibrosis registry 2023 annual data report. *London: Cystic Fibrosis Trust*, 2023.
- 809 [22] Carlo Castellani, CFTR2 team, et al. Cftr2: how will it help care? *Paediatric*  
810 *respiratory reviews*, 14:2–5, 2013.
- 811 [23] Hartmut Grasemann and Felix Ratjen. Cystic fibrosis. *New England Journal*  
812 *of Medicine*, 389(18):1693–1707, 2023. doi: 10.1056/NEJMra2216474. URL  
813 <https://www.nejm.org/doi/full/10.1056/NEJMra2216474>.
- 814 [24] Kyoko Watanabe, Erdogan Taskesen, Arjen Van Bochoven, and Danielle  
815 Posthuma. Functional mapping and annotation of genetic associations with  
816 FUMA. *Nature Communications*, 8(1):1826, November 2017. ISSN 2041-1723.  
817 doi: 10.1038/s41467-017-01261-5. URL <https://www.nature.com/articles/s41467-017-01261-5>.
- 818 [25] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir,  
819 Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (MSigDB)  
820 3.0. *Bioinformatics*, 27(12):1739–1740, June 2011. ISSN 1367-4811, 1367-  
821 4803. doi: 10.1093/bioinformatics/btr260. URL <https://academic.oup.com/bioinformatics/article/27/12/1739/257711>.
- 822
- 823

- 824 [26] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Tay-  
825 lor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias  
826 Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hass-  
827 abis, Pushmeet Kohli, and Žiga Avsec. Predictions for alphanonsense, September  
828 2023. URL <https://doi.org/10.5281/zenodo.8208688>.
- 829 [27] Dylan Lawless. Variant risk estimate probabilities for iei genes. March 2025. doi:  
830 10.5281/zenodo.15111584. URL <https://doi.org/10.5281/zenodo.15111584>.
- 831 [28] Bradley Efron and Carl Morris. Stein’s Estimation Rule and Its Competitors—  
832 An Empirical Bayes Approach. *Journal of the American Statistical Association*,  
833 68(341):117, March 1973. ISSN 01621459. doi: 10.2307/2284155. URL <https://www.jstor.org/stable/2284155?origin=crossref>.
- 835 [29] Yaowu Liu, Sixing Chen, Zilin Li, Alanna C Morrison, Eric Boerwinkle, and  
836 Xihong Lin. Acat: a fast and powerful p value combination method for rare-  
837 variant analysis in sequencing studies. *The American Journal of Human Genetics*,  
838 104(3):410–421, 2019.
- 839 [30] Xihao Li, Zilin Li, Hufeng Zhou, Sheila M Gaynor, Yaowu Liu, Han Chen, Ryan  
840 Sun, Rounak Dey, Donna K Arnett, Stella Aslibekyan, et al. Dynamic incorpora-  
841 tion of multiple in silico functional annotations empowers rare variant association  
842 analysis of large whole-genome sequencing studies at scale. *Nature genetics*, 52  
843 (9):969–983, 2020.
- 844 [31] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xi-  
845 hong Lin. Rare-variant association testing for sequencing data with the sequence  
846 kernel association test. *The American Journal of Human Genetics*, 89(1):82–93,  
847 2011.
- 848 [32] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J  
849 Rieder, Deborah A Nickerson, David C Christiani, Mark M Wurfel, and Xihong  
850 Lin. Optimal unified approach for rare-variant association testing with applica-  
851 tion to small-sample case-control whole-exome sequencing studies. *The American  
852 Journal of Human Genetics*, 91(2):224–237, 2012.
- 853 [33] Augustine Kong, Gudmar Thorleifsson, Michael L Frigge, Bjarni J Vilhjalmsson,  
854 Alexander I Young, Thorgeir E Thorgeirsson, Stefania Benonisdottir, Asmundur  
855 Oddsson, Bjarni V Halldorsson, Gisli Masson, et al. The nature of nurture:  
856 Effects of parental genotypes. *Science*, 359(6374):424–428, 2018.
- 857 [34] Laurence J Howe, Michel G Nivard, Tim T Morris, Ailin F Hansen, Humaira  
858 Rasheed, Yoonsu Cho, Geetha Chittoor, Penelope A Lind, Teemu Palviainen,  
859 Matthijs D van der Zee, et al. Within-sibship gwas improve estimates of direct  
860 genetic effects. *BioRxiv*, pages 2021–03, 2021.
- 861 [35] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-  
862 Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al.

- 863 Standards and guidelines for the interpretation of sequence variants: a joint  
864 consensus recommendation of the american college of medical genetics and ge-  
865 nomics and the association for molecular pathology. *Genetics in medicine*, 17  
866 (5):405–423, 2015.
- 867 [36] Sean V Tavtigian, Steven M Harrison, Kenneth M Boucher, and Leslie G  
868 Biesecker. Fitting a naturally scaled point system to the acmng/amp variant  
869 classification guidelines. *Human mutation*, 41(10):1734–1737, 2020.
- 870 [37] Quan Li and Kai Wang. Intervar: clinical interpretation of genetic variants by  
871 the 2015 acmng-amp guidelines. *The American Journal of Human Genetics*, 100  
872 (2):267–280, 2017.
- 873 [38] Brent S Pedersen, Joe M Brown, Harriet Dashnow, Amelia D Wallace, Matt  
874 Velinder, Martin Tristani-Firouzi, Joshua D Schiffman, Tatiana Tvrzik, Rong  
875 Mao, D Hunter Best, et al. Effective variant filtering and expected candidate  
876 variant yield in studies of rare human disease. *NPJ Genomic Medicine*, 6(1):1–8,  
877 2021.
- 878 [39] Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon,  
879 Andrew P Morris, and Krina T Zondervan. Data quality control in genetic  
880 case-control association studies. *Nature protocols*, 5(9):1564–1573, 2010. URL  
881 <https://doi.org/10.1038/nprot.2010.116>.
- 882 [40] David T Miller, Kristy Lee, Noura S Abul-Husn, Laura M Amendola, Kyle Broth-  
883 ers, Wendy K Chung, Michael H Gollob, Adam S Gordon, Steven M Harrison,  
884 Ray E Hershberger, et al. Acmng sf v3. 2 list for reporting of secondary findings  
885 in clinical exome and genome sequencing: a policy statement of the american  
886 college of medical genetics and genomics (acmng). *Genetics in Medicine*, 25(8):  
887 100866, 2023.

888 **6 Supplemental**

889 **6.1 Validation studies**

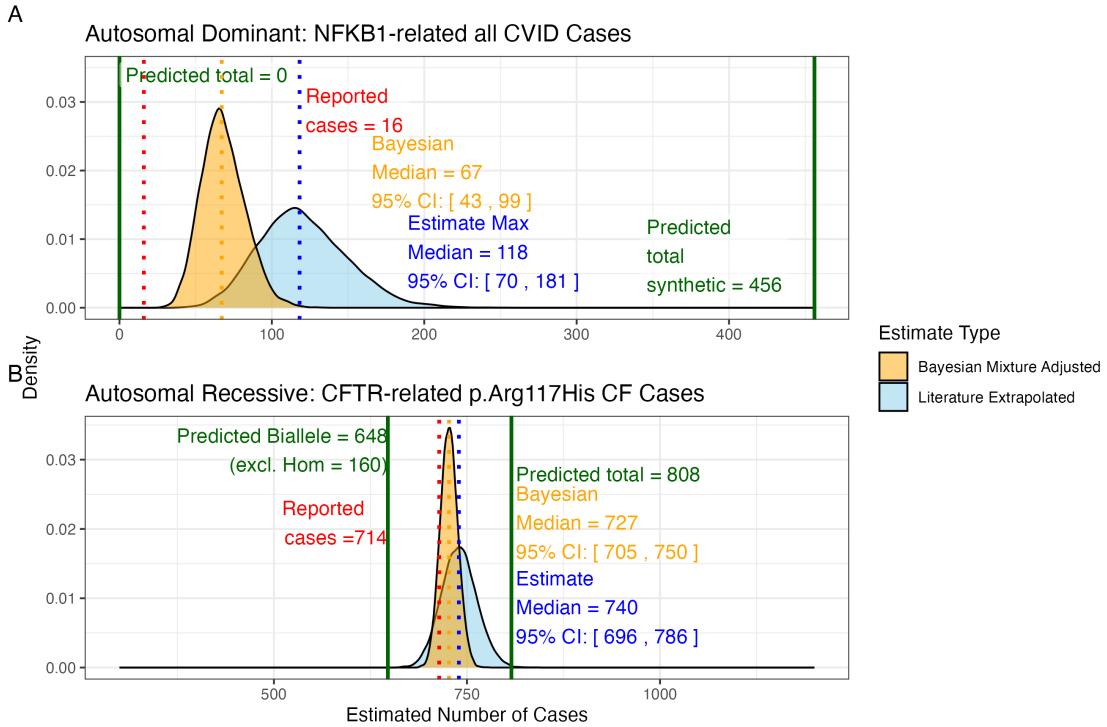
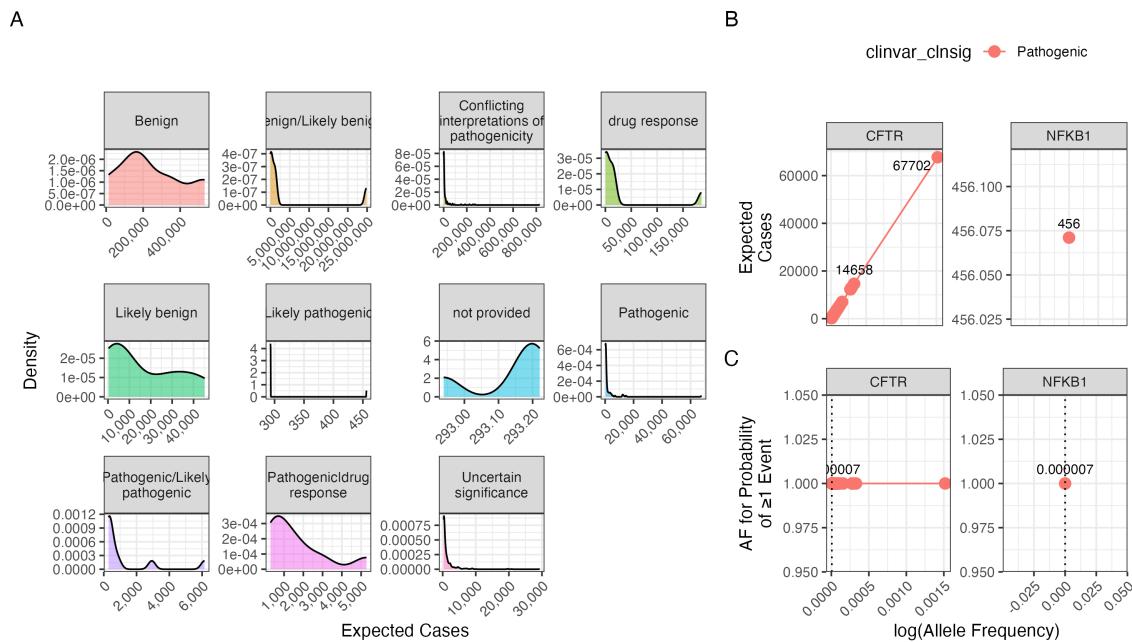
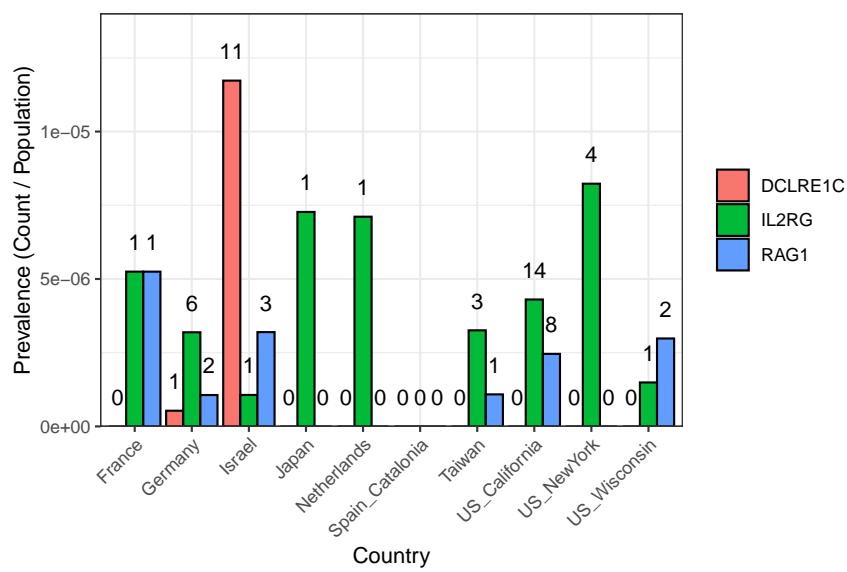


Figure S1: **Prior probabilities compared to validation disease cohort metrics.** (A) Density distributions for the number of *NFKB1*-related CVID cases in the UK. Our model (green) predicted 456 cases, which falls between the observed cohort count (red) of 390 and the upper extrapolated values. The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

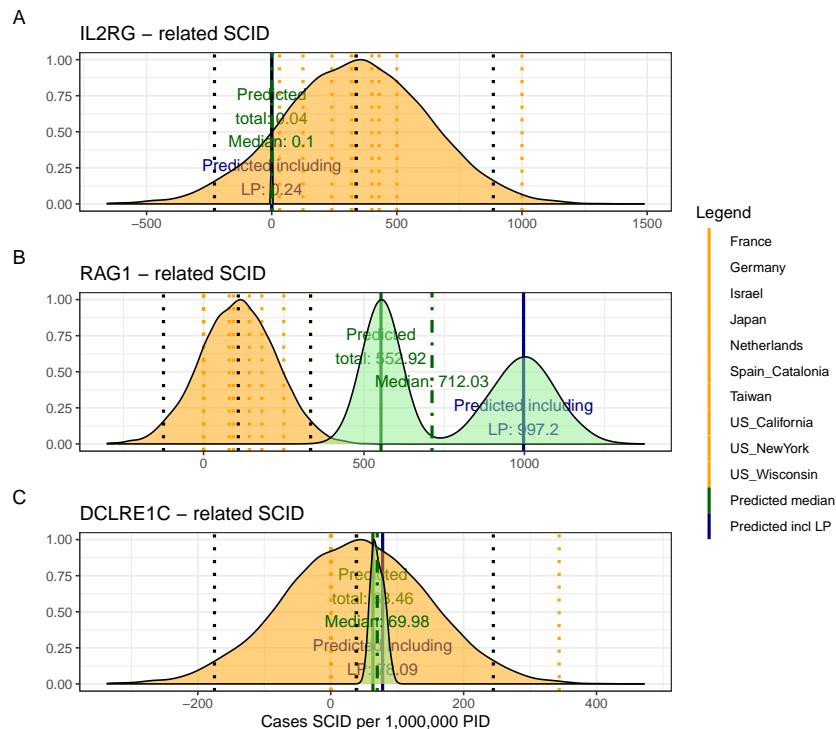
Condition: population size 69433632, phenotype PID-related, genes CFTR and NFKB1.



**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.

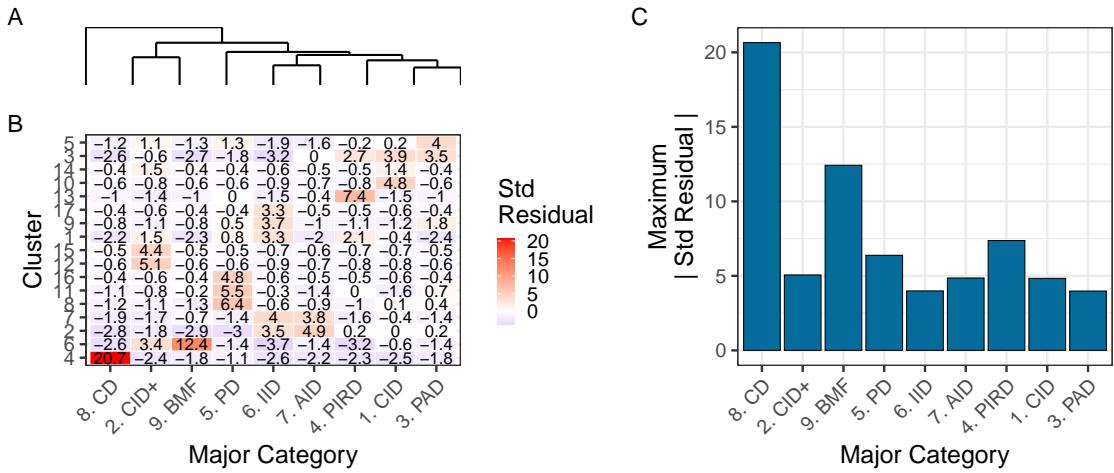


**Figure S3: SCID-specific gene comparison across regions.** The bar plot shows the prevalence of SCID-related cases (count divided by population) for each gene and country (or region), with numbers printed above the bars representing the actual counts in the original cohort (ranging from 0 to 11 per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

890 **6.2 Hierarchical Clustering of Enrichment Scores for Major**  
 891 **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 values, indicating significant enrichment or depletion ( $\text{residuals} > |2|$ ). Definitions in **Box 2.1**.

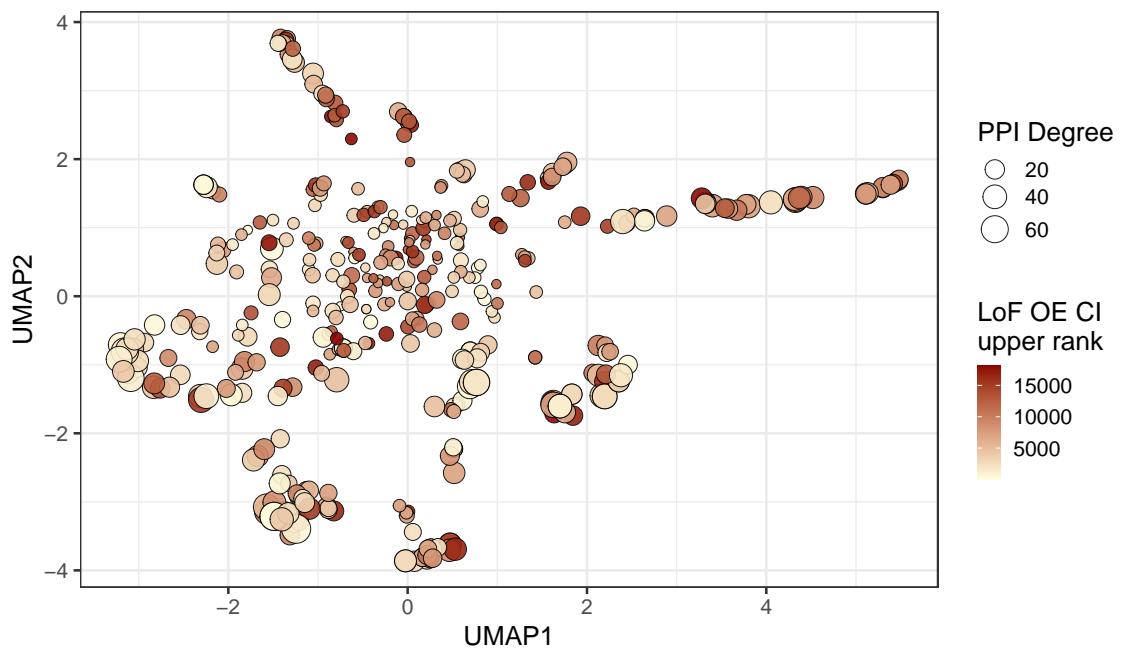
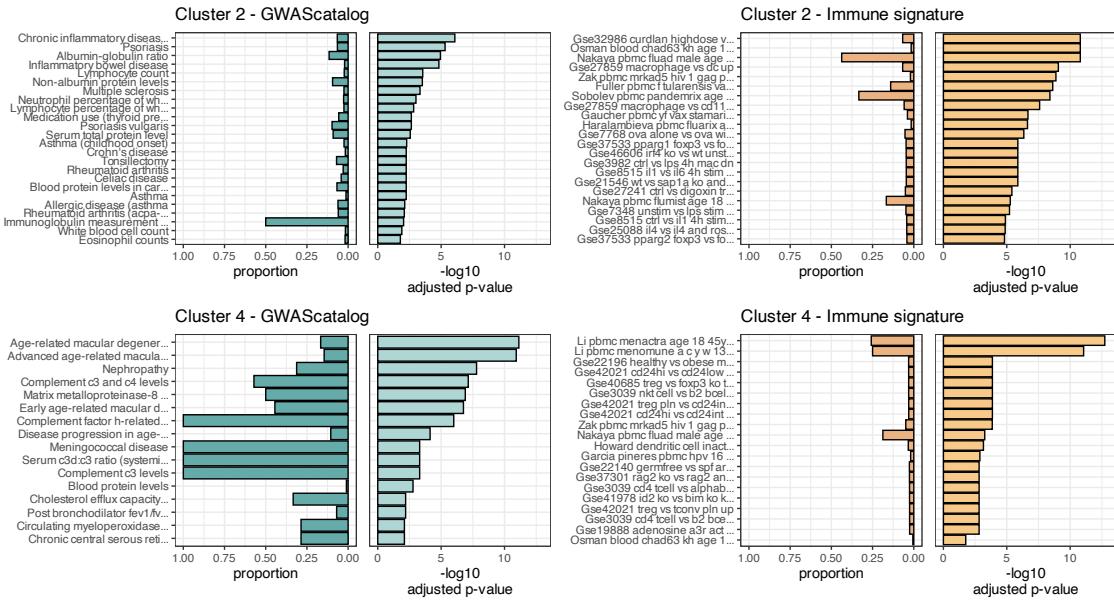
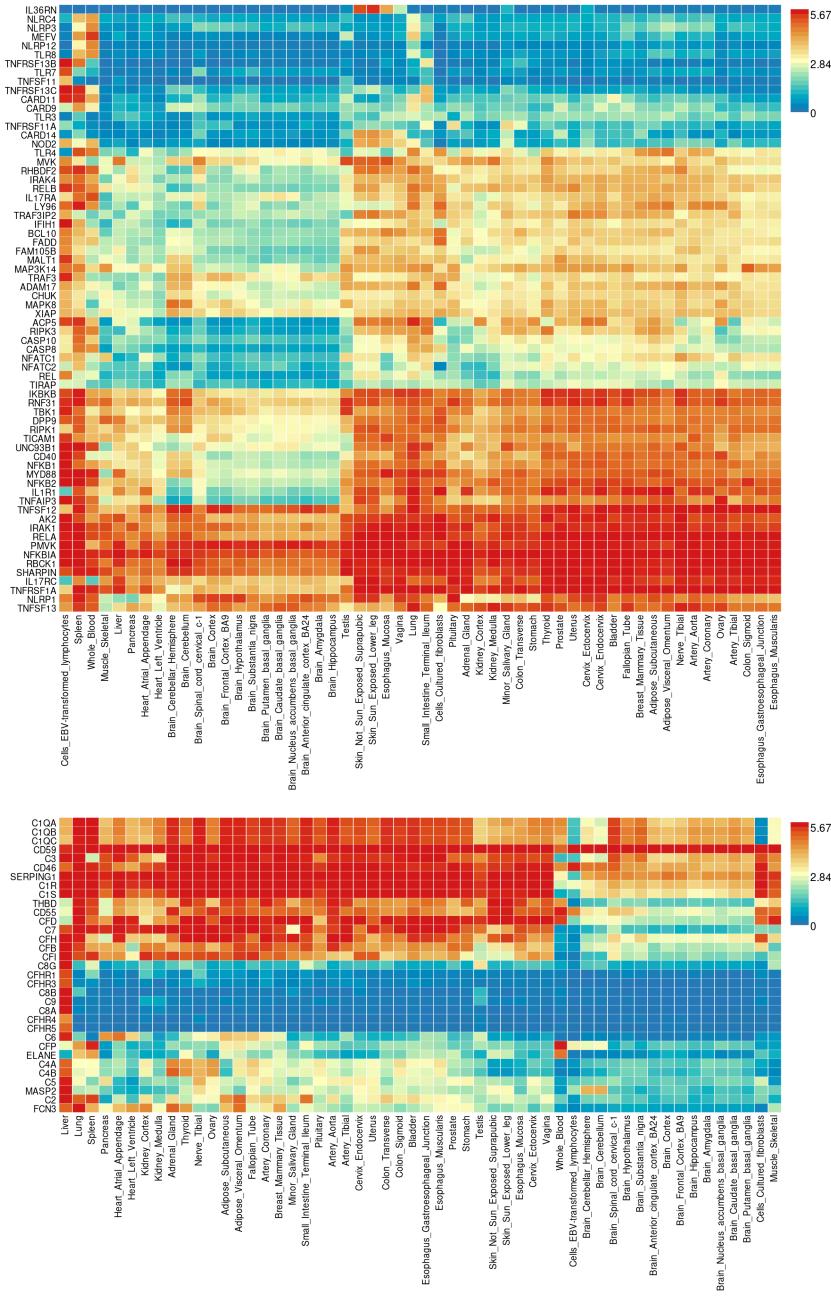


Figure S6: **Analysis of PPI degree versus LOEUF upper rank with UMAP embedding of the PPI network.** The relationship between PPI degree (size) and LOEUF upper rank (color) across gene clusters. No clear patterns are evident.



**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.



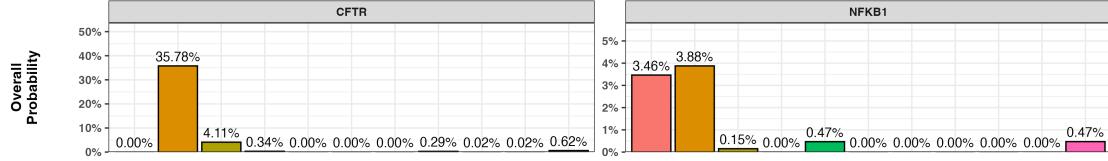
**Figure S8: Gene Expression Heatmaps for IEI Genes.** GTEx v8 data from 54 tissue types display the average expression per tissue label ( $\log_2$  transformed) for the IEI gene panels. Top: Cluster 2; Bottom: Cluster 4.

## 6.3 Interpretation of ClinVar Variant Observations

Recessive and Dominant Disease Genes

A

Overall Probability of an Affected Birth by ClinVar Category



B

Example, Total Expected Cases in UK population size (~ 69.4M)

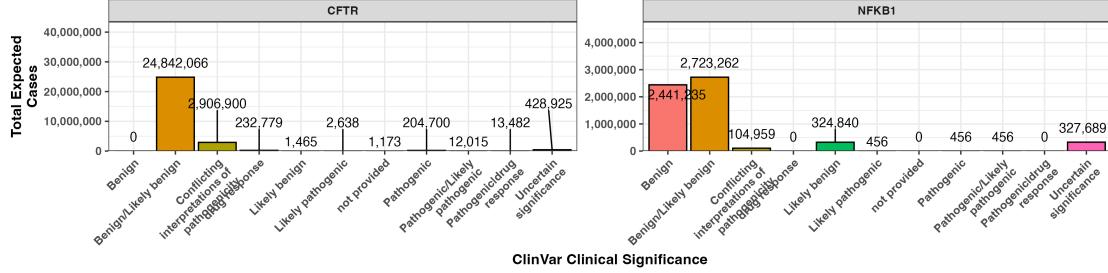
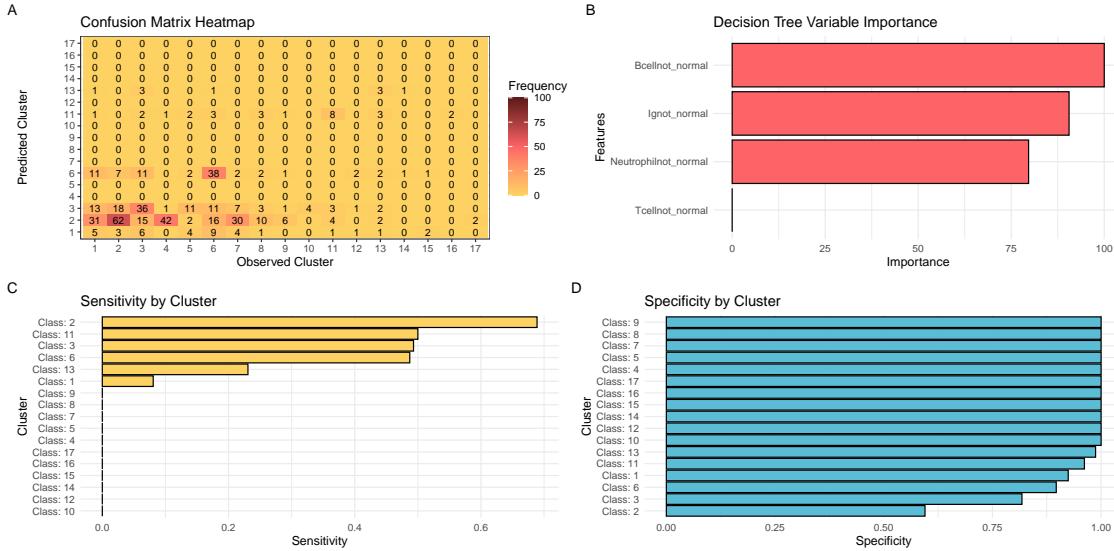


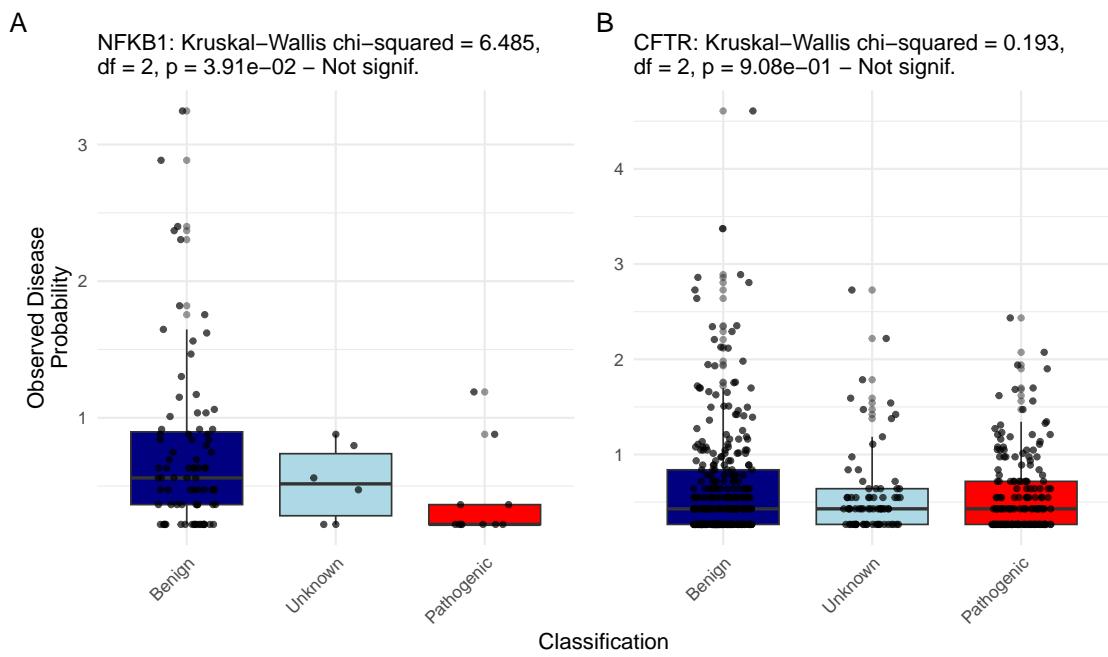
Figure S9: Combined bar charts summarizing the genome-wide analysis of ClinVar clinical significance for the PID gene panel. Panel (A) shows the overall probability of an affected birth by variant classification, and (B) displays the total expected number of cases per classification, both stratified 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.

## 894 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.

## 895 7 Clinical Genetics Application

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

### 906 7.1 Methods

#### 907 7.1.1 Quality Control:

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

#### 916 7.1.2 Prior Probability Calculation:

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

1. **Convert the ACMG Score:** The evidence score  $S$  is normalised to a frac-  
tional support level:

$$S_{\text{adj}} = \frac{S}{100}$$

921 This value reflects the strength of the pathogenic support.

2. **Assign a Minimal Risk ( $\epsilon$ ):** 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:

$$\epsilon = \frac{1}{\max(AN) + 1} \times S_{\text{adj}}$$

922 This step ensures that even low-frequency variants receive a baseline risk pro-  
923 portional to the qualitative evidence.

- 924
- 925
3. **Adjust the Allele Frequency:** The observed allele frequency  $p_i$  is then in-  
creased by  $\epsilon$  to yield an adjusted frequency:

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

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

- 926
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

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

927

### 7.1.3 Deriving the Confidence Interval (CI)

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

- 930
- 931
- 932
- The observed allele frequency and its adjustment via  $\epsilon$ .
  - 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.

933

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

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

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

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

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

941 where  $z = 1.96$  for a 95% confidence level. This interval integrates uncertainty from  
 942 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  $\alpha$  and  $\beta$  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:

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

943 where  $\text{BetaInv}(q; \alpha, \beta)$  denotes the quantile function of the Beta distribution at prob-  
 944 ability  $q$ .

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

## 950 7.2 Results

951 We illustrate our method with two examples:

952   **Example 1: Missing Sequence Information** In one case, a known pathogenic  
953 nucleotide position in *GENE\_XYZ* exhibited low quality reads and did not pass QC.  
954 This missing information prevents confirmation of the absence of the known variant (a  
955 potential false negative), thereby widening the uncertainty in our probability estimate.  
956 In such cases, the adjusted allele frequency is calculated with additional variance,  
957 leading to a broader CI. For instance, if the observed allele frequency is  $1.0 \times 10^{-5}$   
958 and after adjusting with the ACMG score the estimated occurrence probability is  
959  $1.0 \times 10^{-5}$ , the propagated uncertainty might yield a 95% CI of [0.70, 0.85]. This  
960 broader interval reflects the impact of missing sequence data on our confidence.

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

972   Table S1 shows the final variant results for a male patient carrying an X-linked  
973 loss-of-function (LOF) variant in *GENE\_XYZ* where all known pathogenic positions  
974 were confirmed as reference alleles. For the variant c. 1234del (p.Glu412Argfs\*5), the  
975 observed allele frequency is  $1.2 \times 10^{-5}$ . After applying the ACMG evidence score  
976 adjustment (for a VUS lacking a ClinVar classification), the adjusted allele frequency  
977 remains consistent with the observed data. The resulting occurrence probability is  
978  $1.2 \times 10^{-5}$ , and by propagating the uncertainty from the allele frequency, evidence  
979 score adjustment, and the full range of possible outcomes (TP, FP, TN, FN), we  
980 derive a 95% CI for causality of [0.92, 0.97]. This confirms the variant as the top  
981 causal variant in this patient, with no evidence of additional alternative pathogenic  
982 variants.

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

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