

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

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

<sup>1</sup>The quantitative omic epidemiology group.

<sup>2</sup>Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Switzerland.

<sup>3</sup>Division of Immunology and the Children's Research Center, University Children's Hospital Zurich, University of Zurich, Zurich, Switzerland.

<sup>4</sup>Division of Metabolism and Children's Research Center, University Children's Hospital Zürich, University of Zurich, Zurich, Switzerland.

<sup>5</sup>Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, Leeds, UK.

<sup>6</sup>Department of Intensive Care and Neonatology, University Children's Hospital Zurich, University of Zurich, Zurich, Switzerland.

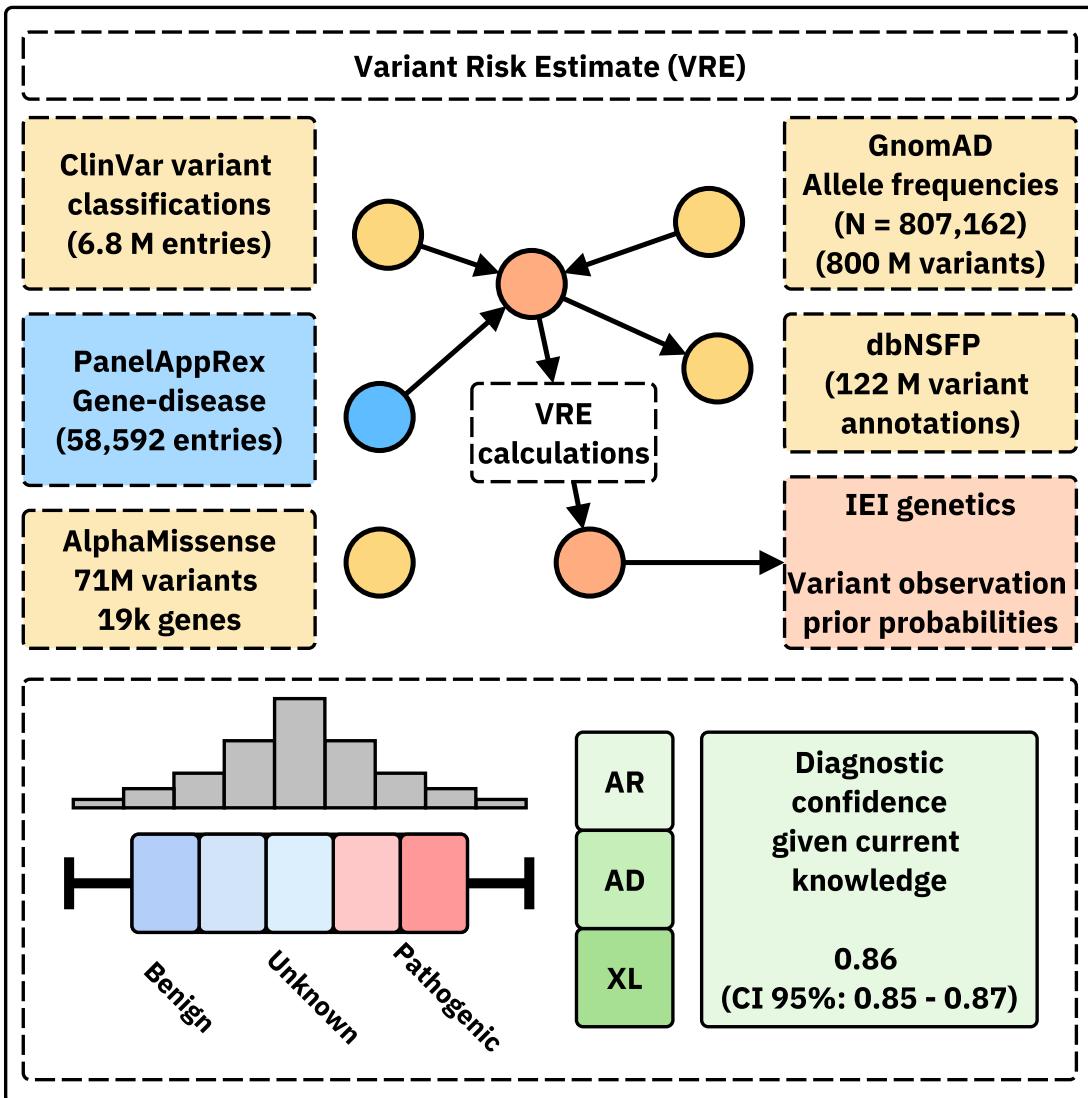
May 6, 2025

## Abstract

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

<sup>1</sup> \* Addresses for correspondence: Dylan.Lawless@kispi.uzh.ch.

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



18

19 Graphical abstract.

20 **Acronyms**

21	<b>ACMG</b> American College of Medical Genetics and Genomics.....	40
22	<b>ACAT</b> Aggregated Cauchy Association Test .....	40
24	<b>AD</b> Autosomal Dominant.....	6
26	<b>AF</b> Allele Frequency.....	6
28	<b>ANOVA</b> Analysis of Variance .....	19
30	<b>AR</b> Autosomal Recessive .....	6
32	<b>BMF</b> Bone Marrow Failure.....	29
34	<b>CD</b> Complement Deficiencies .....	30
36	<b>CI</b> Confidence Interval.....	27
38	<b>CrI</b> Credible Interval .....	13
40	<b>CF</b> Cystic Fibrosis .....	17
42	<b>CFTR</b> Cystic Fibrosis Transmembrane Conductance Regulator.....	8
44	<b>CVID</b> Common Variable Immunodeficiency .....	15
46	<b>DCLRE1C</b> DNA Cross-Link Repair 1C .....	8
48	<b>dbNSFP</b> database for Non-Synonymous Functional Predictions .....	8
50	<b>GE</b> Genomics England .....	7
52	<b>gnomAD</b> Genome Aggregation Database .....	8
54	<b>gVCF</b> genomic variant call format .....	13
56	<b>HGVS</b> Human Genome Variation Society.....	8
58	<b>HPC</b> High-Performance Computing.....	12
60	<b>HSD</b> Honestly Significant Difference .....	19
62	<b>HWE</b> Hardy-Weinberg Equilibrium .....	6
64	<b>IEI</b> Inborn Errors of Immunity .....	6
66	<b>Ig</b> Immunoglobulin .....	33
68	<b>IL2RG</b> Interleukin 2 Receptor Subunit Gamma.....	8
70	<b>InDel</b> Insertion/Deletion .....	8
72	<b>IUIS</b> International Union of Immunological Societies .....	6
74		

75	<b>LD</b> Linkage Disequilibrium .....	32
76	<b>LOEUF</b> Loss-Of-function Observed/Expected Upper bound Fraction .....	19
78	<b>LOF</b> Loss-of-Function .....	19
80	<b>MOI</b> Mode of Inheritance .....	6
82	<b>NFKB1</b> Nuclear Factor Kappa B Subunit 1 .....	8
84	<b>OMIM</b> Online Mendelian Inheritance in Man .....	37
86	<b>PID</b> Primary Immunodeficiency .....	6
88	<b>PPI</b> Protein-Protein Interaction .....	8
90	<b>pLI</b> Probability of being Loss-of-function Intolerant .....	19
92	<b>QC</b> Quality Control .....	13
94	<b>RAG1</b> Recombination activating gene 1 .....	8
96	<b>SCID</b> Severe Combined Immunodeficiency .....	8
98	<b>SNV</b> Single Nucleotide Variant .....	6
100	<b>SKAT</b> Sequence Kernel Association Test.....	40
102	<b>STRINGdb</b> Search Tool for the Retrieval of Interacting Genes/Proteins.....	8
104	<b>TP</b> true positive.....	6
106	<b>FP</b> false positive.....	6
108	<b>TN</b> true negative .....	6
110	<b>FN</b> false negative .....	6
112	<b>TNFAIP3</b> Tumor necrosis factor, alpha-induced protein 3 .....	8
114	<b>UMAP</b> Uniform Manifold Approximation and Projection .....	20
116	<b>UniProt</b> Universal Protein Resource.....	7
118	<b>VCF</b> variant call format .....	13
120	<b>VEP</b> Variant Effect Predictor.....	8
122	<b>VRE</b> variant risk estimate .....	9
124	<b>XL</b> X-Linked .....	6
126		

## <sup>127</sup> 1 Introduction

<sup>128</sup> Accurately determining the probability that a patient harbours a disease-causing  
<sup>129</sup> genetic variant remains a foundational challenge in clinical and statistical genetics.  
<sup>130</sup> For over a century, the primary focus has been on identifying true positive (TP)s,  
<sup>131</sup> pathogenic causal variants observed in affected individuals. Peer review and classifi-  
<sup>132</sup> cation frameworks also work to suppress false positive (FP)s. However, two critical  
<sup>133</sup> components of the genetic landscape have received far less attention: false nega-  
<sup>134</sup> tive (FN)s, where pathogenic variants are missed due to technical or interpretive  
<sup>135</sup> limitations, and true negative (TN)s, which represent the vast majority of benign  
<sup>136</sup> or non-causal variants. TNs are more commonly used in contexts such as cancer  
<sup>137</sup> screening, where a negative result can provide reassurance that a panel of known  
<sup>138</sup> actionable variants has been checked. Yet outside these specific uses, their broader  
<sup>139</sup> statistical and clinical value is rarely leveraged. From a statistical perspective, FNs  
<sup>140</sup> and TNs are an untapped goldmine. They hold essential information about what is  
<sup>141</sup> not observed, what should be expected under baseline assumptions, and how confident  
<sup>142</sup> one can be in the absence of a pathogenic finding. Yet these dimensions are rarely  
<sup>143</sup> quantified, leaving a bias in current variant interpretation frameworks towards known  
<sup>144</sup> TPs and lacking principled priors for genome-wide disease probability estimation.

<sup>145</sup> In this study, we focused on reporting the probability of disease observation  
<sup>146</sup> through genome-wide assessments of gene-disease combinations. Our central hypoth-  
<sup>147</sup> esis was that by using highly curated annotation data including population Allele  
<sup>148</sup> Frequency (AF)s, disease phenotypes, Mode of Inheritance (MOI) patterns, and vari-  
<sup>149</sup> ant classifications and by applying rigorous calculations based on Hardy-Weinberg  
<sup>150</sup> Equilibrium (HWE), we could accurately estimate the expected probabilities of ob-  
<sup>151</sup> serving disease-associated variants. Among other benefits, this knowledge can be used  
<sup>152</sup> to derive genetic diagnosis confidence by incorporating these new priors.

<sup>153</sup> We focused on known Inborn Errors of Immunity (IEI) genes, sometimes called  
<sup>154</sup> the “Primary Immunodeficiency (PID) or Monogenic Inflammatory Bowel Disease  
<sup>155</sup> genes” (1–3), to validate our approach and demonstrate its clinical relevance. This  
<sup>156</sup> application to a well-established genotype-phenotype set, comprising over 500 gene-  
<sup>157</sup> disease associations, underscores its utility. The most recent update on the classification  
<sup>158</sup> of IEI from the International Union of Immunological Societies (IUIS) expert committee  
<sup>159</sup> was reported by Poli et al. (1), with an accompanying diagnostic guide (4).

<sup>160</sup> Quantifying the risk that a patient inherits a disease-causing variant is a fun-  
<sup>161</sup> damental challenge in genomics. Classical statistical approaches grounded in HWE  
<sup>162</sup> (5; 6) have long been used to calculate genetic MOI probabilities for Single Nucleotide  
<sup>163</sup> Variant (SNV)s. However, applying these methods becomes more complex when  
<sup>164</sup> accounting for different MOI, such as Autosomal Recessive (AR) versus Autosomal  
<sup>165</sup> Dominant (AD) or X-Linked (XL) disorders. In AR conditions, for example, the  
<sup>166</sup> occurrence probability must incorporate both the homozygous state and compound  
<sup>167</sup> heterozygosity, whereas for AD and XL disorders, a single pathogenic allele is sufficient  
<sup>168</sup> to cause disease. Advances in genetic research have revealed that MOI can be even

more complex (7). Mechanisms such as dominant negative effects, haploinsufficiency, mosaicism, and digenic or epistatic interactions can further modulate disease risk and clinical presentation, underscoring the need for nuanced approaches in risk estimation. Karczewski et al. (8) made significant advances; however, the remaining challenge lies in applying the necessary statistical genomics data across all MOI for any gene-disease combination, which our current work aims to address. Similar approaches have been reported for diseases such Wilson disease, mucopolysaccharidoses, primary ciliary dyskinesia, and treatable metabolic diseases, (9; 10), as reviewed by Hannah et al. (11).

To our knowledge, all approaches to date have been limited to single MOI, specific to the given disease, or restricted to a small number of genes. We argue that our integrated approach is highly powerful because the resulting probabilities can serve as informative priors in a Bayesian framework for variant and disease probability estimation; a perspective that is often overlooked in clinical and statistical genetics. Such a framework not only refines classical HWE-based risk estimates but also has the potential to enrich clinicians' understanding of what to expect in a patient and to enhance the analytical models employed by bioinformaticians. The dataset also holds value for AI and reinforcement learning applications, providing an enriched version of the data underpinning frameworks such as AlphaFold (12) and AlphaMissense (13).

This gap is not only due to conceptual limitations, but to the historical absence of large, harmonised reference datasets. Only recently have resources become available to support rigorous genome-wide probability estimation. These include high-resolution population allele frequencies (e.g. gnomAD v4 (8)), curated variant classifications (e.g. ClinVar (14)), functional annotations (e.g. UniProt (15)), and pathogenicity prediction models (e.g. AlphaMissense (13)). We previously introduced PanelAppRex to aggregate gene panel data from multiple sources, including Genomics England (GE) PanelApp, ClinVar, and Universal Protein Resource (UniProt), thereby enabling advanced natural searches for clinical and research applications (2; 3; 14; 15). This earlier work relied on expert-curated panels, such as those from the NHS National Genomic Test Directory and the 100,000 Genomes Project, converted into machine-readable formats for rapid variant discovery and interpretation. Together, these resources now make it possible to model the expected distribution of variant types, frequencies, and classifications across the genome.

By reframing variant interpretation as a problem of calibrated expectation rather than solely reactive confirmation, our framework empowers clinicians and researchers to anticipate both observed and unobserved pathogenic burdens. This scalable, genome-wide approach promises to streamline diagnostic workflows, reduce uncertainty in inconclusive cases, inform statistical models and genetic epidemiology studies, and accelerate the integration of genetic insights into patient care.

208 **2 Methods**

209 **2.1 Dataset**

210 Data from Genome Aggregation Database (gnomAD) v4 comprised 807,162 individuals, including 730,947 exomes and 76,215 genomes (8). This dataset provided  
211 786,500,648 SNVs and 122,583,462 Insertion/Deletion (InDel)s, with variant type  
212 counts of 9,643,254 synonymous, 16,412,219 missense, 726,924 nonsense, 1,186,588  
213 frameshift and 542,514 canonical splice site variants. ClinVar data were obtained from  
214 the variant summary dataset (as of: 16 March 2025) available from the NCBI FTP site,  
215 and included 6,845,091 entries, which were processed into 91,319 gene classification  
216 groups and a total of 38,983 gene classifications; for example, the gene *A1BG* contained  
217 four variants classified as likely benign and 102 total entries (14). For our analysis  
218 phase we also used database for Non-Synonymous Functional Predictions (dbNSFP)  
219 which consisted of a number of annotations for 121,832,908 SNVs (16). The Pan-  
220 elAppRex core model contained 58,592 entries consisting of 52 sets of annotations,  
221 including the gene name, disease-gene panel ID, diseases-related features, confidence  
222 measurements. (2) Protein-Protein Interaction (PPI) network data was provided by  
223 Search Tool for the Retrieval of Interacting Genes/Proteins (STRINGdb), consisting  
224 of 19,566 proteins and 505,968 interactions (17). The Human Genome Variation  
225 Society (HGVS) nomenclature is used with Variant Effect Predictor (VEP)-based  
226 codes for variant IDs. AlphaMissense includes pathogenicity prediction classifications  
227 for 71 million variants in 19 thousand human genes (13; 18). We used these scores  
228 to compared against the probability of observing the same given variants. **Box 2.1**  
229 list the definitions from the IUIS IEI for the major disease categories used throughout  
230 this study (1).

232 The following genes were used for disease cohort validations and examples. We  
233 used the two most commonly reported genes from the IEI panel Nuclear Factor Kappa  
234 B Subunit 1 (*NFKB1*) (19–22) and Cystic Fibrosis Transmembrane Conductance  
235 Regulator (*CFTR*) (23–25) to demonstrate applications in AD and AR disease genes,  
236 respectively. We used Severe Combined Immunodeficiency (SCID)-specific genes  
237 AR DNA Cross-Link Repair 1C (*DCLRE1C*), AR Recombination activating gene  
238 1 (*RAG1*), XL Interleukin 2 Receptor Subunit Gamma (*IL2RG*) to demonstrate a  
239 IEI subset disease phenotype of SCID. We also used AD Tumor necrosis factor,  
240 alpha-induced protein 3 (*TNFAIP3*) for other examples comparable to *NFKB1* since it  
241 is also causes AD pro-inflammatory disease but has more known ClinVar classifications  
242 at higher AF then *NFKB1*.

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	

243

244 **2.2 Variant classification occurrence probability**

To quantify the likelihood that an individual harbours a variant with a given disease classification, we compute the variant-level occurrence probability (variant risk estimate (VRE)) for each variant. As a starting point, we considered the classical HWE for a biallelic locus:

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

245 where  $p$  is the allele frequency,  $q = 1 - p$ ,  $p^2$  represents the homozygous dominant,  
 246  $2pq$  the heterozygous, and  $q^2$  the homozygous recessive genotype frequencies. For  
 247 disease phenotypes, particularly under AR MOI, the risk is traditionally linked to  
 248 the homozygous state ( $p^2$ ); however, to account for compound heterozygosity across  
 249 multiple variants, we allocated the overall gene-level risk proportionally among variants.

250 Our computational pipeline estimated the probability of observing a disease-  
 251 associated genotype for each variant and aggregated these probabilities by gene and  
 252 ClinVar classification. This approach included all variant classifications, not limited  
 253 solely to those deemed “pathogenic”, and explicitly conditioned the classification on  
 254 the given phenotype, recognising that a variant could only be considered pathogenic  
 255 relative to a defined clinical context. The core calculations proceeded as follows:

256 **1. Allele frequency and total variant frequency.** For each variant  $i$  in a gene,  
 257 the allele frequency was denoted as  $p_i$ . For each gene (any genomic region or set), we  
 258 defined the total variant frequency (summing across all reported variants in that gene)  
 259 as:

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

260 Note that, because each calculation is confined to one gene, no additional scaling  
 261 was required for our primary analyses ( $P_{\text{tot}}$ ). However, if this same unscaled summation  
 262 is applied across regions or variant sets of differing size or dosage sensitivity, it can  
 263 bias burden estimates. In such cases, normalisation by region length or incorporation  
 264 of gene- or region-specific dosage constraints is recommended.

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

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

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

270 **2. Occurrence probability based on MOI.** The probability that an individual  
 271 is affected by a variant depends on the MOI. For **AD** and **XL** variants, a single  
 272 pathogenic allele suffices:

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

273 For **AR** variants, disease manifests when two pathogenic alleles are present, either  
 274 as homozygotes or as compound heterozygotes. We use:

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

275 Under HWE, the overall gene-level probability of an AR genotype is

$$P_{\text{AR}} = P_{\text{tot}}^2 = \sum_i p_i^2 + 2 \sum_{i < j} p_i p_j,$$

276 where  $P_{\text{tot}} = \sum_i p_i$ . A naïve per-variant assignment

$$p_i^2 + 2 p_i (P_{\text{tot}} - p_i)$$

277 would, when summed over all  $i$ , double-count the compound heterozygous terms. To  
 278 partition  $P_{\text{AR}}$  among variants without double counting, we allocate risk in proportion  
 279 to each variant's allele frequency:

$$p_{\text{disease},i} = \frac{p_i}{P_{\text{tot}}} \times P_{\text{tot}}^2 = p_i P_{\text{tot}}.$$

280 This ensures

$$\sum_i p_{\text{disease},i} = \sum_i p_i P_{\text{tot}} = P_{\text{tot}}^2,$$

281 recovering the correct AR risk while attributing each variant its fair share of  
282 homozygous and compound-heterozygous contributions.

283 More simply, for AD or XL conditions a single pathogenic allele suffices, so the  
284 classification risk (e.g. benign, pathogenic) equals its population frequency. For AR  
285 conditions two pathogenic alleles are required - either two copies of the same variant  
286 or one copy each of two different variants, so we divide the overall recessive risk among  
287 variants according to each variant's share of the total classification frequency in that  
288 gene.

289 **3. Expected case numbers and case detection probability.** Given a popu-  
290 lation with  $N$  births (e.g. as seen in our validation studies,  $N = 69\,433\,632$ ), the  
291 expected number of cases attributable to variant  $i$  was calculated as:

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

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

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

294 **4. Aggregation by gene and ClinVar classification.** For each gene and for each  
295 ClinVar classification (e.g. “Pathogenic”, “Likely pathogenic”, “Uncertain significance”,  
296 etc.), we aggregated the results across all variants. The classification grouping can  
297 be substituted by any alternative score system. The total expected cases for a given  
298 group was:

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

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

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

301 **5. Data processing and implementation.** We implemented the calculations  
302 within a High-Performance Computing (HPC) pipeline and provided an example for a  
303 single dominant disease gene, *TNFAIP3*, in the source code to enhance reproducibility.  
304 Variant data were imported in chunks from the annotation database for all chromosomes  
305 (1-22, X, Y, M).

306 For each data chunk, the relevant fields were gene name, position, allele number,  
307 allele frequency, ClinVar classification, and HGVS annotations. Missing classifications  
308 (denoted by “.”) were replaced with zeros and allele frequencies were converted to  
309 numeric values. Subsequently, the variant data were merged with gene panel data from  
310 PanelAppRex to obtain the disease-related MOI mode for each gene. For each gene, if  
311 no variant was observed for a given ClinVar classification (i.e.  $p_i = 0$ ), a minimal risk  
312 was assigned as described above. Finally, we computed the occurrence probability,  
313 expected cases, and the probability of observing at least one case of disease using the  
314 equations presented.

315 The final results were aggregated by gene and ClinVar classification and used to  
316 generate summary statistics that reviewed the predicted disease observation probabili-  
317 ties. We define the *VRE* as the prior probability of observing a variant classified as  
318 the cause of disease

319 **6. Score-positive-total.** For use as a simple summary statistic on the resulting  
320 user-interface, we defined the *score-positive-total* as the total number of positively  
321 scored variant classifications within a given region (gene, locus, or variant set). Using  
322 the ClinVar classification assigned to a scale from -5 (benign) to +5 (pathogenic),  
323 we included only scores  $> 0$ , corresponding to some evidence of pathogenicity. The  
324 score-positive-total yields a non-normalised estimate of the prior probability that a  
325 phenotype is explained by known pathogenic variants.

326 **7. Classification scoring system.** Each ClinVar classification was assigned an  
327 integer score: pathogenic = +5, likely pathogenic = +4, pathogenic (low penetrance)  
328 = +3, likely pathogenic (low penetrance) = +2, conflicting pathogenicity = +2,  
329 likely risk allele/risk factor/association = +1, drug response/uncertain significance/no  
330 classification/affects/other/not provided/uncertain risk allele = 0, protective = -3,  
331 likely benign = -4, benign = -5. No further normalisation was applied. The resulting  
332 distribution (**Figure S1 A-B**) is naturally comparable to a zero-centred average  
333 rank (**C-D**). This straightforward, modular approach can be readily replaced by

any comparable evidence-based classification system. Variants with scores  $\leq 0$  were omitted, since benign classifications do not inform disease likelihood in the score-positive-total summary.

### 2.3 Integrating observed true positives and unobserved false negatives into a single, actionable conclusion

In this section, we detail our approach to integrating sequencing data with prior classification evidence (e.g. pathogenic on ClinVar) to calculate the posterior probability of a complete successful genetic diagnosis. Our method is designed to account for possible outcomes of TP, TN, and FN, by first ensuring that all nucleotides corresponding to known variant classifications (benign, pathogenic, etc.) have been accurately sequenced. This implies the use of genomic variant call format (gVCF)-style data which refer to variant call format (VCF)s that contain a record for every position in the genome (or interval of interest) regardless of whether a variant was detected at that site or not. Only after confirming that these positions match the reference alleles (or novel unaccounted variants are classified) do we calculate the probability that additional, alternative pathogenic variants (those not observed in the sequencing data) could be present. Our Credible Interval (CrI) for pathogenicity thus incorporates uncertainty from the entire process, including the tally of TP, TN, and FN outcomes. We ignore the contribution of FPs as a separate task to be tackled in the future.

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

**(i) Data pre-processing.** We synthesized an example patient in a disease cohort of 200 cases. We made several scenarios where a causal genetic diagnosis based on the available data is either simple, difficult, or impossible. Our example focused on a proband two representative genes for AD IEI: *NFKB1* and *TNFAIP3*. All coding and canonical splice-region variants for *NFKB1* were extracted from the gVCF. We assumed a typical Quality Control (QC) scenario, where sites corresponding to previously reported pathogenic alleles were checked for read depth  $\geq 10$  and genotype quality  $\geq 20$ . Positions that failed this check were labelled *missing*, thus separating true reference calls from non-sequenced or uninformative sequence.

**(ii) Evidence mapping and occurrence probability.** PanelAppRex variants were annotated with ClinVar clinical significance. Each label was converted to an ordinal evidence score  $S_i \in [-5, 5]$  and rescaled to a pathogenic weight  $W_i = \text{rescale}(S_i; -5, 5 \rightarrow 0, 1)$ . This scoring system can be replaced with any comparable alternative. The HWE-based pipeline of Section 2.2 supplied a per-variant occurrence probability  $p_i$ . The adjusted prior was

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

371 **(iii) Prior specification.** In a hypothetical cohort of  $n = 200$  diploid individuals  
372 the count of allele  $i$  follows a Beta-Binomial model. Marginalising the Binomial yields  
373 the Beta prior

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

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

375 **(iv) Posterior simulation and aggregation.** For each variant  $i$  we drew  $M =$   
376 10 000 realisations  $\pi_i^{(m)}$  and normalised within each iteration,

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

377 Variants with  $S_i > 4$  were deemed to have evidence as *causal* (pathogenic or likely  
378 pathogenic). We note that an alternative evidence score or conditional threshold can  
379 be substituted for this step. Their mean posterior share  $\bar{\pi}_i = M^{-1} \sum_m \tilde{\pi}_i^{(m)}$  and 95%  
380 CrI were retained. The probability that a damaging causal allele is physically present  
381 was obtained by a second layer:

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

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

385 **(v) Scenario analysis.** The three scenarios were explored for a causal genetic  
386 diagnosis that is either simple, difficult, or impossible given the existing data. The  
387 proband spiked data had either: (1) known classified variants only, including only one  
388 known TP pathogenic variant, *NFKB1* p.Ser237Ter, (2) inclusion of an additional  
389 plausible yet non-sequenced splice-donor allele *NFKB1* c.159+1G>A (likely pathogenic)  
390 as a FN, and (3) where no known causal variants were present for a patient, one rep-  
391 resentative variant from each distinct ClinVar classification was selected and marked  
392 as unsequenced to emulate a range of putative FNs. The selected variants were:

393 *TNFAIP3* p.Cys243Arg (pathogenic), p.Tyr246Ter (likely pathogenic), p.His646Pro  
394 (conflicting interpretations of pathogenicity), p.Thr635Ile (uncertain significance),  
395 p.Arg162Trp (not provided), p.Arg280Trp (likely benign), p.Ile207Leu (benign/-  
396 likely benign), and p.Lys304Glu (benign). All subsequent steps were identical.

## 397 2.4 Validation of autosomal dominant estimates using *NFKB1*

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

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

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

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

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

412 **3. National estimate based on literature.** Based on literature (19; 20; 22), the  
413 prevalence of CVID in the general population was estimated as

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

414 For a UK population of  $N_{\text{UK}} \approx 69\,433\,632$ , the expected total number of CVID  
415 cases was

$$416 E_{\text{CVID}} \approx \frac{69\,433\,632}{25\,000} \approx 2777.$$

417 Assuming that the proportion of CVID cases attributable to *NFKB1* is equivalent to  
418 the cohort estimate, the literature extrapolated estimate is Estimated *NFKB1* cases  $\approx$   
419  $2777 \times 0.041 \approx 114$ , with a median value of approximately 118 and a 95% confidence  
420 interval of 70 to 181 cases (derived from posterior sampling).

421    **4. Bayesian adjustment.** Recognising that the sequenced cohort cases likely  
422    captures the majority of *NFKB1*-related patients (apart from close relatives), but may  
423    still miss rare or geographically dispersed variants, we combined the cohort-based and  
424    literature-based estimates using two complementary Bayesian approaches:

- 425    1. **Weighted adjustment (emphasising the cohort,  $w = 0.9$ ):** We assigned  
426    90% weight to the directly observed cohort count (16) and 10% to the extrapolated  
427    population estimate (114), thereby accounting, illustratively, for a small fraction of unobserved cases while retaining confidence in our well-characterised cohort:  
429

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

430    yielding a 95% CrI of roughly 21 to 33 cases.

- 431    2. **Mixture adjustment (equal weighting,  $w = 0.5$ ):** To reflect greater uncertainty about how representative the cohort is, we combined cohort and population prevalences equally. We sampled from the posterior distribution of the cohort prevalence,  
434

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

435    and mixed this with the literature-based rate at 50% each (19; 20; 22). This  
436    yields a median estimate of 67 cases and a wider 95% CrI of approximately  
437    43 to 99 cases, capturing uncertainty in both under-ascertainment and over-generalisation.  
438

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

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

449 **2.5 Validation study for autosomal recessive CF using *CFTR***

450 To validate our framework for AR diseases, we focused on Cystic Fibrosis (CF). For  
 451 comparability sizes between the validation studies, we analysed the most common SNV  
 452 in the *CFTR* gene, typically reported as p.Arg117His (GRCh38 Chr 7:117530975  
 453 G/A, MANE Select HGVSp ENST00000003084.11: p.Arg117His). Our goal was to  
 454 validate our genome-wide probability estimates by comparing the expected number  
 455 of CF cases attributable to the p.Arg117His variant in *CFTR* with the nationally  
 456 reported case count in a well-characterised disease cohort (23–25).

**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:

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

457 **2. Mortality adjustment.** Since CF patients experience increased mortality, we  
 458 adjusted the expected genotype counts using an exponential survival model (23–25).  
 459 With an annual mortality rate  $\lambda \approx 0.004$  and a median age of 22 years, the survival  
 460 factor was computed as:

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

461 Thus, the mortality-adjusted expected genotype count became:

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

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

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

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

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

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

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

475 **2.6 Validation of SCID-specific estimates using PID–SCID  
476 genes**

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

481 **1. Incidence conversion.** Based on literature, PID occurs in approximately 1 in  
482 1,000 births, whereas SCID occurs in approximately 1 in 100,000 births. Consequently,  
483 in a population of 1,000,000 births there are about 1,000 PID cases and 10 SCID cases.  
484 To express SCID-related variant counts on a per-million PID scale, the observed SCID  
485 counts were multiplied by 100. For example, if a gene is expected to cause SCID in 10  
486 cases within the total PID population, then on a per-million PID basis the count is  
487  $10 \times 100 = 1,000$  cases (across all relevant genes).

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

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

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

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

## 510 **2.7 Protein network and genetic constraint interpretation**

511 A PPI network was constructed using protein interaction data from STRINGdb (17).  
512 We previously prepared and reported on this dataset consisting of 19,566 proteins and  
513 505,968 interactions (<https://github.com/DylanLawless/ProteoMCLustR>). Node  
514 attributes were derived from log-transformed score-positive-total values, which in-  
515 formed both node size and colour. Top-scoring nodes (top 15 based on score) were  
516 labelled to highlight prominent interactions. To evaluate group differences in score-  
517 positive-total across major disease categories, one-way Analysis of Variance (ANOVA)  
518 was performed followed by Tukey Honestly Significant Difference (HSD) post hoc  
519 tests (and non-parametric Dunn's test for confirmation). GnomAD v4.1 constraint  
520 metrics data was used for the PPI analysis and was sourced from Karczewski et al.  
521 (8). This provided transcript-level metrics, such as observed/expected ratios, Loss-Of-  
522 function Observed/Expected Upper bound Fraction (LOEUF), Probability of being  
523 Loss-of-function Intolerant (pLI), and Z-scores, quantifying Loss-of-Function (LOF)  
524 and missense intolerance, along with confidence intervals and related annotations for  
525 211,523 observations.

526 **2.8 Gene set enrichment test**

527 To test for overrepresentation of biological functions, the prioritised genes were  
528 compared against gene sets from MsigDB (including hallmark, positional, curated,  
529 motif, computational, GO, oncogenic, and immunologic signatures) and WikiPathways  
530 using hypergeometric tests with FUMA (26; 27). The background set consisted of  
531 24,304 genes. Multiple testing correction was applied per data source using the  
532 Benjamini-Hochberg method, and gene sets with an adjusted P-value  $\leq 0.05$  and more  
533 than one overlapping gene are reported.

534 **2.9 Deriving novel PID classifications by genetic PPI and  
535 clinical features**

536 We recategorised 315 immunophenotypic features from the original IUIS IEI annotations,  
537 reducing the original multi-level descriptors (e.g. “decreased CD8, normal or  
538 decreased CD4”) first to minimal labels (e.g.“low”) and second to binary outcomes  
539 (normal vs. not-normal) for T cells, B cells, neutrophils, and immunoglobulins. Each  
540 gene was mapped to its PPI cluster derived from STRINGdb and Uniform Manifold  
541 Approximation and Projection (UMAP) embeddings from previous steps. We first  
542 tested for non-random associations between these four binary immunophenotypes  
543 and PPI clusters using  $\chi^2$  tests. To generate a data-driven PID classification, we  
544 trained a decision tree (rpart) to predict PPI cluster membership from the four  
545 immunophenotypic features plus the traditional IUIS Major and Subcategory la-  
546 bels. Hyperparameters (complexity parameter=0.001, minimum split=10, minimum  
547 bucket=5, maximum depth=30) were optimised via five-fold cross validation using  
548 the caret framework. Terminal node assignments were then relabelled according to  
549 each group’s predominant abnormal feature profile.

550 **2.10 Probability of observing AlphaMissense pathogenicity**

551 We obtained the subset pathogenicity predictions from AlphaMissense via the Al-  
552 phaFold database and whole genome data from the studies data repository(13; 18). The  
553 AlphaMissense data (genome-aligned and amino acid substitutions) were merged with  
554 the panel variants based on genomic coordinate and HGVS annotation. Occurrence  
555 probabilities were log-transformed and adjusted (y-axis displaying  $\log_{10}(\text{occurrence}$   
556 prob + 1e-5) + 5), to visualise the distribution of pathogenicity scores across the  
557 residue sequence. A Kruskal-Wallis test was used to compare the observed disease  
558 probability across clinical classification groups.

559 **2.11 Probability model definitions**

Estimating disease risk requires accounting for both variant penetrance,  $P(D | G)$ , where  $D$  denotes the disease state and  $G$  the genotype, and the fraction of cases

attributable to a given variant,  $P(G | D)$ . In a fully penetrant single-variant model ( $P(D | G) = 1$ ), the lifetime risk  $P(D)$  equals the genotype frequency  $P(G)$ . For an allele with population frequency  $p$ , this gives  $P(D) = p^2$  for a recessive mode of inheritance and  $P(D) = 2p(1 - p) \approx 2p$  for a dominant mode. With incomplete penetrance,  $P(D) = P(G) P(D | G)$ , and when multiple variants contribute to disease,

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

Because both  $P(D | G)$  and  $P(G | D)$  are often uncertain, we integrate ClinVar clinical classifications, population allele frequencies and curated gene–disease associations, assuming James-Stein shrinkage to derive robust aggregate priors. By focusing on a filtered set of variants  $\mathcal{V}$  where each  $P(G_i | D)$  is the probability that disease  $D$  is attributable to variant  $i$  and assuming  $\sum_{i \in \mathcal{V}} P(G_i | D) \approx 1$ , we obtain calibrated estimates of genotype frequency  $P(G)$  despite uncertainty in individual parameters.

566 

### 3 Results

567 

#### 3.1 Occurrence probability across disease genes

568 Our study integrated large-scale annotation databases with gene panels from PanelAp-  
569 pRex to systematically assess disease genes by MOI (2). By combining population  
570 allele frequencies with ClinVar clinical classifications, we computed an expected occur-  
571 rence probability for each SNV, representing the likelihood of encountering a variant  
572 of a specific pathogenicity for a given phenotype. We report these probabilities for  
573 54,814 ClinVar variant classifications across 557 genes (linked dataset (28)).

574 We focused on panels related to Primary Immunodeficiency or Monogenic In-  
575 flammatory Bowel Disease, using PanelAppRex panel ID 398. **Figure 1** displays all  
576 reported ClinVar variant classifications for this panel. The resulting natural scaling  
577 system (-5 to +5) accounts for the frequently encountered combinations of classification  
578 labels (e.g. benign to pathogenic). The resulting dataset (28) is briefly shown in  
579 **Table 1** to illustrate that our method yielded estimates of the probability of observing  
580 a variant with a particular ClinVar classification.

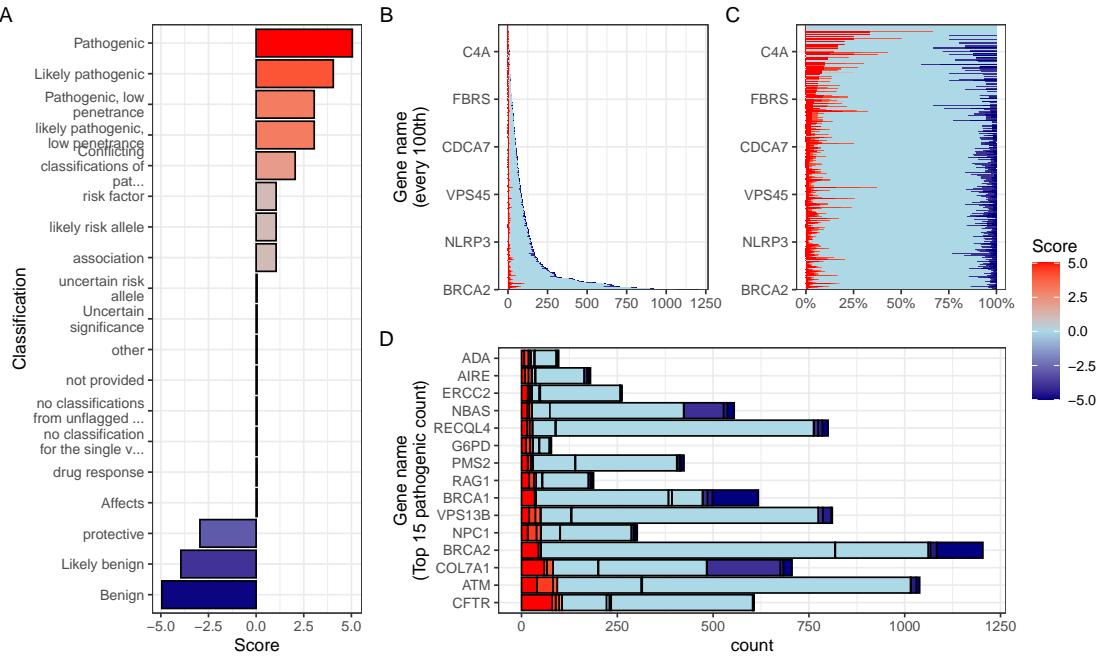
Table 1: Example of the first several rows from our main results for 557 genes of PanelAppRex’s panel: (ID 398) Primary immunodeficiency or monogenic inflammatory bowel disease. “ClinVar Significance” indicates the pathogenicity classification assigned by ClinVar, while “Occurrence Prob” represents our calculated probability of observing the corresponding variant class for a given phenotype. MOI shows the gene-disease-specific mode of inheritance. Additional columns, such as population allele frequency, are not shown. (28)

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

581 

#### 3.2 Integrating observed true positives and unobserved false 582 negatives into a single, actionable conclusion

583 Having established a probabilistic framework for estimating the prior probability  
584 of observing disease-associated variants under different inheritance modes, we then  
585 applied this model to an example patient to demonstrate its potential for clinical  
586 genetics. The algorithm first verified that all known pathogenic positions have been  
587 sequenced and observed as reference (true negatives), and identified any positions  
588 that were either observed as variant (true positives) or not assessable due to missing



**Figure 1: Summary of ClinVar clinical significance classifications in the PID gene panel.** (A) Shows the numeric score coding for each classification (i.e. -5 benign to +5 pathogenic) as defined in methods Section 2.2. (B) Displays the stacked absolute count of classifications per gene. The same counts are shown as percentages per gene in (C). (D) For demonstration, the top 15 genes ranked by absolute count of pathogenic (score 5) variant classifications, indicating those most frequently occurring in the population as disease-causing.

589 sequence data of failed QC. These missing sites represent potential false negatives. By  
 590 jointly modelling the observed and unobserved space, the method yielded a calibrated,  
 591 evidence-weighted probability that at least one damaging causal variant could be  
 592 present within a gene.

### 593 3.3 Scenario one - simple diagnosis

594 We present the results from three scenarios for an example single-case patient being  
 595 investigated for the genetic diagnosis of IEI. **Figure S2** shows the results of the first  
 596 simple scenario, in which only one known pathogenic variant, *NFKB1* p.Ser237Ter,  
 597 was observed and all other previously reported pathogenic positions were successfully  
 598 sequenced and confirmed as reference. In this setting, the model assigned the full  
 599 posterior probability to the observed allele, yielding 100 % confidence that all present  
 600 evidence supported a single, true positive causal explanation. The most strongly  
 601 supported observed variant was p.Ser237Ter (posterior: 0.594). The strongest  
 602 (probability of observing) non-sequenced variant was a benign variant p.Thr567Ile  
 603 (posterior: 0). The total probability of a causal diagnosis given the available evidence

604 was 1 (95% CI: 1–1) (**Table S1**).

### 605 3.4 Scenario two - complex diagnosis

606 **Figure 2** shows the second more complex scenario, where the same pathogenic variant  
607 *NFKB1* p.Ser237Ter was present, but coverage was incomplete at three additional  
608 sites of known classified variants. Among these was the likely-pathogenic splice-site  
609 variant *NFKB1* c.159+1G>A, which was not captured in the sequencing data. The  
610 panels of **Figure 2** (A–F) illustrate the stepwise integration of observed and missing  
611 evidence, culminating in a posterior probability that reflects both confirmed findings  
612 and residual uncertainty. **Table 2** demonstrates our main goal and lists the final  
613 conclusion for reporting the clinical genetics results. **Table S2** lists the main metrics  
614 used to reach the conclusion (as illustrated in **Figure 2**).

615 Bayesian integration of every annotated allele yielded the quantitative CrIs for  
616 pathogenic attribution that (i) preserve Hardy-Weinberg expectations, (ii) accommo-  
617 date AD, AR, XL inheritance, and (iii) carry explicit uncertainty for non-sequenced  
618 (or failed QC) genomic positions. **Figure 2** (A) depicts the prior landscape where  
619 occurrence probabilities are partitioned by observed or missing status and by causal  
620 or non-causal evidence, with colour reflecting the underlying ClinVar score. **Figure**  
621 **2** (B) shows posterior normalisation which concentrates probability density on two  
622 high-confidence (high evidence score) alleles since the benign variants are, by definition,  
623 non-causal. **Figure 2** (C) shows the resulting per-variant probability of being simul-  
624 taneously damaging and causal; only the confirmed present (true positive) nonsense  
625 variant p.Ser237Ter and the (false negative) hypothetical splice-donor c.159+1G>A  
626 retain substantial support. Restricting the view to causal candidates in **Figure 2**  
627 (D) confirms that posterior mass is distributed across these two variants. **Figure 2**  
628 (E) decomposes the total damaging probability into observed (approximately 40 %)  
629 and missing (approximately 34 %) sources, whereas **Figure 2** (F) summarises the  
630 gene-level posterior: inclusion of the splice-site allele (scenario 2) produces a median  
631 probability of 0.542 with a 95 % CrI of 0.264–0.8.

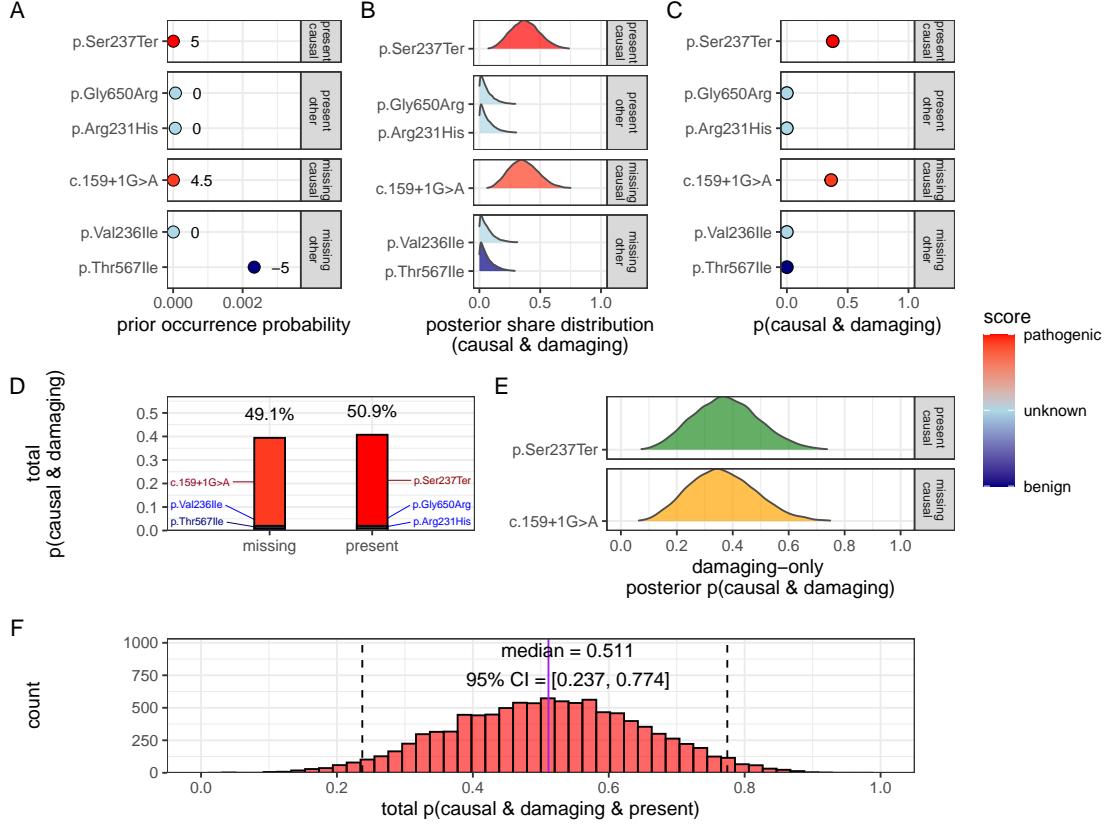
632 Numerically, the present variant p.Ser237Ter accounts for 0.399 of the posterior  
633 share, and the potentially causal but missing splice-donor allele c.159+1G>A contributes  
634 0.339. The remaining alleles together explain a negligible share (**Table S2**). Thus, we  
635 can report that in this patient’s scenario the probability of correct genetic diagnosis  
636 due to *NFKB1* p.Ser237Ter is 0.542 (95 % CrI of 0.264–0.8) given that a likely  
637 alternative remains to be confirmed for this patient. Upon confirmation that the  
638 second variant is not present, the confidence will rise to 1 (95 % CrI of 1–1) as shown  
639 in scenario one.

Table 2: Final variant report for clinical genetics scenario 2. Reported causal: p.Ser237Ter (posterior 0.377). Undetected causal: c.159+1G>A (posterior 0.364). The total probability of a causal diagnosis given the available evidence was 0.511 (95% CI: 0.237–0.774).

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

### 640 3.5 Scenario three - currently impossible diagnosis

641 **Figure S3** shows the third scenario, in which no observed variants were detected in the  
 642 proband for *NFKB1*. Instead, a broad range of plausible FN were detected as missing  
 643 for the gene *TNFAIP3*. The strongest (probability of observing and pathogenic) of  
 644 these non-sequenced variants was p.Cys243Arg (posterior: 0.347). However, the total  
 645 probability of a causal diagnosis for the patient *given the available evidence* was 0  
 646 (95% CI: 0–0) since these missing variants must be accounted for (**Table S3**). Upon  
 647 confirmation, these probabilities can update, as per scenario two.

Gene: *NFKB1*

**Figure 2: Quantification of present (TP) and missing (FN) causal genetic variants for disease in *NFKB1* (scenario 2).** The example proband carried three known heterozygous variants, including pathogenic p.Ser237Ter, and had incomplete coverage at three additional loci, including likely-pathogenic splice-site variant c.159+1G>A. The sequential steps towards the posterior probability of complete genetic diagnosis are shown: (A) Prior occurrence probabilities, stratified by observed/missing and causal/non-causal status. Pathogenicity scores (-5 to +5) are annotated. (B) Posterior distributions of normalised variant weights  $\tilde{\pi}_i$ . (C) Per-variant posterior probability of being both damaging and causal. (D) Posterior distributions for causal variants only. (E) Decomposition of total pathogenic probability into observed (green) and missing (orange) sources. (F) Gene-level posterior showing the probability that at least one damaging causal allele is present; median 0.54, 95% CrI 0.26-0.80. This result can be compared to scenarios one and three in **Figures S2** and **S3**, respectively.

648 **3.6 Validation studies**

649 **3.6.1 Validation of dominant disease occurrence with *NFKB1***

650 To validate our genome-wide probability estimates for AD disorders, we focused on  
651 *NFKB1*. We used a reference dataset from Tuijnenburg et al. (19), in which whole-  
652 genome sequencing of 846 PID patients identified *NFKB1* as one of the genes most  
653 strongly associated with the disease, with 16 *NFKB1*-related CVID cases attributed to  
654 AD heterozygous variants. Our goal was to compare the predicted number of *NFKB1*-  
655 related CVID cases with the reported count in this well-characterised national-scale  
656 cohort.

657 Our model calculated 0 known pathogenic variant *NFKB1*-related CVID cases  
658 in the UK with a minimal risk of 456 unknown de novo variants. In the reference  
659 cohort, 16 *NFKB1* CVID cases were reported. We additionally wanted to account for  
660 potential under-reporting in the reference study. We used an extrapolated national  
661 CVID prevalence which yielded a median estimate of 118 cases (95% CI: 70–181), while  
662 a Bayesian-adjusted mixture estimate produced a median of 67 cases (95% CI: 43–99).  
663 **Figure S5 (A)** illustrates that our predicted values reflect these ranges and are closer  
664 to the observed count. This case supports the validity of our integrated probability  
665 estimation framework for AD disorders, and represents a challenging example where  
666 pathogenic SNV are not reported in the reference population of gnomAD. Our min-max  
667 values successfully contained the true reported values.

668 **3.6.2 Validation of recessive disease occurrence with *CFTR***

669 Our analysis predicted the number of CF cases attributable to carriage of the  
670 p.Arg117His variant (either as homozygous or as compound heterozygous with  
671 another pathogenic allele) in the UK. Based on HWE calculations and mortality  
672 adjustments, we predicted approximately 648 cases arising from biallelic variants and  
673 160 cases from homozygous variants, resulting in a total of 808 expected cases.

674 In contrast, the nationally reported number of CF cases was 714, as recorded  
675 in the UK Cystic Fibrosis Registry 2023 Annual Data Report (23). To account for  
676 factors such as reduced penetrance and the mortality-adjusted expected genotype, we  
677 derived a Bayesian-adjusted estimate via posterior simulation. Our Bayesian approach  
678 yielded a median estimate of 740 cases (95% Confidence Interval (CI): 696, 786) and  
679 a mixture-based estimate of 727 cases (95% CI: 705, 750). **Figure S5 (B)** illustrates  
680 the close concordance between the predicted values, the Bayesian-adjusted estimates,  
681 and the national report supports the validity of our approach for estimating disease.

682 **Figure S4** shows the final values for these genes of interest in a given population  
683 size and phenotype. It reveals that an allele frequency threshold of approximately  
684 0.000007 is required to observe a single heterozygous carrier of a disease-causing  
685 variant in the UK population for both genes. However, owing to the AR MOI pattern  
686 of *CFTR*, this threshold translates into more than 100,000 heterozygous carriers,

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

### 692 3.6.3 Interpretation of ClinVar variant occurrences

693 **Figure S6** shows the two validation study PID genes, representing AR and dominant  
694 MOI. **Figure S6 (A)** illustrates the overall probability of an affected birth by ClinVar  
695 variant classification, whereas **Figure S6 (B)** depicts the total expected number of  
696 cases per classification for an example population, here the UK, of approximately 69.4  
697 million.

### 698 3.6.4 Validation of SCID-specific disease occurrence

699 Given that SCID is a subset of PID, our probability estimates reflect the likelihood of  
700 observing a genetic variant as a diagnosis when the phenotype is PID. However, we  
701 additionally tested our results against SCID cohorts in **Figure S8**. The summarised  
702 raw cohort data for SCID-specific gene counts are summarised and compared across  
703 countries in **Figure S7**. True counts for *IL2RG* and *DCLRE1C* from ten distinct  
704 locations yielded 95% CI surrounding our predicted values. For *IL2RG*, the prediction  
705 was low (approximately 1 case per 1,000,000 PID), as expected since loss-of-function  
706 variants in this XL gene are highly deleterious and rarely observed in gnomAD.  
707 In contrast, the predicted value for *RAG1* was substantially higher (553 cases per  
708 1,000,000 PID) than the observed counts (ranging from 0 to 200). We attributed  
709 this discrepancy to the lower penetrance and higher background frequency of *RAG1*  
710 variants in recessive inheritance, whereby reference studies may underreport the true  
711 national incidence. Overall, we report that agreement within an order of magnitude  
712 was tolerable given the inherent uncertainties from reference studies arising from  
713 variable penetrance and allele frequencies.

## 714 3.7 Genetic constraint in high-impact protein networks

715 We next examined genetic constraint in high-impact protein networks across the whole  
716 IEI gene set of over 500 known disease-gene phenotypes (1). By integrating ClinVar  
717 variant classification scores with PPI data, we quantified the pathogenic burden per  
718 gene and assessed its relationship with network connectivity and genetic constraint  
719 (8; 17).

720 **3.7.1 Score-positive-total within IEI PPI network**

721 The ClinVar classifications reported in **Figure 1** were scaled -5 to +5 based on their  
722 pathogenicity. We were interested in positive (potentially damaging) but not negative  
723 (benign) scoring variants, which are statistically incidental in this analysis. We tallied  
724 gene-level positive scores to give the score-positive-total metric. **Figure S9 (A)** shows  
725 the PPI network of disease-associated genes, where node size and colour encode the  
726 score-positive-total (log-transformed). The top 15 genes with the highest total prior  
727 probabilities of being observed with disease are labelled (as per **Figure 1**).

728 **3.7.2 Association analysis of score-positive-total across IEI categories**

729 We checked for any statistical enrichment in score-positive-totals, which represents the  
730 expected observation of pathogenicity, between the IEI categories. One-way ANOVA  
731 revealed an effect of major disease category on score-positive-total ( $F(8, 500) =$   
732 2.82,  $p = 0.0046$ ), indicating that group means were not identical, which we observed  
733 in **Figure S9 (B)**. However, despite some apparent differences in median scores across  
734 categories (i.e. 9. Bone Marrow Failure (BMF)), the Tukey HSD post hoc comparisons  
735 **Figure S9 (C)** showed that all pairwise differences had 95% CIs overlapping zero,  
736 suggesting that individual group differences were not significant.

737 **3.7.3 UMAP embedding of the PPI network**

738 To address the density of the PPI network for the IEI gene panel, we applied UMAP  
739 (**Figure 3**). Node sizes reflect interaction degree, a measure of evidence-supported  
740 connectivity (17). We tested for a correlation between interaction degree and score-  
741 positive-total. In **Figure 3**, gene names with degrees above the 95th percentile are  
742 labelled in blue, while the top 15 genes by score-positive-total are labelled in yellow  
743 (as per **Figure 1**). Notably, genes with high pathogenic variant loads segregated from  
744 highly connected nodes, suggesting that LOF in hub genes is selectively constrained,  
745 whereas damaging variants in lower-degree genes yield more specific effects. This  
746 observation was subsequently tested empirically.

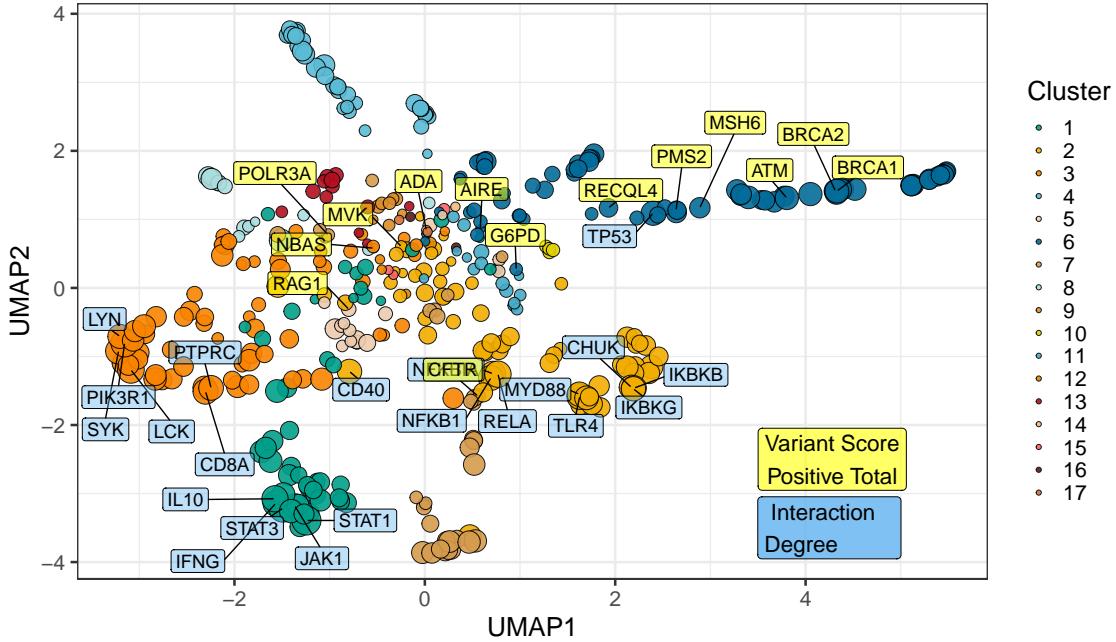


Figure 3: **UMAP embedding of the PPI network.** The plot projects the high-dimensional protein-protein interaction network into two dimensions, with nodes coloured by cluster and sized by interaction degree. Blue labels indicate hub genes (degree above the 95th percentile) and yellow labels mark the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

747 **3.7.4 Hierarchical clustering of enrichment scores for major disease cate-**  
 748 **gories**

749 **Figure S10** presents a heatmap of standardised residuals for major disease categories  
 750 across network clusters, as per **Figure 3**. A dendrogram clusters similar disease  
 751 categories, while the accompanying bar plot displays the maximum absolute standard-  
 752 ised residual for each category. Notably, (8) Complement Deficiencies (CD) shows  
 753 the highest maximum enrichment, followed by (9) BMF. While all maximum values  
 754 exceed 2, the threshold for significance, this likely reflects the presence of protein  
 755 clusters with strong damaging variant scores rather than uniform significance across  
 756 all categories (i.e. genes from cluster 4 in 8 CD).

757 **3.7.5 PPI connectivity, LOEUF constraint and enriched network cluster**  
 758 **analysis**

759 Based on the preliminary insight from **Figure S10**, we evaluated the relationship  
 760 between network connectivity (PPI degree) and LOEUF constraint (LOEUF upper rank)  
 761 Karczewski et al. (8) using Spearman's rank correlation. Overall, there was a weak  
 762 but significant negative correlation ( $\rho = -0.181, p = 0.00024$ ) at the global scale,

763 indicating that highly connected genes tend to be more constrained. A supplementary  
764 analysis (**Figure S11**) did not reveal distinct visual associations between network  
765 clusters and constraint metrics, likely due to the high network density. However  
766 once stratified by gene clusters, the natural biological scenario based on quantitative  
767 PPI evidence (17), some groups showed strong correlations; for instance, cluster 2  
768 ( $\rho = -0.375$ ,  $p = 0.000994$ ) and cluster 4 ( $\rho = -0.800$ ,  $p < 0.000001$ ), while others  
769 did not. This indicated that shared mechanisms within pathway clusters may underpin  
770 genetic constraints, particularly for LOF intolerance. We observe that the score-  
771 positive-total metric effectively summarises the aggregate pathogenic burden across  
772 IEI genes, serving as a robust indicator of genetic constraint and highlighting those  
773 with elevated disease relevance.

774 **Figure S12 (C, D)** shows the re-plotted PPI networks for clusters with significant  
775 correlations between PPI degree and LOEUF upper rank. In these networks, node  
776 size is scaled by a normalised variant score, while node colour reflects the variant score  
777 according to a predefined palette.

### 778 3.8 New insight from functional enrichment

779 To interpret the functional relevance of our prioritised IEI gene sets with the highest  
780 load of damaging variants (i.e. clusters 2 and 4 in **Figure S12**), we performed  
781 functional enrichment analysis for known disease associations using MsigDB with  
782 FUMA (i.e. GWAScatalog and Immunologic Signatures) (26). Composite enrichment  
783 profiles (**Figure S13**) reveal that our enriched PPI clusters were associated with  
784 distinct disease-related phenotypes, providing functional insights beyond traditional  
785 IUIS IEI groupings (1). The gene expression profiles shown in **Figure S14** (GTEx v8  
786 54 tissue types) offer the tissue-specific context for these associations. Together, these  
787 results enable the annotation of IEI gene sets with established disease phenotypes,  
788 supporting a data-driven classification of IEI.

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

798 **3.9 Genome-wide gene distribution and linkage disequilibrium**  
 799

800 **Figure 4 (A)** shows a genome-wide karyoplot of all IEI panel genes across GRCh38,  
 801 with colour-coding based on MOI. Figures (B) and (C) display zoomed-in locus plots  
 802 for *NFKB1* and *CFTR*, respectively. In **Figure 4 (B)**, the probability of observing  
 803 variants with known classifications is high only for variants such as p.Ala475Gly, which  
 804 are considered benign in the AD *NFKB1* gene that is intolerant to LOF. In **Figure 4 (C)**,  
 805 high probabilities of observing patients with pathogenic variants in *CFTR* are  
 806 evident, reproducing this well-established phenomenon. Furthermore, the analysis of  
 807 Linkage Disequilibrium (LD) using  $R^2$  shows that high LD regions can be modelled  
 808 effectively, allowing independent variant signals to be distinguished.

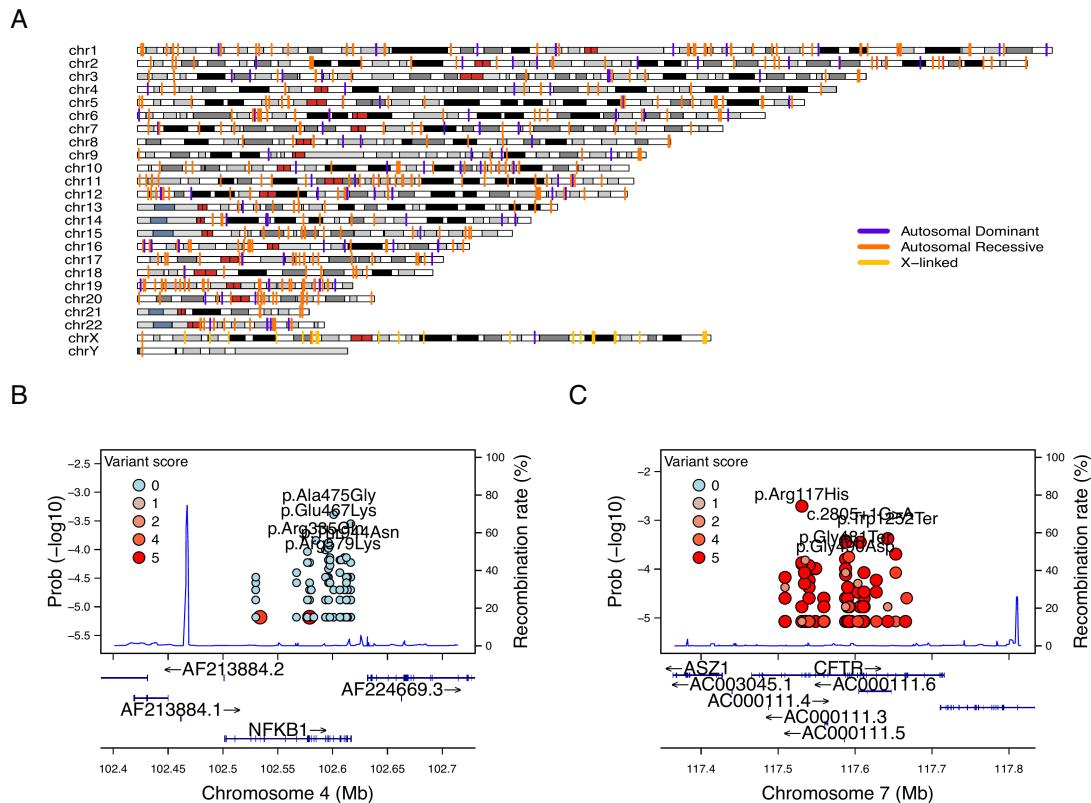


Figure 4: **Genome-wide IEI, variant occurrence probability and LD by  $R^2$ .**  
 (A) Genome-wide karyoplot of all IEI panel genes mapped to GRCh38, with colours  
 indicating MOI. (B) Zoomed-in locus plot example for *NFKB1* showing variant  
 occurrence probabilities; only benign variants such exhibit high probabilities in this  
 AD gene intolerant to LOF. (C) Locus plot example for *CFTR* displaying high  
 probabilities for pathogenic variants; due to the dense clustering of pathogenic variants,  
 score filter  $>0$  was applied. Top five variants are labelled per gene.

809    **3.10 Novel PID classifications derived from genetic PPI and**  
810    **clinical features**

811    We recategorised 315 immunophenotypic features from the original IUIS IEI annotations,  
812    reducing detailed descriptions (e.g. “decreased CD8, normal or decreased  
813    CD4”) to minimal labels (e.g. “low”) and then binarising them (normal vs. not-normal)  
814    for T cells, B cells, Immunoglobulin (Ig) and neutrophils (**Figure S15**). These sim-  
815    plified profiles were mapped onto STRINGdb PPI clusters, revealing non-random  
816    distributions ( $\chi^2 < 1e-13$ ; **Figure S16**), indicating that network context captures key  
817    immunophenotypic variation.

818    We next compared four classifiers under 5-fold cross-validation to determine which  
819    features predicted PPI clustering. As shown in **Figure S17**, the fully combined model  
820    achieved the highest accuracy among the four: (i) phenotypes only (33 %) (i.e. T  
821    cell, B cell, Ig, Neutrophil); (ii) phenotypes+IUIS major category (50 %) (e.g. CID.  
822    See **Box 2.1** for more); (iii) IUIS major+subcategory only (59 %) (e.g. CID, T-B+  
823    SCID); and (iv) phenotypes+IUIS major+subcategory (61 %). This demonstrated that  
824    incorporating both traditional IUIS IEI classifications and core immunophenotypic  
825    markers into the PPI-based framework yielded the most robust discrimination of PID  
826    gene clusters. Variable importance analysis highlighted abnormality status for Ig and  
827    T cells were among the top ten features in addition to the other IUIS major and sub  
828    categories. Per-class specificity remained uniform across the classes while sensitivity  
829    dropped.

830    The PPI and immunophenotype model yielded 17 data-driven PID groups, whereas  
831    incorporating the full complement of IUIS categories expanded this to 33 groups. For  
832    clarity, we only demonstrate the decision tree from the smaller 17-group model in  
833    **Figure 5**. Each terminal node is annotated by its predominant immunophenotypic  
834    signature (for example, “group 65 with abnormal T cell and B cell features”), and  
835    the full resulting gene counts per 33 class are plotted in **Figure 5**. Although, less  
836    user-friendly, this data-driven taxonomy both aligns with and refines traditional IUIS  
837    IEI classifications to provide a scaffold for large-scale computational analyses. Because  
838    this framework is fully reproducible, alternative PPI embeddings that incorporate  
839    additional molecular annotations can readily swapped to continue building on these  
840    IEI classification schemes.

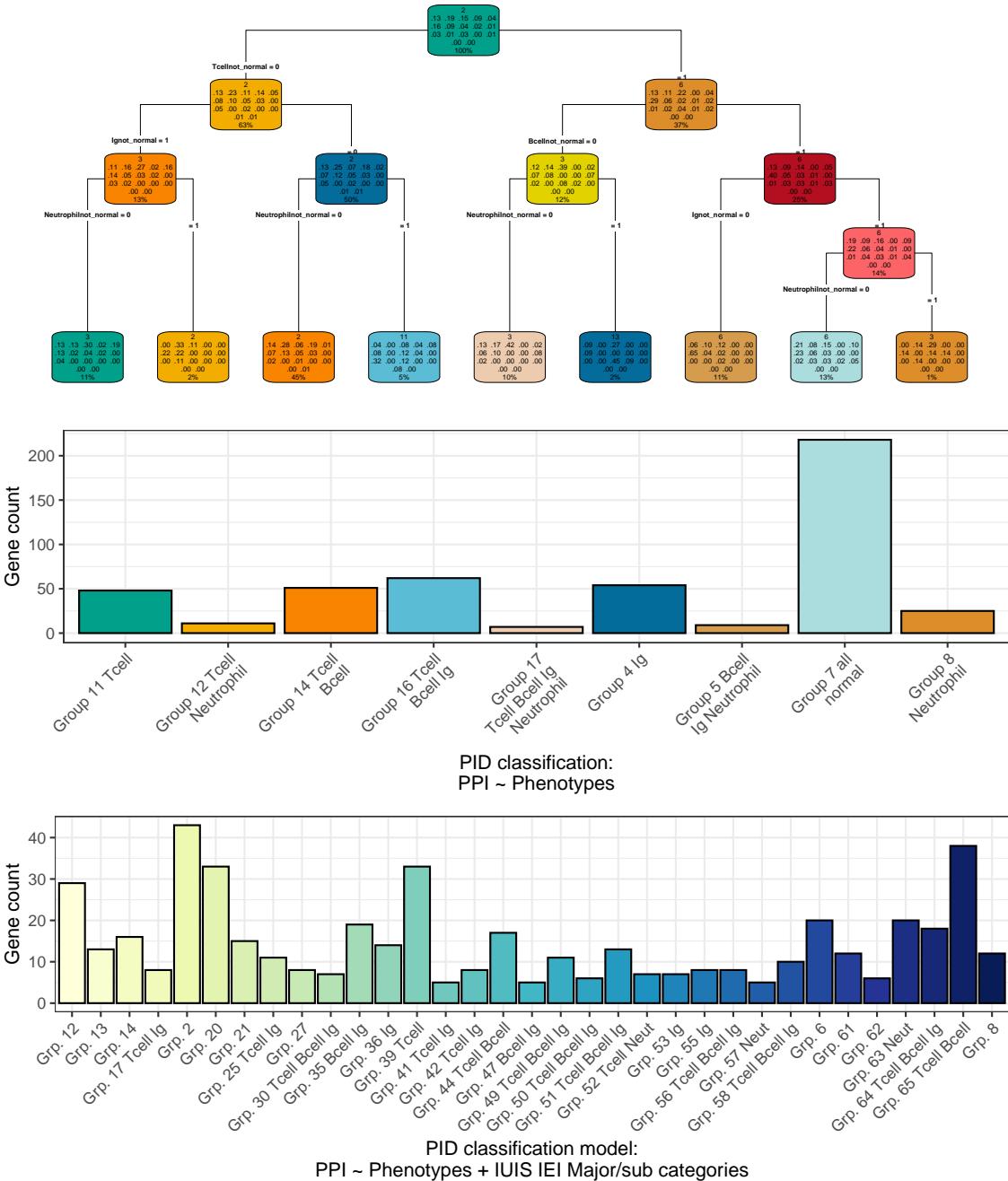
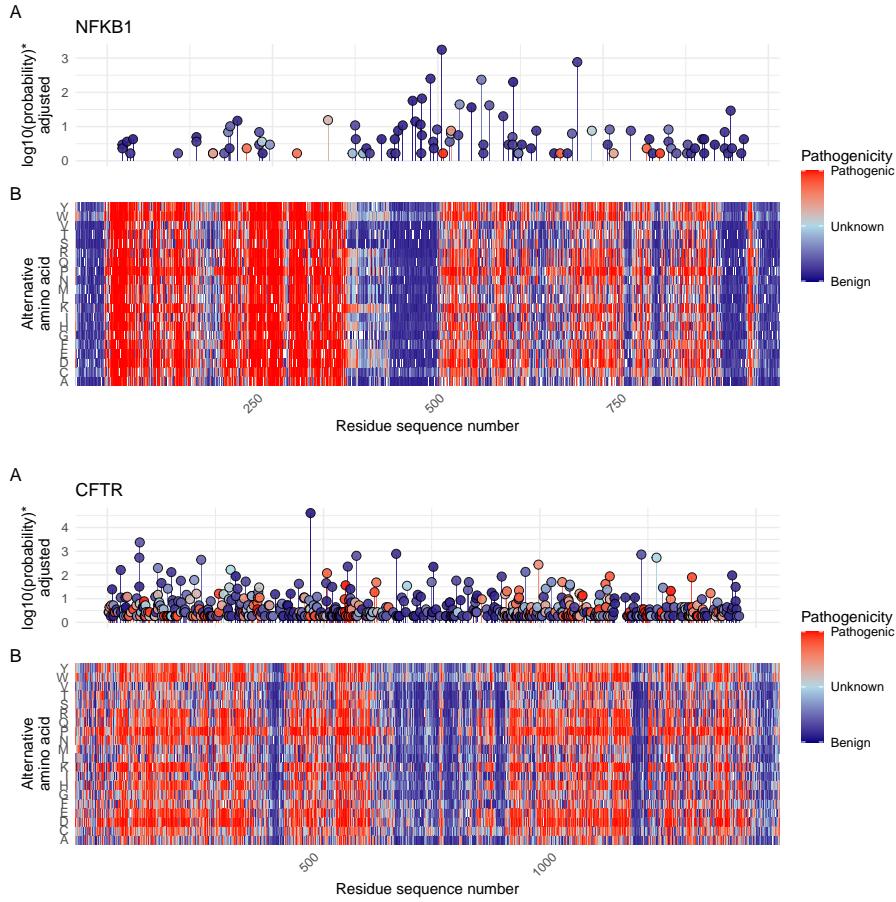


Figure 5: **Fine-tuned model for PID classification.** (Top) In each terminal node, the top block indicates the number of genes in the node; the middle block shows the fitted class probabilities (which sum to 1); and the bottom block displays the percentage of the total sample in that node. These metrics 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.

841 **3.11 Probability of observing AlphaMissense pathogenicity**

842 AlphaMissense provides pathogenicity scores for all possible amino acid substitutions;  
843 however, our results in **Figure 6** show that the most probable observations in patients  
844 occur predominantly for benign or unknown variants. This finding places the likelihood  
845 of disease-associated substitutions into perspective and offers a data-driven foundation  
846 for future improvements in variant prediction. The values in **Figure 6 (A)** can  
847 be directly compared to **Figure 1 (D)** to view the distribution of classifications.  
848 A Kruskal-Wallis test was used to compare the observed disease probability across  
849 clinical classification groups and no significant differences were detected. In general,  
850 most variants in patients are classified as benign or unknown, indicating limited  
851 discriminative power in the current classification, such that pathogenicity prediction  
852 does not infer occurrence prediction (**Figure S18**). Inverse correlation likely depends  
853 on factors like MOI and intolerance to LOF.

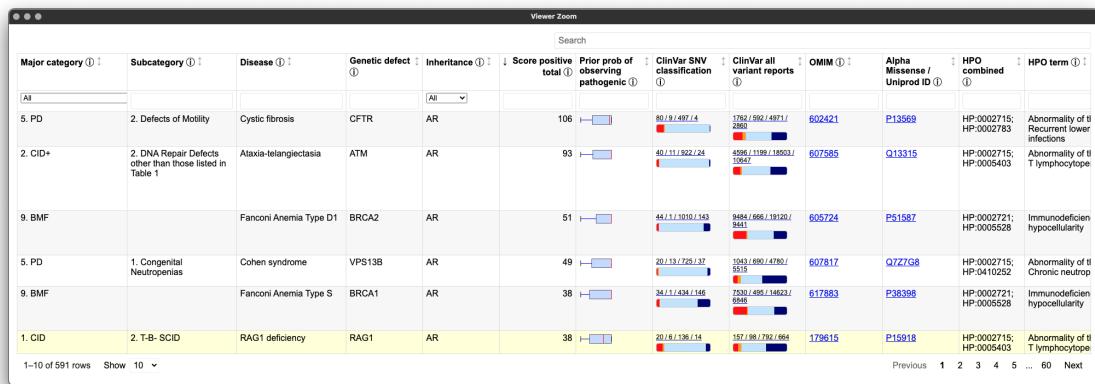


**Figure 6: (A) Probabilities of observing a patient with (B) AlphaMissense-derived pathogenicity scores.** Although AlphaMissense provides scores for every possible amino acid substitution, the most frequently observed variants in patients tend to be classified as benign or of unknown significance. This juxtaposition contextualises the likelihood of disease-associated substitutions and underlines prospects for refining predictive models. \*Axis scaling for visibility near zero. Higher point indicates higher probability.

### 3.12 Integration of variant probabilities into IEI genetics data

We integrated the computed prior probabilities for observing variants in all known genes associated with a given phenotype (1), across AD, AR, and XL MOI, into our IEI genetics framework. These calculations, derived from gene panels in PanelAppRex, have yielded novel insights for the IEI disease panel. The final result comprised of machine- and human-readable datasets, including the table of variant classifications and priors available via a the linked repository (28), and a user-friendly web interface that incorporates these new metrics.

**Figure 7** shows the interface summarising integrated variant data. We include pre-calculated summary statistics and clinical significance as numerical metrics. Key quantiles (min, Q1, median, Q3, max) for each gene are rendered as sparkline box plots, and dynamic URLs link table entries to external databases (e.g. ClinVar, Online Mendelian Inheritance in Man (OMIM), AlphaFold) as per **Section 3.1**. The prepared data are available for bioinformatic application (28) as per **Section 3.2**.



**Figure 7: 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.

868 **4 Discussion**

869 Our study presents, to our knowledge, the first comprehensive framework for calculating  
870 prior probabilities of observing disease-associated variants and the first to demonstrate  
871 the method for an evidence-aware genetic diagnosis with CrI (9; 11). By integrating  
872 large-scale genomic annotations, including population allele frequencies from gnomAD  
873 (8), variant classifications from ClinVar (14), and functional annotations from resources  
874 such as dbNSFP, with classical HWE-based calculations, we derived robust estimates  
875 for 54,814 ClinVar variant classifications across 557 IEI genes implicated in PID and  
876 monogenic inflammatory bowel disease (1; 2). Although our results focus on IEI, the  
877 genome-wide framework also supports all inheritance patterns: AD and XL require a  
878 single pathogenic allele, whereas AR demands homozygous or compound heterozygous  
879 states. Classical HWE-based estimates thus furnish baseline occurrence probabilities  
880 and serve as robust priors for Bayesian risk models, a practice underutilised until the  
881 advent of large-scale databases (2; 8; 13; 14).

882 A major deficit in current clinical genetics is the prevailing focus on confirming  
883 only the presence of TP variants. Our approach yielded three key results to overcome  
884 this hurdle. We generated per-variant priors across all MOI. The patient's results of  
885 observed and unobserved variants were integrated into a single posterior probability  
886 of carrying a damaging causal allele. As demonstrated in **Table S2** and **Figure 2**,  
887 this key result delivers a clinically applicable, interpretable probability that combines  
888 both detected and potentially unobserved variants. When whole-genome sequencing  
889 analyses are not yet available, the score-positive-total metric can serve as an optional  
890 decision aid, enabling manual, evidence-based ranking of candidate genes to prioritise  
891 diagnoses in patients with overlapping phenotypes.

892 We acknowledge that our framework is currently focused (but not restricted) on  
893 SNVs and does not incorporate numerous other complexities of genetic disease, such  
894 as structural variants, de novo variants, hypomorphic alleles, overdominance, variable  
895 penetrance, tissue-specific expression, the Wahlund effect, pleiotropy, and others (7).  
896 In certain applications, more refined estimates would benefit from including factors  
897 such as embryonic lethality, condition-specific penetrance, and age of onset (11). Our  
898 analysis also relies on simplifying assumptions of random mating, an effectively infinite  
899 population, and the absence of migration, novel mutations, or natural selection. We  
900 demonstrated the genome-wide gene distribution and MOI for the IEI panel relative  
901 to LD showing that it is an important consideration and is feasible. However, LD is  
902 a challenging feature that requires accurate implementation which depends on the  
903 whole genome population-based pairwise genotype matrices for the given population.  
904 We used the reference global population AFs, which is more generalisable but less  
905 accurate than population-specific AF values.

906 In the example single-case diagnosis scenarios, our approach enabled high-confidence  
907 attribution to a known pathogenic variant while also capturing the potential impact  
908 of a likely-pathogenic splice-site allele that was missed by sequencing. Scenario two  
909 showed a common diagnostic challenge where a strong candidate exists alongside an

910 unconfirmed but plausible alternative. Our method distributes confidence across both  
 911 possibilities. Conventional approaches focus only on detecting TP and cannot provide  
 912 this insight. By quantifying residual uncertainty, we can generate structured reports  
 913 that clearly distinguish supported, excluded, and plausible-but-unseen variants. We  
 914 call this “evidence-aware” interpretation. When combined with genome-wide priors  
 915 from the full range of disease-gene panels, this approach applies to any phenotype  
 916 from PanelAppRex. By combining variant classification, allele frequency, MOI, and se-  
 917 quencing quality metrics, our method creates a scalable foundation for evidence-aware  
 918 diagnostics in clinical genomics.

919 Estimating disease risk in genetic studies is complicated by uncertainties in key  
 920 parameters such as variant penetrance and the fraction of cases attributable to specific  
 921 variants (7). In the simplest model, where a single, fully penetrant variant causes  
 922 disease, the lifetime risk  $P(D)$  is equivalent to the genotype frequency  $P(G)$ . For an  
 923 allele with frequency  $p$  (ignoring LD for AR), this translates to:

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

924 When penetrance is incomplete, defined as  $P(D | G)$ , the risk becomes:  $P(D) =$   
 925  $P(G) P(D | G)$ . In more realistic scenarios where multiple variants contribute to  
 926 disease,  $P(G | D)$  denotes the fraction of cases attributable to a given variant. This  
 927 leads to:

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

928 Because both penetrance and  $P(G | D)$  are often uncertain, solving this equation  
 929 systematically poses a major challenge, which we incidentally tackled in the validation  
 930 studies (29; 30).

931 Our framework addresses this challenge by combining variant classifications, popu-  
 932 lation allele frequencies, and curated gene-disease associations. While imperfect on  
 933 an individual level, these sources exhibit predictable aggregate behaviour, supported  
 934 by James-Stein estimation principles (31). Curated gene-disease associations help  
 935 identify genes that are explainable for most disease cases, allowing us to approximate  
 936  $P(G | D)$  close to one. In this way, we obtain robust estimates of  $P(G)$  (the frequency  
 937 of disease-associated genotypes), even when exact values of penetrance and case  
 938 attribution remain uncertain.

939 This approach allows us to pre-calculate priors and summarise the overall pathogenic  
 940 burden. By focusing on a subset  $\mathcal{V}$  of variants that pass stringent filtering, where each  
 941  $P(G_i | D)$  is the probability that a case of disease  $D$  is attributable to variant(s)  $i$ ,  
 942 we assume that, in aggregate,

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

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

947 For the IEI-specific investigation, we showed that immunophenotypic and network-  
 948 derived features can be used to train and test models that predict PPIs. From this, we  
 949 derived a new, simplified classification of immune features for IEI genes. We have listed  
 950 the new immunophenotypic categories (e.g. T cell low) in the user database, however  
 951 we have not included the detailed cluster assignments (e.g. PPI groups) because  
 952 they are too complex for direct interpretation manually. Instead, our demonstration  
 953 provides worked examples that bioinformaticians can use to perform more refined  
 954 clustering in larger studies.

955 Moreover, because variant sets can be collapsed instead of relying on the gene-  
 956 level, our method complements existing statistical approaches for aggregating variant  
 957 effects with methods like Sequence Kernel Association Test (SKAT) and Aggregated  
 958 Cauchy Association Test (ACAT) (32–35) and multi-omics integration techniques  
 959 (36; 37). It also remains consistent with established variant interpretation guidelines  
 960 from the American College of Medical Genetics and Genomics (ACMG) (38) and  
 961 complementary frameworks (39; 40), as well as QC protocols (41; 42). Standardised  
 962 reporting for qualifying variant sets, such as ACMG Secondary Findings v3.2 (43),  
 963 further contextualises the integration of these probabilities into clinical decision-  
 964 making.

965 We compared our occurrence probabilities with AlphaMissense pathogenicity scores  
 966 and observed that common variants are predominantly scored as benign or of uncertain  
 967 significance. While this aligns with their allele frequencies, any pathogenic variant  
 968 seen in a patient warrants evaluation against its prior observation probability to assess  
 969 causality. Predictive tools such as AlphaMissense could ostensibly enhance their  
 970 embedding of variant features by incorporating gene-disease associations and MOI  
 971 data, which may not be fully represented by raw population allele frequencies.

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

976 **5 Conclusion**

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

980 **Acknowledgements**

981 We would like to thank all the patients and families who have been providing advice  
982 on SwissPedHealth and its projects, as well as the clinical and research teams at the  
983 participating institutions. We acknowledge Genomics England for providing public  
984 access to the PanelApp data. The use of data from Genomics England panelapp was  
985 licensed under the Apache License 2.0. The use of data from UniProt was licensed  
986 under Creative Commons Attribution 4.0 International (CC BY 4.0). ClinVar asks its  
987 users who distribute or copy data to provide attribution to them as a data source in  
988 publications and websites (14). dbNSFP version 4.4a is licensed under the Creative  
989 Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-  
990 ND 4.0); while we cite this dataset as used our research publication, it is not used  
991 for the final version which instead used ClinVar and gnomAD directly. GnomAD is  
992 licensed under Creative Commons Zero Public Domain Dedication (CC0 1.0 Universal).  
993 GnomAD request that usages cites the gnomAD flagship paper (8) and any online  
994 resources that include the data set provide a link to the browser, and note that tool  
995 includes data from the gnomAD v4.1 release. AlphaMissense asks to cite Cheng et al.  
996 (13) for usage in research, with data available from Cheng et al. (18).

997 **Funding**

998 This project was supported through the grant NDS-2021-911 (SwissPedHealth) from  
999 the Swiss Personalized Health Network and the Strategic Focal Area 'Personalized  
1000 Health and Related Technologies' of the ETH Domain (Swiss Federal Institutes of  
1001 Technology).

1002 **Contributions**

1003 DL performed main analysis and wrote the manuscript. SB, AS, MS, and JT designed  
1004 analysis and wrote the manuscript. JF, LJS supervised the work, and applied for  
1005 funding. The Quant Group members on this project were DL, SB, AS, and MS.

1006 **Competing interest**

1007 We declare no competing interest.

1008 **References**

- 1009 [1] M. Cecilia Poli, Ivona Aksentijevich, Ahmed Aziz Bousfiha, Charlotte  
1010 Cunningham-Rundles, Sophie Hambleton, Christoph Klein, Tomohiro Morio,  
1011 Capucine Picard, Anne Puel, Nima Rezaei, Mikko R.J. Seppänen, Raz Somech, He-  
1012 len C. Su, Kathleen E. Sullivan, Troy R. Torgerson, Isabelle Meyts, and Stuart G.  
1013 Tangye. Human inborn errors of immunity: 2024 update on the classification from  
1014 the International Union of Immunological Societies Expert Committee. *Journal of*  
1015 *Human Immunity*, 1(1):e20250003, May 2025. ISSN 3065-8993. doi: 10.70962/jhi.  
1016 20250003. URL <https://rupress.org/jhi/article/1/1/e20250003/277390/Human-inborn-errors-of-immunity-2024-update-on-the>.
- 1017
- 1018 [2] Dylan Lawless. PanelAppRex aggregates disease gene panels and facilitates  
1019 sophisticated search. March 2025. doi: 10.1101/2025.03.20.25324319. URL  
1020 <http://medrxiv.org/lookup/doi/10.1101/2025.03.20.25324319>.
- 1021 [3] Antonio Rueda Martin, Eleanor Williams, Rebecca E. Foulger, Sarah Leigh,  
1022 Louise C. Daugherty, Olivia Niblock, Ivone U. S. Leong, Katherine R. Smith,  
1023 Oleg Gerasimenko, Eik Haraldsdottir, Ellen Thomas, Richard H. Scott, Emma  
1024 Baple, Arianna Tucci, Helen Brittain, Anna De Burca, Kristina Ibañez, Dalia  
1025 Kasperaviciute, Damian Smedley, Mark Caulfield, Augusto Rendon, and Ellen M.  
1026 McDonagh. PanelApp crowdsources expert knowledge to establish consensus  
1027 diagnostic gene panels. *Nature Genetics*, 51(11):1560–1565, November 2019.  
1028 ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-019-0528-2. URL <https://www.nature.com/articles/s41588-019-0528-2>.
- 1029
- 1030 [4] Ahmed Aziz Bousfiha, Leïla Jeddane, Abderrahmane Moundir, M. Cecilia  
1031 Poli, Ivona Aksentijevich, Charlotte Cunningham-Rundles, Sophie Hambleton,  
1032 Christoph Klein, Tomohiro Morio, Capucine Picard, Anne Puel, Nima Rezaei,  
1033 Mikko R.J. Seppänen, Raz Somech, Helen C. Su, Kathleen E. Sullivan, Troy R.  
1034 Torgerson, Stuart G. Tangye, and Isabelle Meyts. The 2024 update of IUIS  
1035 phenotypic classification of human inborn errors of immunity. *Journal of Hu-  
1036 man Immunity*, 1(1):e20250002, May 2025. ISSN 3065-8993. doi: 10.70962/jhi.  
1037 20250002. URL <https://rupress.org/jhi/article/1/1/e20250002/277374/The-2024-update-of-IUIS-phenotypic-classification>.
- 1038
- 1039 [5] Oliver Mayo. A Century of Hardy–Weinberg Equilibrium. *Twin Research  
1040 and Human Genetics*, 11(3):249–256, June 2008. ISSN 1832-4274, 1839-2628.  
1041 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).
- 1042

- 1043 [6] Nikita Abramovs, Andrew Brass, and May Tassabehji. Hardy-Weinberg Equi-  
1044 librium in the Large Scale Genomic Sequencing Era. *Frontiers in Genetics*,  
1045 11:210, March 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.00210. URL  
1046 <https://www.frontiersin.org/article/10.3389/fgene.2020.00210/full>.
- 1047 [7] Johannes Zschocke, Peter H. Byers, and Andrew O. M. Wilkie. Mendelian  
1048 inheritance revisited: dominance and recessiveness in medical genetics. *Nature  
1049 Reviews Genetics*, 24(7):442–463, July 2023. ISSN 1471-0056, 1471-0064.  
1050 doi: 10.1038/s41576-023-00574-0. URL <https://www.nature.com/articles/s41576-023-00574-0>.
- 1051 [8] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings,  
1052 Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea  
1053 Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified  
1054 from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.
- 1055 [9] Sarah L. Bick, Aparna Nathan, Hannah Park, Robert C. Green, Monica H.  
1056 Wojcik, and Nina B. Gold. Estimating the sensitivity of genomic newborn  
1057 screening for treatable inherited metabolic disorders. *Genetics in Medicine*, 27  
1058 (1):101284, January 2025. ISSN 10983600. doi: 10.1016/j.gim.2024.101284. URL  
1059 <https://linkinghub.elsevier.com/retrieve/pii/S1098360024002181>.
- 1060 [10] Benjamin D. Evans, Piotr Słowiński, Andrew T. Hattersley, Samuel E. Jones, Seth  
1061 Sharp, Robert A. Kimmitt, Michael N. Weedon, Richard A. Oram, Krasimira  
1062 Tsaneva-Atanasova, and Nicholas J. Thomas. Estimating disease prevalence in  
1063 large datasets using genetic risk scores. *Nature Communications*, 12(1):6441,  
1064 November 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-26501-7. URL  
1065 <https://www.nature.com/articles/s41467-021-26501-7>.
- 1066 [11] William B. Hannah, Mitchell L. Drumm, Keith Nykamp, Tiziano Prampano,  
1067 Robert D. Steiner, and Steven J. Schrodi. Using genomic databases to de-  
1068 termine the frequency and population-based heterogeneity of autosomal reces-  
1069 sive conditions. *Genetics in Medicine Open*, 2:101881, 2024. ISSN 29497744.  
1070 doi: 10.1016/j.gimo.2024.101881. URL <https://linkinghub.elsevier.com/retrieve/pii/S2949774424010276>.
- 1071 [12] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov,  
1072 Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna  
1073 Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J.  
1074 Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishabh  
1075 Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy,  
1076 Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer,  
1077 Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray  
1078 Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein  
1079 structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.  
1080 ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03819-2. URL <https://www.nature.com/articles/s41586-021-03819-2>.
- 1081
- 1082
- 1083

- 1084 [13] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor  
1085 Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant,  
1086 Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet  
1087 Kohli, and Žiga Avsec. Accurate proteome-wide missense variant effect prediction  
1088 with AlphaMissense. *Science*, 381(6664):eadg7492, September 2023. ISSN 0036-  
1089 8075, 1095-9203. doi: 10.1126/science.adg7492. URL <https://www.science.org/doi/10.1126/science.adg7492>.
- 1090
- 1091 [14] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao,  
1092 Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee  
1093 Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adri-  
1094 ana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou,  
1095 J Bradley Holmes, Brandi L Kattman, and Donna R Maglott. ClinVar: impro-  
1096 ving access to variant interpretations and supporting evidence. *Nucleic Acids  
1097 Research*, 46(D1):D1062–D1067, January 2018. ISSN 0305-1048, 1362-4962. doi:  
1098 10.1093/nar/gkx1153. URL <http://academic.oup.com/nar/article/46/D1/D1062/4641904>.
- 1099
- 1100 [15] The UniProt Consortium, Alex Bateman, Maria-Jesus Martin, Sandra Orchard,  
1101 Michele Magrane, Aduragbemi Adesina, Shadab Ahmad, Bowler-Barnett, and  
1102 Others. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids  
1103 Research*, 53(D1):D609–D617, January 2025. ISSN 0305-1048, 1362-4962. doi:  
1104 10.1093/nar/gkae1010. URL <https://academic.oup.com/nar/article/53/D1/D609/7902999>.
- 1105
- 1106 [16] Xiaoming Liu, Chang Li, Chengcheng Mou, Yibo Dong, and Yicheng Tu.  
1107 dbNSFP v4: a comprehensive database of transcript-specific functional pre-  
1108 dictions and annotations for human nonsynonymous and splice-site SNVs.  
1109 *Genome Medicine*, 12(1):103, December 2020. ISSN 1756-994X. doi: 10.  
1110 1186/s13073-020-00803-9. URL <https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00803-9>.
- 1111
- 1112 [17] Damian Szklarczyk, Katerina Nastou, Mikaela Koutrouli, Rebecca Kirsch, Farrokh  
1113 Mehryary, Radja Hachilif, Dewei Hu, Matteo E Peluso, Qingyao Huang, Tao  
1114 Fang, et al. The string database in 2025: protein networks with directionality of  
1115 regulation. *Nucleic Acids Research*, 53(D1):D730–D737, 2025.
- 1116
- 1117 [18] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor  
1118 Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant,  
1119 Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet  
1120 Kohli, and Žiga Avsec. Predictions for alphamissense, September 2023. URL  
<https://doi.org/10.5281/zenodo.8208688>.
- 1121
- 1122 [19] Paul Tuijnenburg, Hana Lango Allen, Siobhan O. Burns, Daniel Greene, Machiel H.  
1123 Jansen, and Others. Loss-of-function nuclear factor B subunit 1 (NFKB1) variants  
1124 are the most common monogenic cause of common variable immunodeficiency  
in Europeans. *Journal of Allergy and Clinical Immunology*, 142(4):1285–1296,

- 1125        October 2018. ISSN 00916749. doi: 10.1016/j.jaci.2018.01.039. URL <https://linkinghub.elsevier.com/retrieve/pii/S0091674918302860>.
- 1126
- 1127 [20] WHO Scientific Group et al. Primary immunodeficiency diseases: report of a  
1128        who scientific group. *Clin. Exp. Immunol.*, 109(1):1–28, 1997.
- 1129 [21] Charlotte Cunningham-Rundles and Carol Bodian. Common variable immunode-  
1130        ficiency: clinical and immunological features of 248 patients. *Clinical immunology*,  
1131        92(1):34–48, 1999.
- 1132 [22] Eric Oksenhendler, Laurence Gérard, Claire Fieschi, Marion Malphettes, Gael  
1133        Mouillot, Roland Jaussaud, Jean-François Viallard, Martine Gardembas, Lionel  
1134        Galicier, Nicolas Schleinitz, et al. Infections in 252 patients with common variable  
1135        immunodeficiency. *Clinical Infectious Diseases*, 46(10):1547–1554, 2008.
- 1136 [23] Y Naito, F Adams, S Charman, J Duckers, G Davies, and S Clarke. Uk cystic  
1137        fibrosis registry 2023 annual data report. *London: Cystic Fibrosis Trust*, 2023.
- 1138 [24] Carlo Castellani, CFTR2 team, et al. Cftr2: how will it help care? *Paediatric*  
1139        *respiratory reviews*, 14:2–5, 2013.
- 1140 [25] Hartmut Grasemann and Felix Ratjen. Cystic fibrosis. *New England Journal*  
1141        *of Medicine*, 389(18):1693–1707, 2023. doi: 10.1056/NEJMra2216474. URL  
1142        <https://www.nejm.org/doi/full/10.1056/NEJMra2216474>.
- 1143 [26] Kyoko Watanabe, Erdogan Taskesen, Arjen Van Bochoven, and Danielle  
1144        Posthuma. Functional mapping and annotation of genetic associations with  
1145        FUMA. *Nature Communications*, 8(1):1826, November 2017. ISSN 2041-1723.  
1146        doi: 10.1038/s41467-017-01261-5. URL <https://www.nature.com/articles/s41467-017-01261-5>.
- 1147
- 1148 [27] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir,  
1149        Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (MSigDB)  
1150        3.0. *Bioinformatics*, 27(12):1739–1740, June 2011. ISSN 1367-4811, 1367-  
1151        4803. doi: 10.1093/bioinformatics/btr260. URL <https://academic.oup.com/bioinformatics/article/27/12/1739/257711>.
- 1152
- 1153 [28] Dylan Lawless. Variant risk estimate probabilities for iei genes. March 2025. doi:  
1154        10.5281/zenodo.15111584. URL <https://doi.org/10.5281/zenodo.15111584>.
- 1155 [29] Eric Vallabh Minikel, Sonia M. Vallabh, Monkol Lek, Karol Estrada, Kaitlin E.  
1156        Samocha, J. Fah Sathirapongsasuti, Cory Y. McLean, Joyce Y. Tung, Linda  
1157        P. C. Yu, Pierluigi Gambetti, Janis Blevins, Shulin Zhang, Yvonne Cohen,  
1158        Wei Chen, Masahito Yamada, Tsuyoshi Hamaguchi, Nobuo Sanjo, Hidehiro  
1159        Mizusawa, Yosikazu Nakamura, Tetsuyuki Kitamoto, Steven J. Collins, Alison  
1160        Boyd, Robert G. Will, Richard Knight, Claudia Ponto, Inga Zerr, Theo F. J. Kraus,  
1161        Sabina Eigenbrod, Armin Giese, Miguel Calero, Jesús De Pedro-Cuesta, Stéphane  
1162        Haïk, Jean-Louis Laplanche, Elodie Bouaziz-Amar, Jean-Philippe Brandel, Sabina

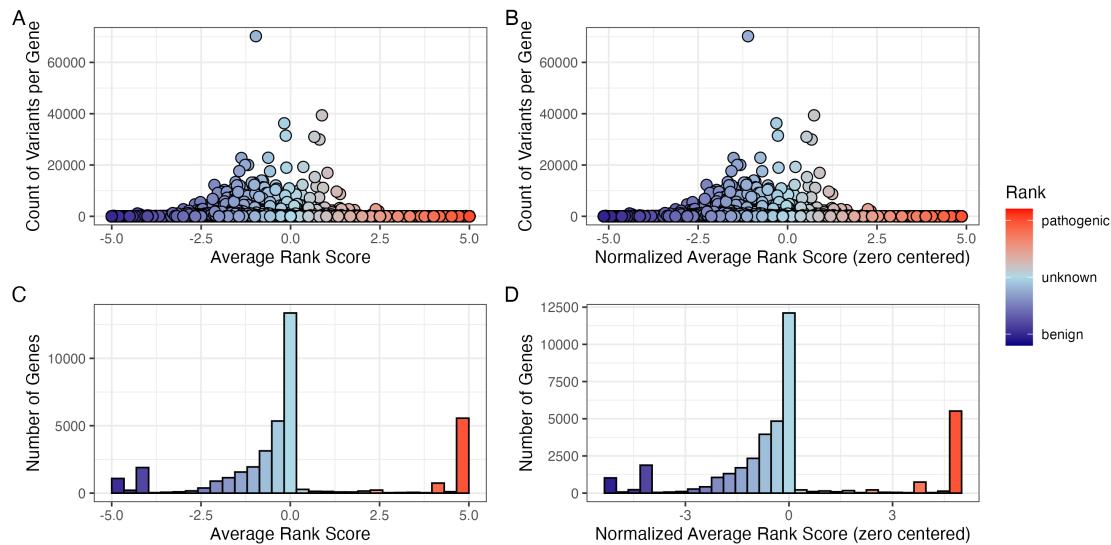
- 1163 Capellari, Piero Parchi, Anna Poleggi, Anna Ladogana, Anne H. O'Donnell-  
1164 Luria, Konrad J. Karczewski, Jamie L. Marshall, Michael Boehnke, Markku  
1165 Laakso, Karen L. Mohlke, Anna Kähler, Kimberly Chamberl, Steven McCarroll,  
1166 Patrick F. Sullivan, Christina M. Hultman, Shaun M. Purcell, Pamela Sklar,  
1167 Sven J. Van Der Lee, Annemieke Rozemuller, Casper Jansen, Albert Hofman,  
1168 Robert Kraaij, Jeroen G. J. Van Rooij, M. Arfan Ikram, André G. Uitterlinden,  
1169 Cornelia M. Van Duijn, Exome Aggregation Consortium (ExAC), Mark J. Daly,  
1170 and Daniel G. MacArthur. Quantifying prion disease penetrance using large  
1171 population control cohorts. *Science Translational Medicine*, 8(322), January  
1172 2016. ISSN 1946-6234, 1946-6242. doi: 10.1126/scitranslmed.aad5169. URL  
1173 <https://www.science.org/doi/10.1126/scitranslmed.aad5169>.
- 1174 [30] Nicola Whiffin, Eric Minikel, Roddy Walsh, Anne H O'Donnell-Luria, Konrad  
1175 Karczewski, Alexander Y Ing, Paul J R Barton, Birgit Funke, Stuart A Cook,  
1176 Daniel MacArthur, and James S Ware. Using high-resolution variant frequencies to  
1177 empower clinical genome interpretation. *Genetics in Medicine*, 19(10):1151–1158,  
1178 October 2017. ISSN 10983600. doi: 10.1038/gim.2017.26. URL <https://linkinghub.elsevier.com/retrieve/pii/S1098360021013678>.
- 1180 [31] Bradley Efron and Carl Morris. Stein's Estimation Rule and Its Competitors—An  
1181 Empirical Bayes Approach. *Journal of the American Statistical Association*, 68  
1182 (341):117, March 1973. ISSN 01621459. doi: 10.2307/2284155. URL <https://www.jstor.org/stable/2284155?origin=crossref>.
- 1184 [32] Yaowu Liu, Sixing Chen, Zilin Li, Alanna C Morrison, Eric Boerwinkle, and  
1185 Xihong Lin. Acat: a fast and powerful p value combination method for rare-  
1186 variant analysis in sequencing studies. *The American Journal of Human Genetics*,  
1187 104(3):410–421, 2019.
- 1188 [33] Xihao Li, Zilin Li, Hufeng Zhou, Sheila M Gaynor, Yaowu Liu, Han Chen,  
1189 Ryan Sun, Rounak Dey, Donna K Arnett, Stella Aslibekyan, et al. Dynamic  
1190 incorporation of multiple in silico functional annotations empowers rare variant  
1191 association analysis of large whole-genome sequencing studies at scale. *Nature  
genetics*, 52(9):969–983, 2020.
- 1193 [34] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong  
1194 Lin. Rare-variant association testing for sequencing data with the sequence kernel  
1195 association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.
- 1196 [35] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J  
1197 Rieder, Deborah A Nickerson, David C Christiani, Mark M Wurfel, and Xihong  
1198 Lin. Optimal unified approach for rare-variant association testing with application  
1199 to small-sample case-control whole-exome sequencing studies. *The American  
1200 Journal of Human Genetics*, 91(2):224–237, 2012.
- 1201 [36] Augustine Kong, Gudmar Thorleifsson, Michael L Frigge, Bjarni J Vilhjalmsson,  
1202 Alexander I Young, Thorgeir E Thorgeirsson, Stefania Benomisdottir, Asmundur

- 1203 Oddsson, Bjarni V Halldorsson, Gisli Masson, et al. The nature of nurture:  
1204 Effects of parental genotypes. *Science*, 359(6374):424–428, 2018.
- 1205 [37] Laurence J Howe, Michel G Nivard, Tim T Morris, Ailin F Hansen, Humaira  
1206 Rasheed, Yoonsu Cho, Geetha Chittoor, Penelope A Lind, Teemu Palviainen,  
1207 Matthijs D van der Zee, et al. Within-sibship gwas improve estimates of direct  
1208 genetic effects. *BioRxiv*, pages 2021–03, 2021.
- 1209 [38] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-  
1210 Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al.  
1211 Standards and guidelines for the interpretation of sequence variants: a joint con-  
1212 sensus recommendation of the american college of medical genetics and genomics  
1213 and the association for molecular pathology. *Genetics in medicine*, 17(5):405–423,  
1214 2015.
- 1215 [39] Sean V Tavtigian, Steven M Harrison, Kenneth M Boucher, and Leslie G Biesecker.  
1216 Fitting a naturally scaled point system to the acmg/amp variant classification  
1217 guidelines. *Human mutation*, 41(10):1734–1737, 2020.
- 1218 [40] Quan Li and Kai Wang. Intervar: clinical interpretation of genetic variants by  
1219 the 2015 acmg-amp guidelines. *The American Journal of Human Genetics*, 100  
1220 (2):267–280, 2017.
- 1221 [41] Brent S Pedersen, Joe M Brown, Harriet Dashnow, Amelia D Wallace, Matt  
1222 Velinder, Martin Tristani-Firouzi, Joshua D Schiffman, Tatiana Tvrzik, Rong  
1223 Mao, D Hunter Best, et al. Effective variant filtering and expected candidate  
1224 variant yield in studies of rare human disease. *NPJ Genomic Medicine*, 6(1):1–8,  
1225 2021.
- 1226 [42] Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon,  
1227 Andrew P Morris, and Krina T Zondervan. Data quality control in genetic  
1228 case-control association studies. *Nature protocols*, 5(9):1564–1573, 2010. URL  
1229 <https://doi.org/10.1038/nprot.2010.116>.
- 1230 [43] David T Miller, Kristy Lee, Noura S Abul-Husn, Laura M Amendola, Kyle  
1231 Brothers, Wendy K Chung, Michael H Gollob, Adam S Gordon, Steven M  
1232 Harrison, Ray E Hershberger, et al. Acmg sf v3. 2 list for reporting of secondary  
1233 findings in clinical exome and genome sequencing: a policy statement of the  
1234 american college of medical genetics and genomics (acmg). *Genetics in Medicine*,  
1235 25(8):100866, 2023.

<sub>1236</sub> **6 Supplemental**

<sub>1237</sub> Supplemental data are presented under the same headings that correspond to their  
<sub>1238</sub> relevant main text sections.

<sub>1239</sub> **6.1 Variant class occurrence probability**



**Figure S1: Global distribution of ClinVar clinical-significance classification scoring.** (A) Number of variants per gene containing the assigned score for each ClinVar classification term ( $-5$  to  $+5$ ). (B) The same data after normalisation by zero centring the average rank score. (C) The tally of genes for their average rank and (D) after normalisation. No normalisation was required for the scoring system as shown by comparison of A-C and B-D.

1240 **6.2 Integrating observed true positives and unobserved false**  
1241 **negatives into a single, actionable conclusion**

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

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

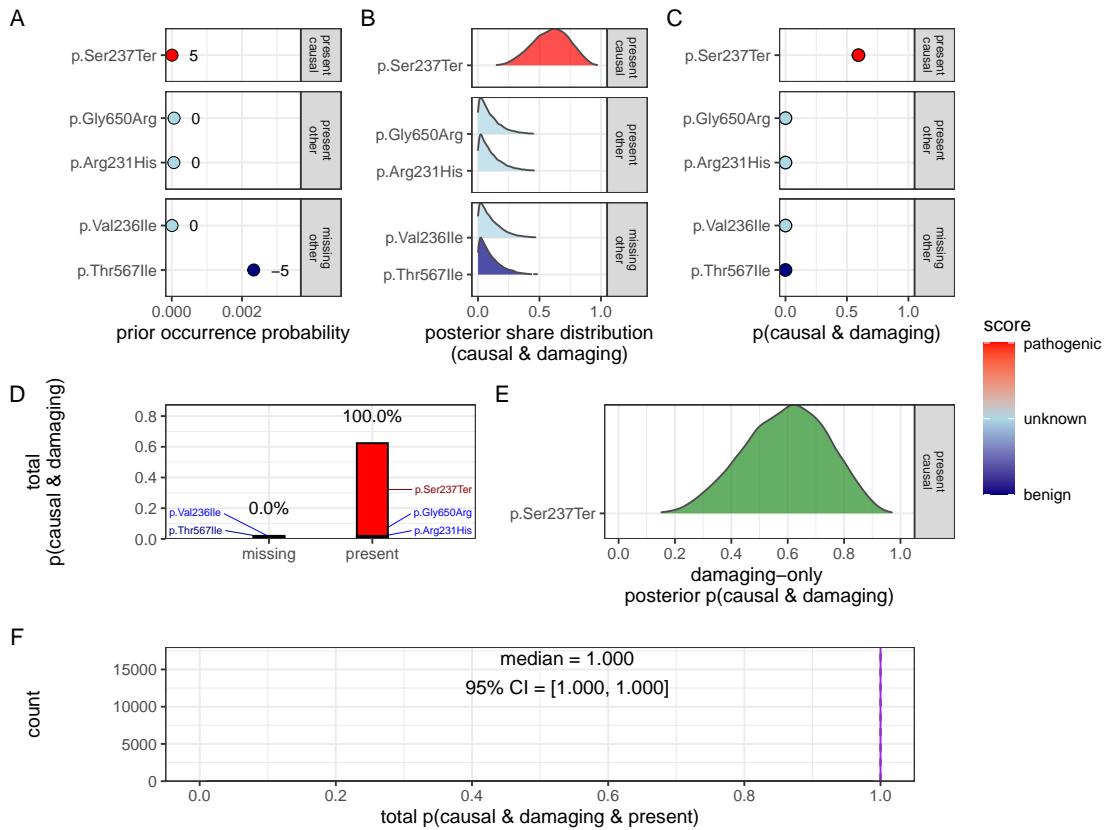
Table S2: Result of clinical genetics diagnosis scenario 2 including metadata. The most strongly supported observed variant was **p.Ser237Ter** (posterior: 0.381). The strongest unsequenced variant was **c.159+1G>A** (posterior: 0.353). The total probability of a causal diagnosis given the available evidence was 0.52 (95% CI: 0.248–0.787).

Variant	Flag	Class	Evi-dence Score	Occur-rence Prob	Adj Occ Prob	Alpha	Beta	Lower	Median	Upper	Poste-rior Share	Prob Causal
p.Ser237Ter	present	causal	5.0	0.000	0	6.0	371	0.003	0.096	0.557	0.381	0.381
c.159+1G>A	missing	causal	4.5	0.000	0	5.5	367	NA	NA	NA	0.353	0.353
p.Thr567Ile	missing	other	-5.0	0.002	0	1.0	365	NA	NA	NA	0.000	0.000
p.Arg231His	present	other	0.0	0.000	0	1.0	359	0.003	0.096	0.557	0.000	0.000
p.Gly650Arg	present	other	0.0	0.000	0	1.0	349	0.003	0.096	0.557	0.000	0.000
p.Val236Ile	missing	other	0.0	0.000	0	1.0	363	NA	NA	NA	0.000	0.000
Total	NA	NA	NA	NA	NA	NA	NA	0.248	0.520	0.787	NA	0.520

Table S3: Result of clinical genetics diagnosis scenario 3 including metadata. No observed variants were detected in this scenario. The strongest unsequenced variant was **p.Cys243Arg** (posterior: 0.366). The total probability of a causal diagnosis given the available evidence was 0 (95% CI: 0–0).

Variant	Flag	Class	Evi-dence Score	Occur-rence Prob	Adj Occ Prob	Alpha	Beta	Lower	Median	Upper	Poste-rior Share	Prob Causal
p.Cys243Arg	missing	causal	5.0	0.000	0.000	6	341	NA	NA	NA	0.366	0.366
p.Tyr246Ter	missing	causal	4.0	0.000	0.000	5	369	NA	NA	NA	0.284	0.284
p.Lys304Glu	missing	other	-5.0	0.000	0.000	1	353	NA	NA	NA	0.000	0.000
p.Ile207Leu	missing	other	-4.5	0.000	0.000	1	359	NA	NA	NA	0.000	0.000
p.His646Pro	missing	other	0.0	0.002	0.001	1	377	NA	NA	NA	0.000	0.000
p.Arg280Trp	missing	other	-4.0	0.000	0.000	1	357	NA	NA	NA	0.000	0.000
p.Thr635Ile	missing	other	0.0	0.000	0.000	1	349	NA	NA	NA	0.000	0.000
p.Arg162Trp	missing	other	0.0	0.000	0.000	1	369	NA	NA	NA	0.000	0.000
Total	NA	NA	NA	NA	NA	NA	NA	0	0	0	NA	0.000

Gene: *NFKB1*



**Figure S2: Quantification of present (TP) and no missing (FN) causal genetic variants for disease in *NFKB1* (scenario 1).** Only one known pathogenic variant, p.Ser237Ter, was observed and all previously reported pathogenic positions were successfully sequenced and confirmed as reference (true negatives). Panels (A–F) follow the same structure as scenario 2 described in **Figure 2**, culminating in a gene-level posterior probability of 1 (95 % CrI: 0.99–1.00), with full support assigned to the observed allele given the available evidence. Pathogenicity scores (-5 to +5) are annotated.

Gene: TNFAIP3

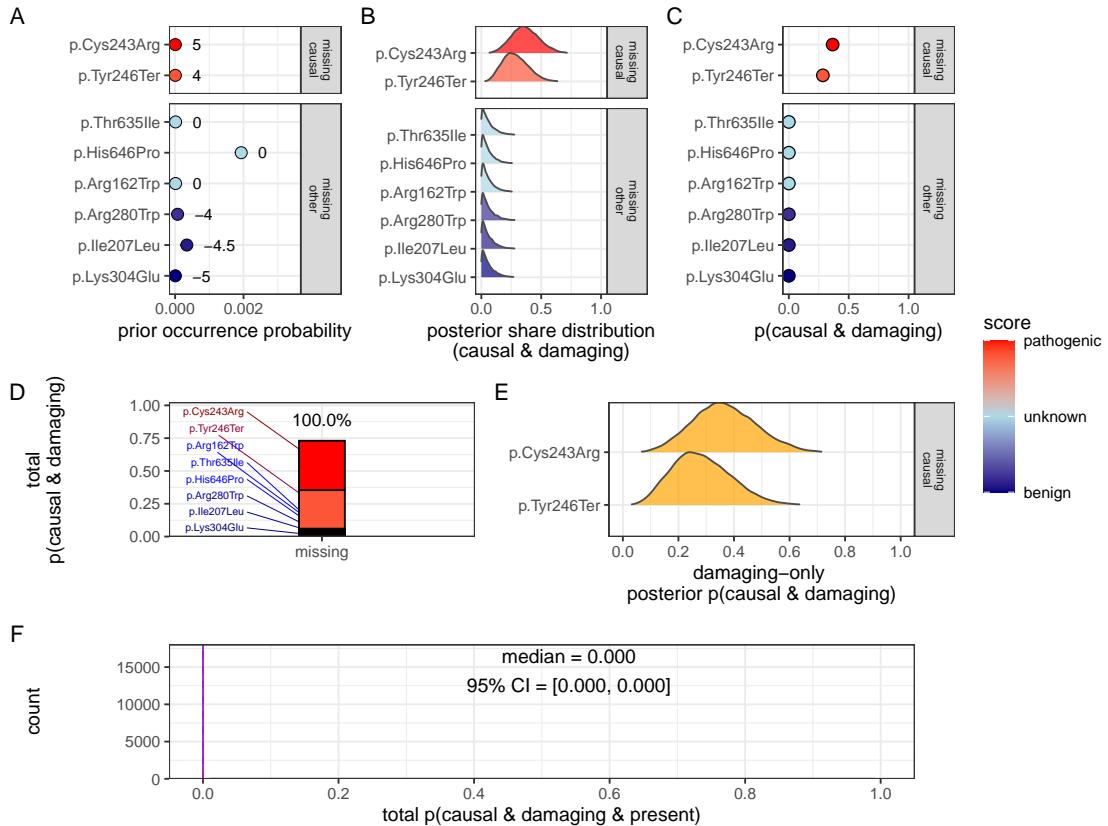
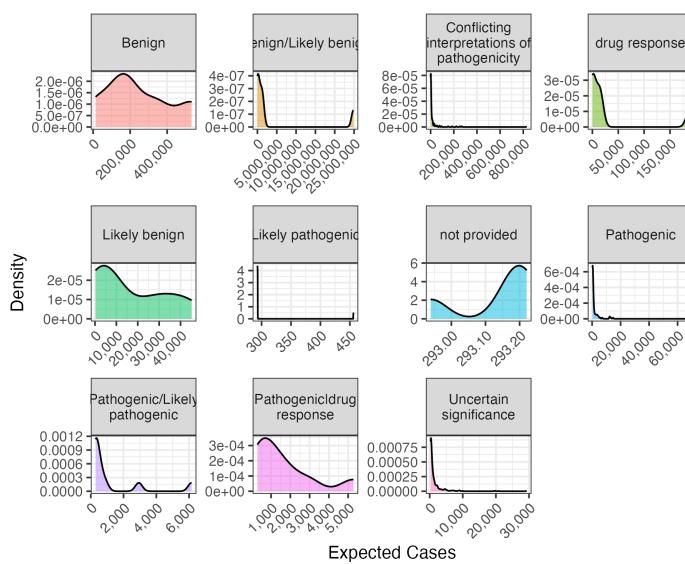


Figure S3: Quantification of no present (TP) in *NFKB1* and only missing (FN) causal genetic variants for disease in *TNFAIP3* (scenario 3). No known causal variants were observed in *NFKB1*, but one representative unsequenced allele was selected from each distinct ClinVar classification and treated as a potential false negative. Panels (A–F) follow the same structure as scenario 2 described in Figure 2. The posterior reflects uncertainty across multiple plausible but unobserved variants, resulting in low CrI (0–0) and 100% missing overall attribution in contrast to scenarios where known pathogenic variants were observed. For this patient, we have no evidence of a causal variant since the only top candidates are not yet accounted for. Pathogenicity scores (-5 to +5) are annotated in (A).

### 6.3 Validation studies

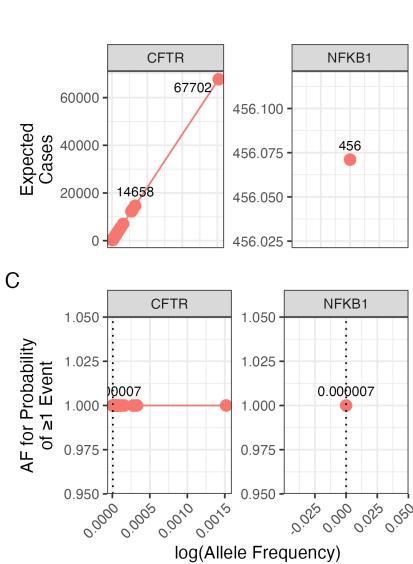
Condition: population size 69433632, phenotype PID-related, genes CFTR and NFKB1.

A

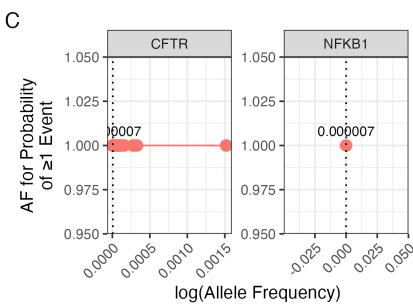


B

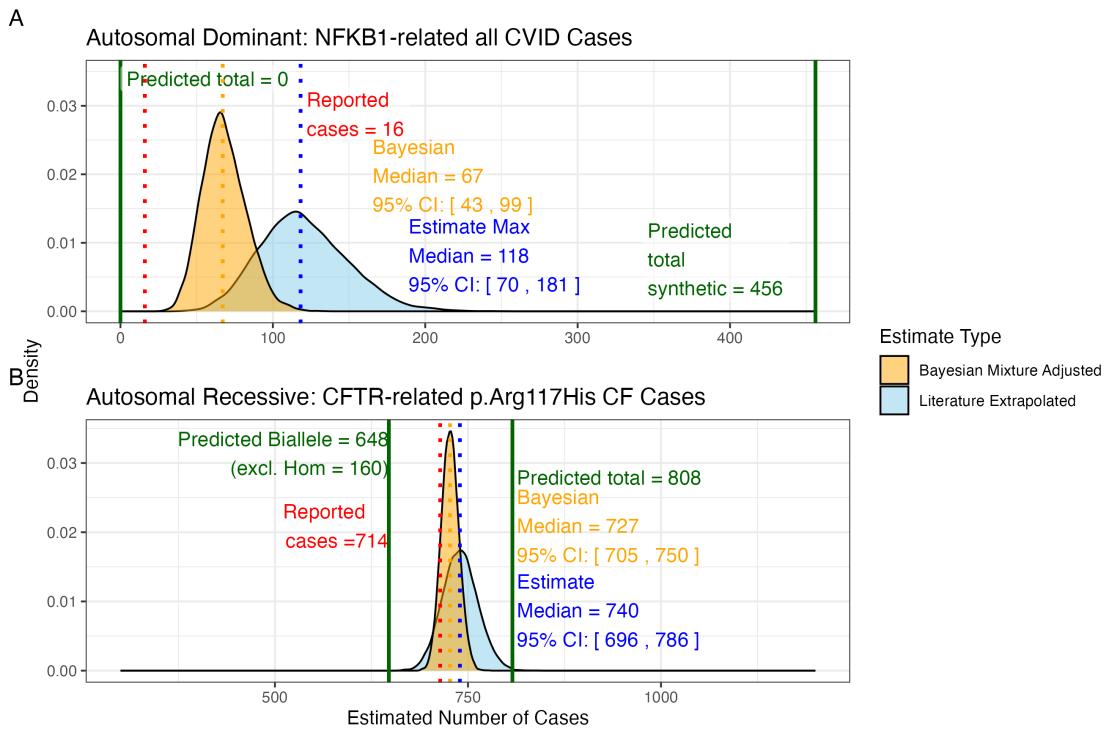
clinvar\_clnsig • Pathogenic



C



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



**Figure S5: 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.

### 6.3.1 Interpretation of ClinVar variant observations

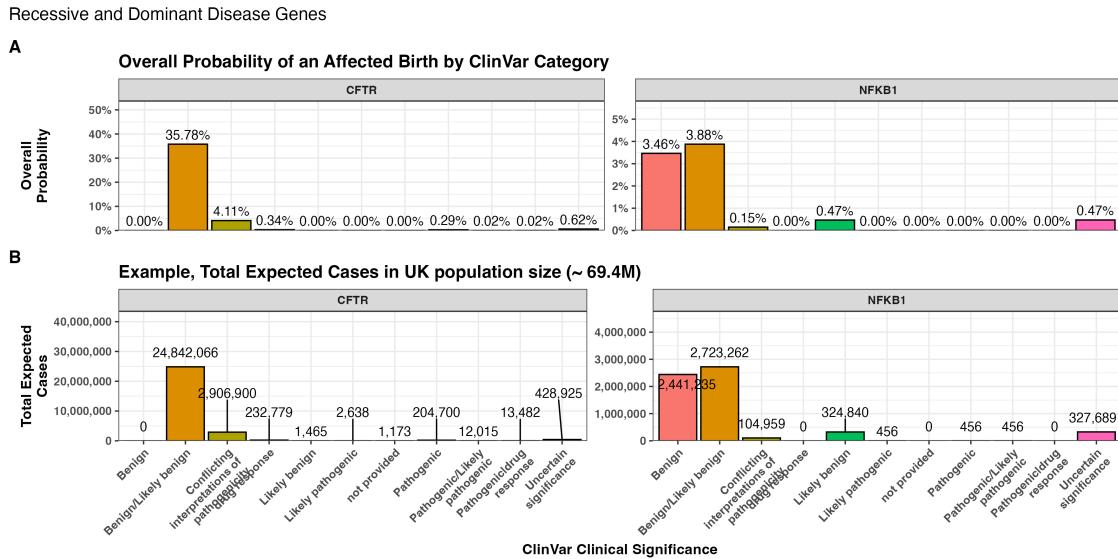
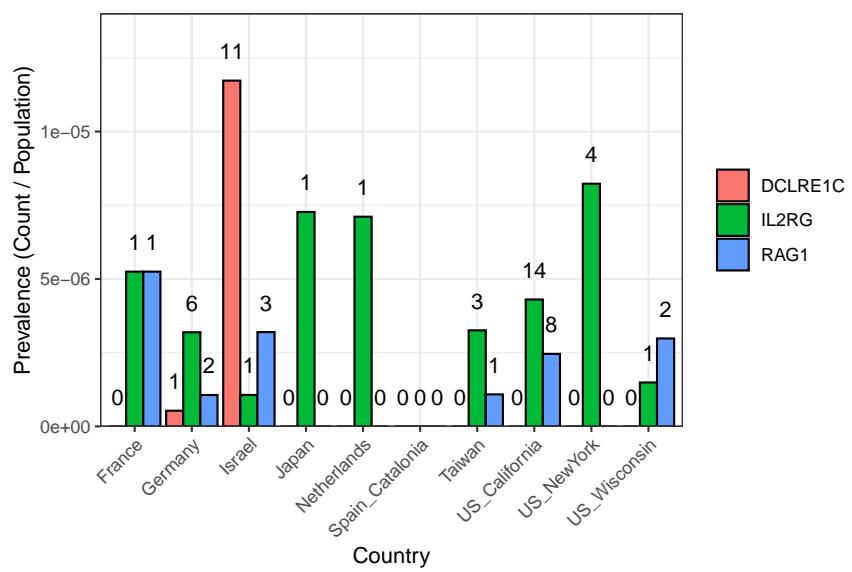
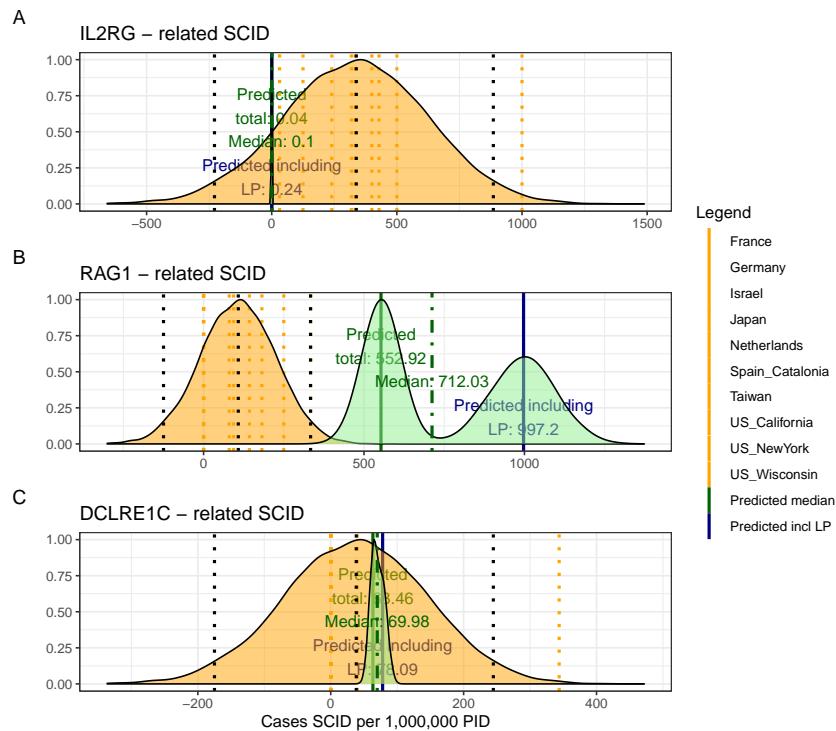


Figure S6: Combined bar charts summarising the genome-wide analysis of ClinVar clinical significance for the PID gene panel. Panel (A) shows the overall probability of an affected birth by variant classification, and (B) displays the total expected number of cases per classification, both stratified by gene. These integrated results illustrate the variability in variant observations across genes and underpin our validation of the probability estimation framework.

### 6.3.2 Validation of SCID-specific disease occurrence



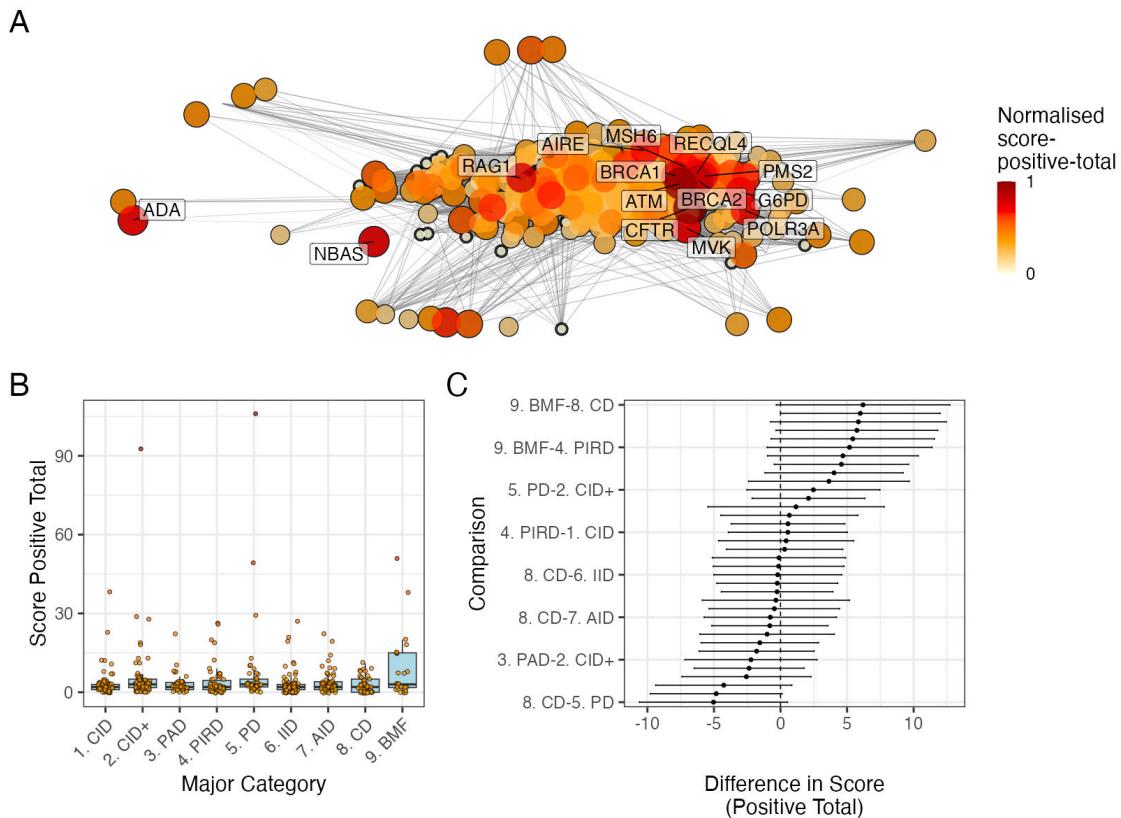
**Figure S7: SCID-specific gene comparison across regions.** The bar plot shows the prevalence of SCID-related cases (count divided by population) for each gene and country (or region), with numbers printed above the bars representing the actual counts in the original cohort (ranging from 0 to 11 per region and gene).



**Figure S8: 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.

<sup>1245</sup> **6.4 Genetic constraint in high-impact protein networks**

<sup>1246</sup> **6.4.1 Score-positive-total within IEI PPI network**



**Figure S9: 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.

1247    **6.4.2 Hierarchical Clustering of Enrichment Scores for Major Disease Cat-**  
 1248    **egories**

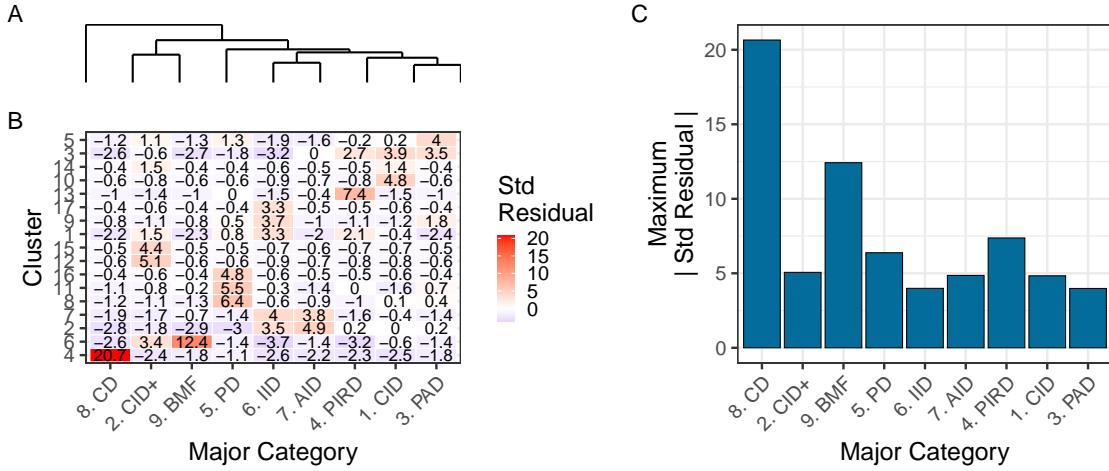


Figure S10: **Hierarchical clustering of enrichment scores.** The heatmap displays standardised residuals for major disease categories (x-axis) across network clusters (y-axis). A dendrogram groups similar disease categories, and the bar plot shows the maximum absolute residual per category. (8) CD and (9)BMF show the highest values, indicating significant enrichment or depletion ( $\text{residuals} > |\text{2}|$ ). Definitions in **Box 2.1**.

1249    6.4.3 PPI connectivity, LOEUF constraint and enriched network cluster  
1250    analysis

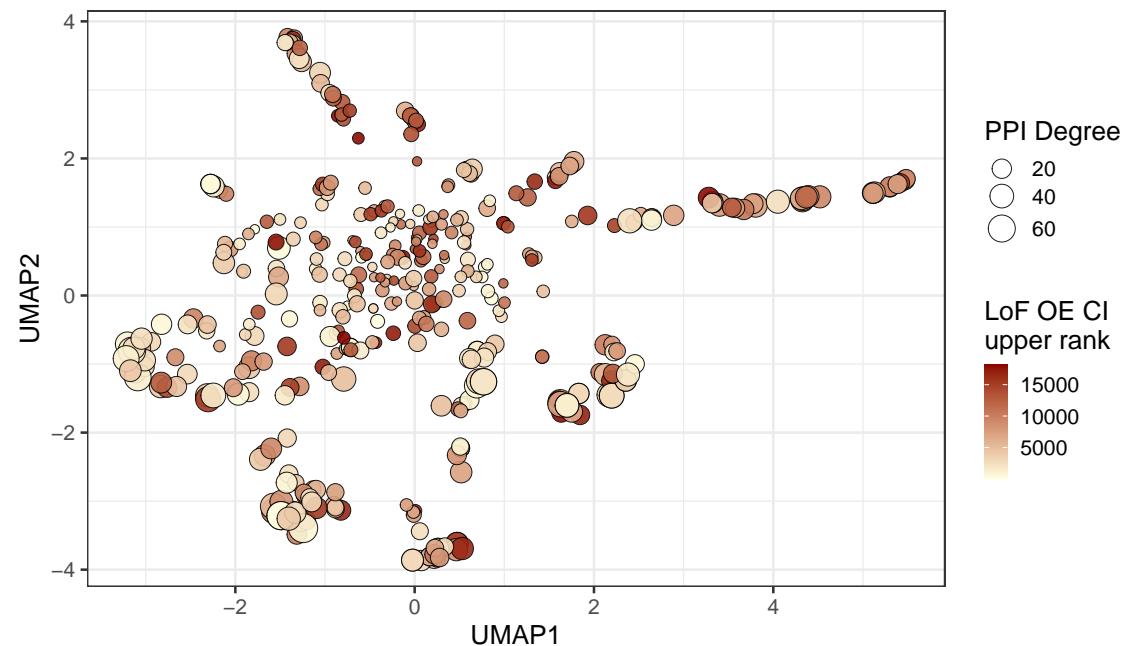
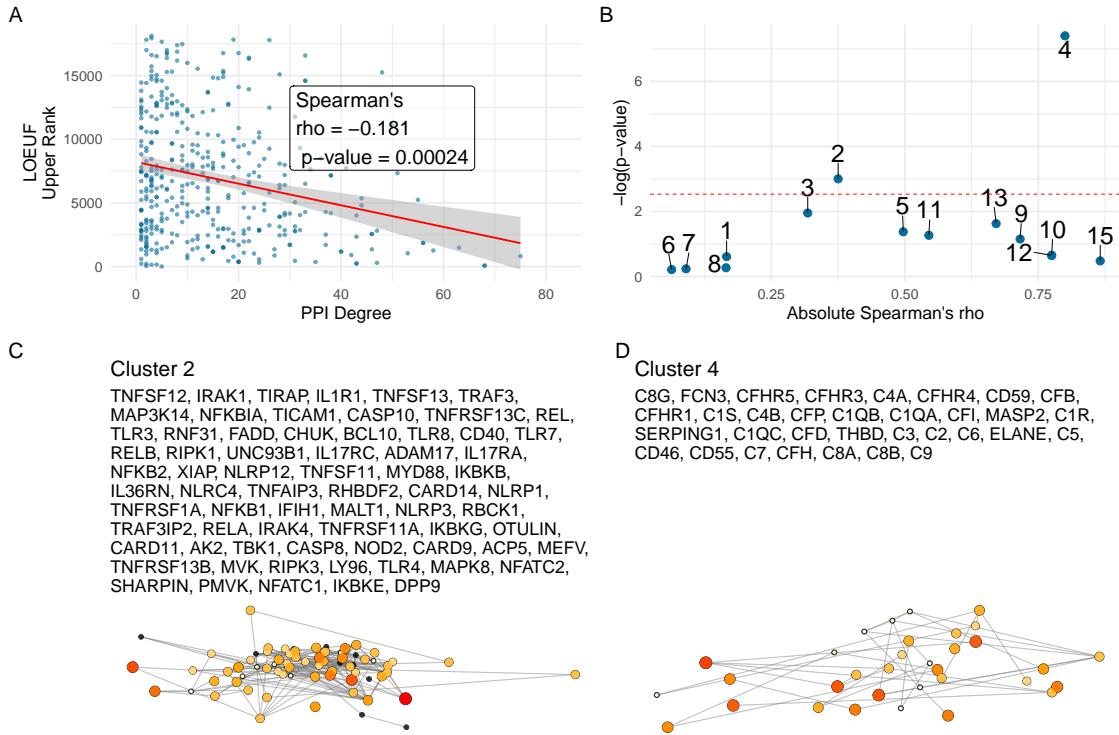
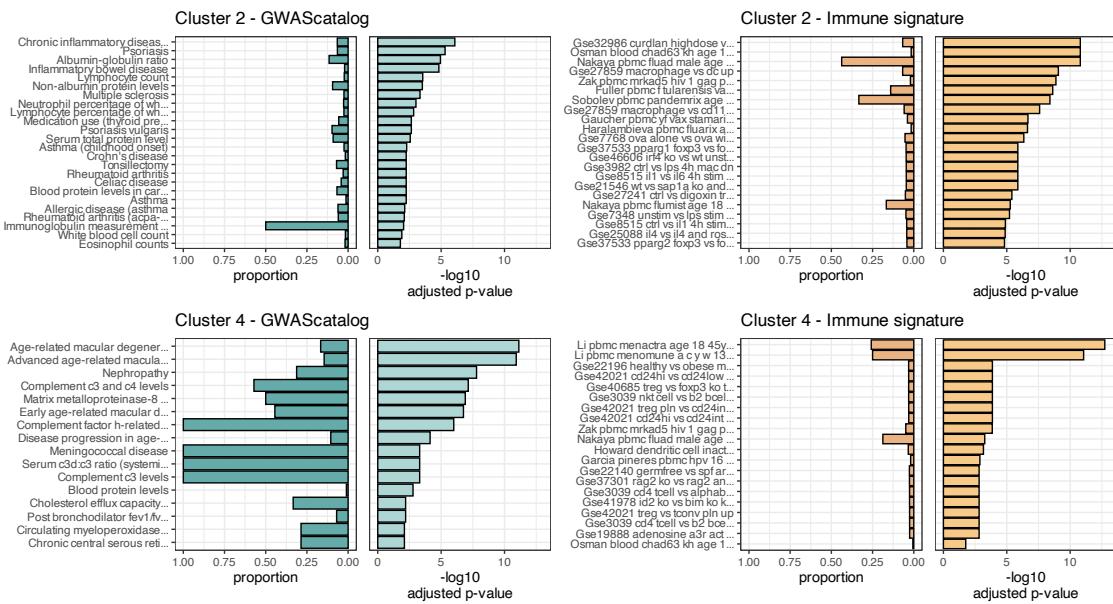


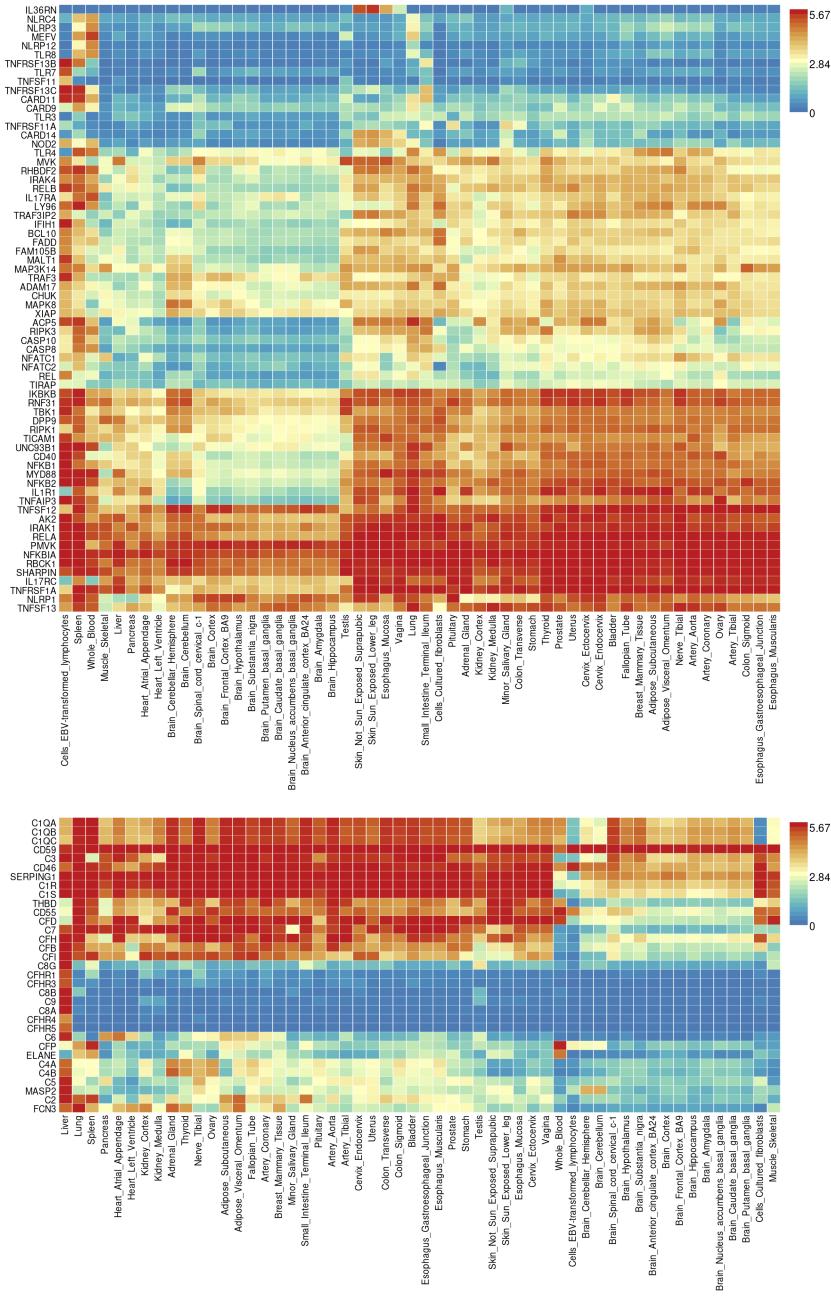
Figure S11: **Analysis of PPI degree versus LOEUF upper rank with UMAP embedding of the PPI network.** The relationship between PPI degree (size) and LOEUF upper rank (color) across gene clusters. No clear patterns are evident.



**Figure S12: Correlation between PPI degree and LOEUF upper rank. (A)** Ananlysis across all genes revealed a weak, significant negative correlation between PPI degree and LOEUF upper rank. **(B)** The cluster-wise analysis showed that clusters 2 and 4 exhibited moderate to strong correlations, while other clusters display weak or non-significant relationships. **(C) and (D)** Shows the new network plots for the significantly enriched clusters based on gnomAD constraint metrics.

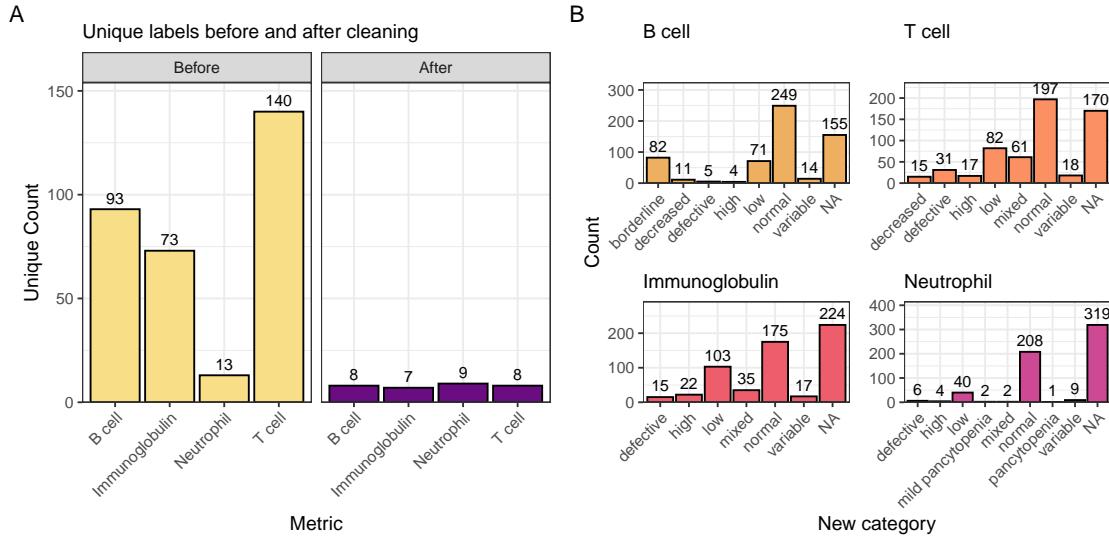


**Figure S13: Composite Enrichment Profiles for IEI Gene Sets.** We selected the top two enriched clusters (as per **Figure S12**) and performed functional enrichment analysis derived from known disease associations. For each gene set, the left panel displays the proportion of input genes overlapping 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.

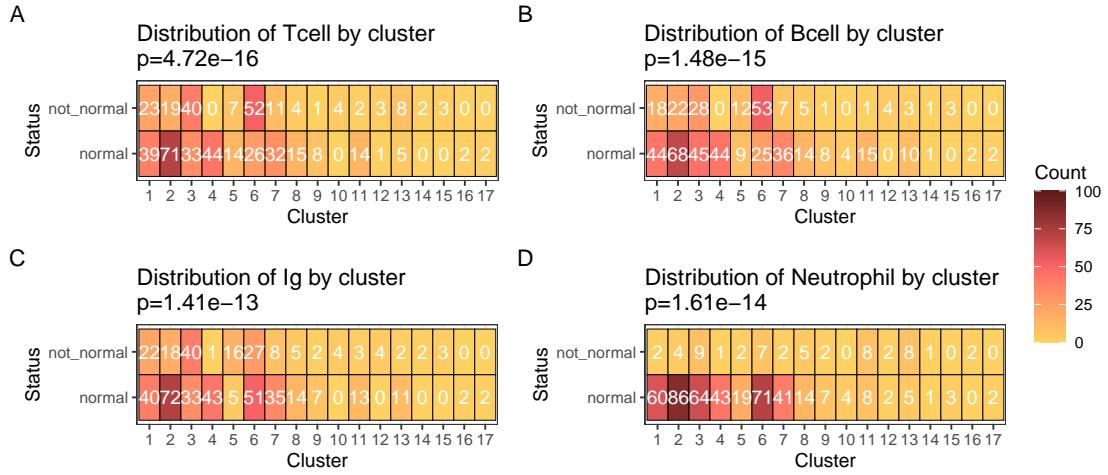


**Figure S14: Gene Expression Heatmaps for IEI Genes.** GTEx v8 data from 54 tissue types display the average expression per tissue label ( $\log_2$  transformed) for the IEI gene panels. Top: Cluster 2; Bottom: Cluster 4.

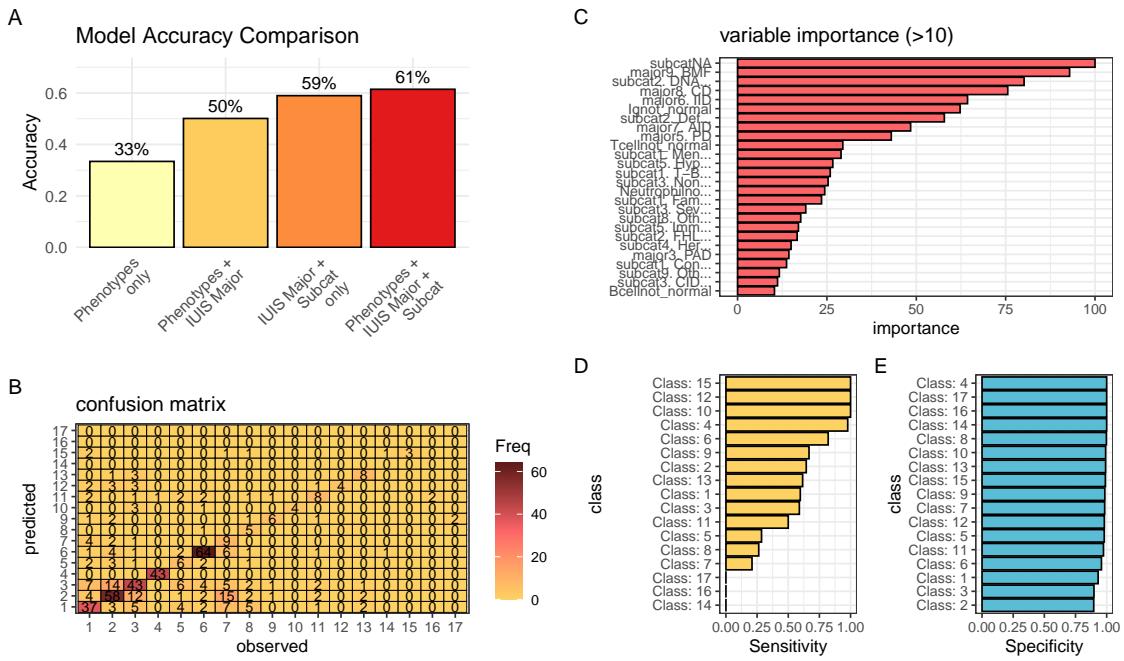
1251 **6.5 Novel PID classifications derived from genetic PPI and  
1252 clinical features**



**Figure S15: 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 S16: 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 S17: 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.

1253 6.6 Probability of observing AlphaMissense pathogenicity

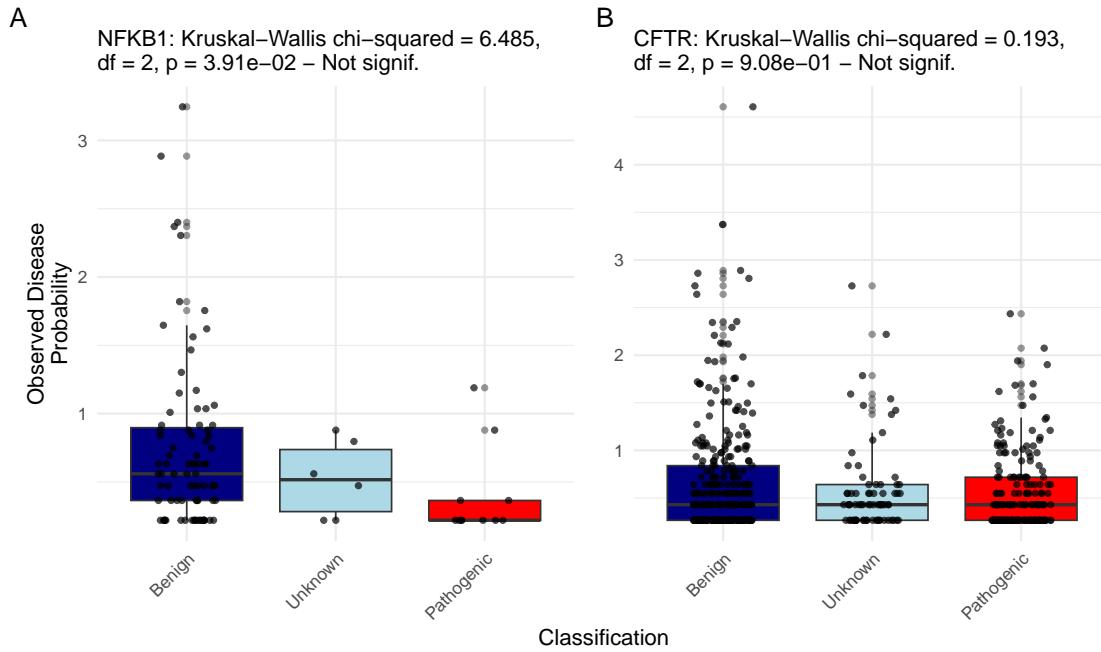


Figure S18: **Observed Disease Probability by Clinical Classification with AlphaMissense.** The figure displays the Kruskal–Wallis test results for NFKB1 and CFTR, showing no significant differences.