Dissecting the Reduced Penetrance of Putative Loss-of-Function Variants in Population-Scale Biobanks

David R. Blair[1,3,#], Neil Risch[2,3]

1. Division of Medical Genetics, Department of Pediatrics
2. Department of Epidemiology & Biostatistics
3. University of California San Francisco
Corresponding Author: David.Blair@ucsf.edu

Abstract:

Loss-of-function variants (LoFs) disrupt the activity of their impacted gene. They are often associated with clinical phenotypes, including autosomal dominant diseases driven by haploinsufficiency. Recent analyses using biobanks have suggested that LoF penetrance for some haploinsufficient disorders may be low, an observation that has important implications for population genomic screening. However, biobanks are also rife with missing data, and the reliability of these findings remains uncertain. Here, we examine the penetrance of putative LoFs (pLoFs) using a cohort of ≈24,000 carriers derived from two population-scale biobanks: the UK Biobank and the All of Us Research Program. We investigate several possible etiologies for reduced pLoF penetrance, including biobank recruitment biases, annotation artifacts, missed diagnoses, and incomplete clinical records. Systematically accounting for these factors increased penetrance, but widespread reduced penetrance remained. Therefore, we hypothesized that other factors must be driving this phenomenon. To test this, we trained machine learning models to identify pLoFs with high penetrance using the genomic features specific to each variant. These models were predictive of penetrance across a range of diseases and pLoF types, including those with prior evidence for pathogenicity. This suggests that reduced pLoF penetrance is in fact common, and care should be taken when counseling asymptomatic carriers.

Introduction:

Exome and genome sequencing are now first-tier tests for rare disease diagnosis[1–6]. Given this success, there is growing interest in applying these technologies to asymptomatic patients[7–17]. The utility of sequencing for screening remains uncertain[18–20]. Generally, a screening test's utility is quantified using its positive predictive value. For genetic testing, this statistic is driven both by the accuracy of the genotype call and its penetrance for the phenotype of interest, where penetrance is defined as the probability that a carrier will manifest the disease. Although imperfect, the accuracy of genotype calling is relatively high[21,22]. Alternatively, penetrance estimates for most genotypes are unknown; they can range anywhere from 0 (no associated disease risk) to 1 (certain disease manifestation). These estimates also vary with age and can be modified by additional factors, including polygenic background[23–25] and environmental exposures[26]. For diagnostic applications, accurate penetrance estimates are less critical. Patients already express a disease phenotype, so laboratories typically

must only determine if variants are pathogenic (i.e. disease-associated) or benign[27]. Variant interpretation for screening applications is more complex. Laboratories and clinicians should be able to express how likely the variants are to cause disease in the future. This risk is of course inextricably tied to penetrance.

Penetrance is notoriously difficult to estimate[28]. For a few pathogenic genotypes that are unusually common (typically due to founder events), accurate penetrance estimation is possible[29–32]. Generally, however, pathogenic genotypes are extremely rare. As a result, penetrance is unknown for most clinically relevant variants. Recently, population-scale biobanks with linked electronic health record (EHR) data have become widely available[33–40], and these datasets have been used to estimate penetrance using a "genotype first" approach[41]. Here, pathogenic variant carriers are identified using the available genetic data, after which their phenotypic expression is assessed retrospectively using EHR data. These analyses have suggested widespread reduced penetrance for pathogenic variants[42–44]. This observation has important implications for genomic screening, as it suggests that the positive predictive value of genetic testing may be unacceptably low. That said, biobanks may have limitations as a resource for estimating penetrance, and biases related to recruitment, coupled with missing data, may lead to deflated estimates[41,45].

Here, we investigate the apparent penetrance for one of the simplest classes of potentially pathogenic mutations: putative loss-of-function (pLoF) variants in haploinsufficient disease genes. To do so, we uniformly processed the genomic data from two biobanks (the UK Biobank[36] and the All of Us Research Program[40]; combined N>700,000), identifying ≈24,000 pLoF carriers at risk for 91 diseases. We then analyzed their relative frequencies across biobanks and diseases, demonstrating that the types and frequencies of pLoFs in these biobanks were likely shaped by recruitment biases. Nevertheless, biobank pLoFs had strong and replicable phenotypic effects, and consistent with prior analyses, penetrance was generally reduced. We then investigated possible factors underlying the reduced penetrance, adjusting estimates accordingly. Examples include annotation artifacts, missed diagnoses, and censored clinical data. Accounting for all these factors increased estimates, but widespread reduced penetrance remained. Therefore, we hypothesized that many of these variants may in fact have intrinsically low penetrance, which may be a function of incomplete or "leaky" loss-of-function. To test this, we trained machine learning models to predict pLoF penetrance using variant-specific genomic features that may correlate with incomplete loss-of-function. These models were predictive of penetrance across a range of diseases and variant types, including those previously annotated to be pathogenic by diagnostic testing laboratories[46]. Consequently, LoF penetrance remains quite uncertain, and accurately communicating this uncertainty to asymptomatic carriers will be crucial for successful genomic screening.

Results

Biobanks are Likely Depleted of pLoFs with Severe Phenotypic Effects

Using the ClinGen database[47], we identified 91 autosomal dominant/pseudo-autosomal dominant Mendelian disorders for which haploinsufficiency is a likely

mechanism of disease (see Methods). This set of diseases covered a broad range of human pathophysiology, including neurodevelopmental disorders, congenital malformation syndromes, and diseases linked to tumor predisposition. Most (76%) typically present during childhood while the remainder occur during various stages of adulthood. The specific diseases analyzed in this study, along with their annotated information, are provided in Supplemental Table 1. Following annotation, we linked the diseases to their associated genes using the Online Mendelian Inheritance in Man database[48] (117 in total). We then systematically identified all putative loss-of-function variants (pLoFs) within these genes in both biobanks, removing those that likely represent technical artifacts based on sequencing depth and quality scores (see Methods). In total, we identified 3,131 unique pLoFs in the UK Biobank (UKBB; total N= 468,672) and 3,889 in the All of Us Research Program (AoU; total N= 245,376), resulting in a total of 6,247 unique pLoFs (773 occured in both). Additional details about the individual variants can be found in Supplemental Tables 2 (UKBB) and 3 (AoU). The distribution of pLoF carrier counts in both datasets are displayed in Figures 1A and B. Most variants were singletons (63% and 67% in the UKBB and AllofUs respectively), consistent with their likely negative impact on fitness and survival.

Even though most individual variants were singletons, the total number of pLoFs per disease was highly variable, ranging over nearly three orders of magnitude (Figure 1C). Moreover, the disease-specific pLoF frequencies were highly correlated across the biobanks (R^2 = 0.84; P-value < 2.2×10^{-16}). Many factors likely drive this shared variability, including properties specific to the populations that constitute each biobank and attributes of the diseases and variants themselves. For example, founder effects and genetic drift almost certainly contribute to the shared variability displayed in Figure 1C. However, the demographic backgrounds for the two biobanks are quite distinct [36,40]. The UKBB mostly contains subjects of European ancestry (≈90%)[36], while AoU is far more diverse[40] (European fraction ≈ 50%). Therefore, while founder variants likely drive some of the shared variability in per-disease pLoF frequencies, their contribution should be limited. Moreover, if genetic drift were driving the shared variability, the pLoF frequencies should correlate with the number of coding sites linked to each disease. While this was true (Figure 1D), the fraction of the variability explained by this phenomenon was only 23%, suggesting that other factors were likely involved.

The per-disease pLoF frequency estimates, after correcting for coding sequence length, were positively correlated with typical disease onset in both biobanks, such that pLoFs linked to childhood-onset diseases were less common than those linked to adult-onset disorders (Figure 1E). This suggested that the biobanks may be depleted of variants linked to childhood-onset diseases, likely due to recruitment biases that favor living and/or healthier adults. Notably, pLoF carriers in general were recruited at a younger age than their non-carrier counterparts (0.87 years on average, Wilcoxon Signed Rank Test Meta-Analysis P-value: 2.07×10^{-5}; Figure 1F), consistent with a more pervasive recruitment bias that favors healthier individuals recruited prior to disease onset (see Supplemental Table 4 for all disease-specific results). This implies that biobanks that recruit younger subjects should harbor more pLoFs. The average recruitment age for AoU was 52 years compared to 57 years for the UKBB (T-test P-value < 2.2×10^{-16}), a consequence of the distinct recruiting strategies for two the studies[49,50]. Thus, it was not surprising to find that the overall pLoF carrier rates were

higher in AoU than the UKBB (4.5% vs 3.0%; *P*-value < 2.2×10^{-16}), an effect that persisted after correcting for differences in ancestry (UKBB: 2.9%; AoU: 4.2%, see Methods).

To avoid such biases, penetrance estimates would ideally be derived from prospective cohorts with millions of subjects, either starting from birth or even during pregnancy. Given that such studies are currently infeasible, biobanks likely represent the best opportunity for systematic penetrance estimation. Based on the above analyses, however, these datasets are likely depleted of pLoFs with severe phenotypic effects, at least compared to younger cohorts. As corollary, they are likely enriched for variants with low penetrance and/or milder phenotypic effects. This implies that aggregate estimates of penetrance derived from biobanks may be systematically deflated when compared to those obtained using other study designs. This is likely to be particularly true for diseases associated with high morbidity and mortality. This does not necessarily imply penetrance estimates from these datasets are meaningless, but care should be taken when applying biobank penetrance estimates to other populations.

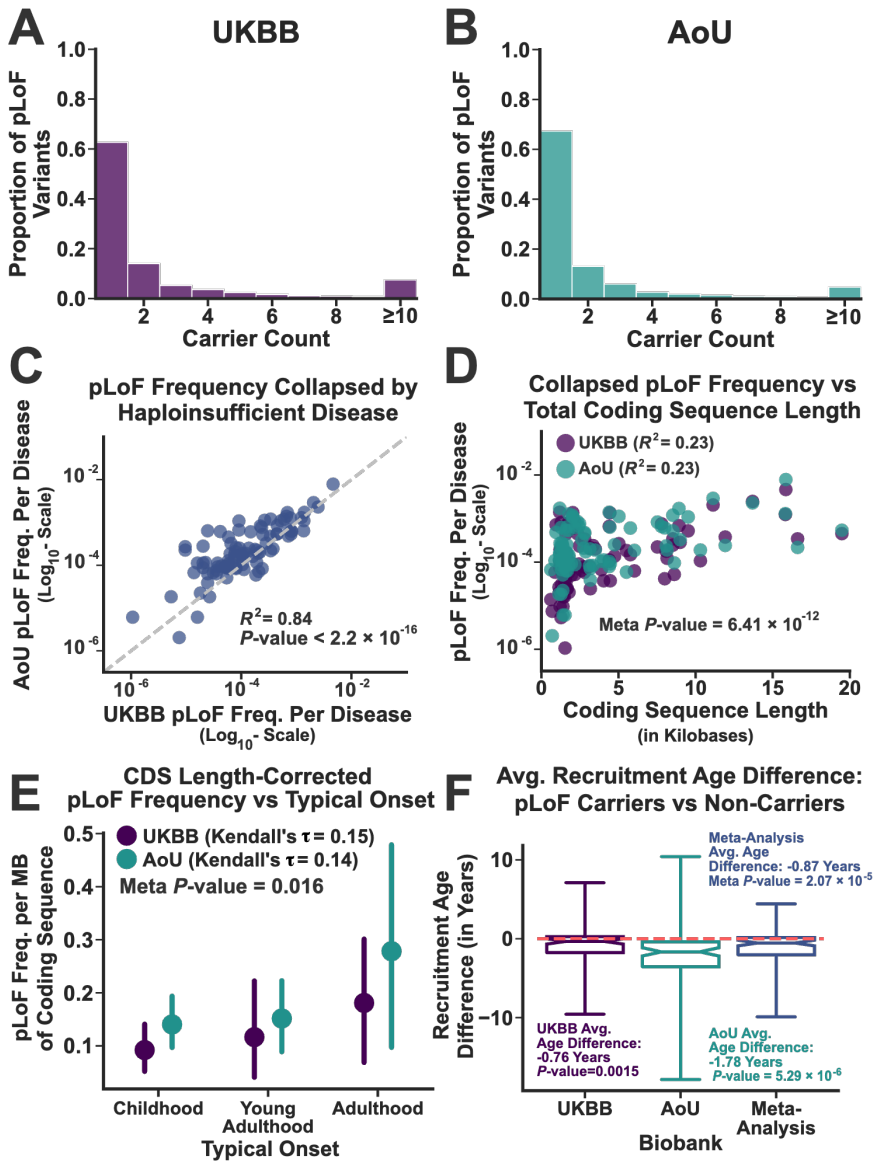


Figure 1: Biobank datasets are likely depleted of pLoFs with severe phenotypic effects. (A, B): The carrier count distributions for the pLoF variants identified in the UK Biobank (UKBB; B) and the All of Us Research Program (AoU; B). (C): The disease-specific pLoF carrier frequencies (*N*=91) compared across the two biobanks: UKBB (x-axis) and AoU (y-axis). Correlation was assessed using Pearson's method (R^2). (D): The coding sequence (CDS) length was computed using the MANE Select transcript for the gene(s) linked to each disease (x-axis), and this was compared to the per-disease pLoF frequency (y-axis) using Pearson's method (R^2). (E): The collapsed pLoF frequencies were normalized by the CDS length (in megabases; MB) and plotted against the three onset classes. Correlation was assessed using Kendall's rank correlation (denoted τ). Error bars represent 95% bootstrapped confidence intervals. (F): Boxplots display the distribution over the differences in recruitment age for the pLoF carriers and their corresponding non-carrier controls (see Methods). This distribution is depicted for each biobank along with a boxplot that summarizes the results of a cross-biobank meta-analysis. The edges of the box define the interquartile range, while the notch indicates the median. The whiskers depict the total range. Statistical significance was assessed using a Wilcoxon signed rank test (two-sided).

pLoFs in Biobanks Have Detectable and Consistent Phenotypic Effects

Although biobanks are likely depleted of pLoFs with severe phenotypic effects, many carriers in these datasets still manifest strong signs of disease expression. To illustrate, we constructed control groups for each disease by identifying biobank subjects that did not carry any rare variant (allele frequency ≤ 0.1%) in their associated genes (see Methods). We then systematically estimated the effects of the pLoF variants on haploinsufficient disease risk by comparing the disease prevalence among carriers and non-carriers using logistic regression. This analysis was limited to those diseases with at least 1 diagnosis in either the carriers or non-carriers across both biobanks (N=28). In a cross-biobank meta-analysis, the pLoF variants had a Bonferroni-corrected statistically significant effect on risk for over two-thirds (20/28) of the haploinsufficient diseases (see Supplemental Table 5 for the full set of results). Moreover, the risk estimates for the statistically significant associations were correlated across biobanks (R^2 = 0.53; P-value = 1.27×10^{-4}, see Figure 2A).

These results indicate that pLoFs have strong and replicable effects on disease prevalence, but the analysis was limited to diseases with diagnostic codes in available in the EHR data. Most haploinsufficient diseases were not amenable to this analysis (N=63), as they lacked the diagnostic data needed to assess their risk. To overcome this limitation, we also quantified the disease expression of the pLoF variants using covariate-corrected Phenotype Risk Scores (PheRS)[51,52]. These scores measure the extent to which a subject is a phenotypic outlier based on their pattern of expressed disease-specific symptoms, which is possible even in the absence of a formal diagnosis (see Methods). Figures 2B (UKBB) and 2C (AoU) display the distributions over the median PheRS estimates for each disease, which were standardized using the PheRS distributions observed in non-carriers (Methods). Although the average relative increase

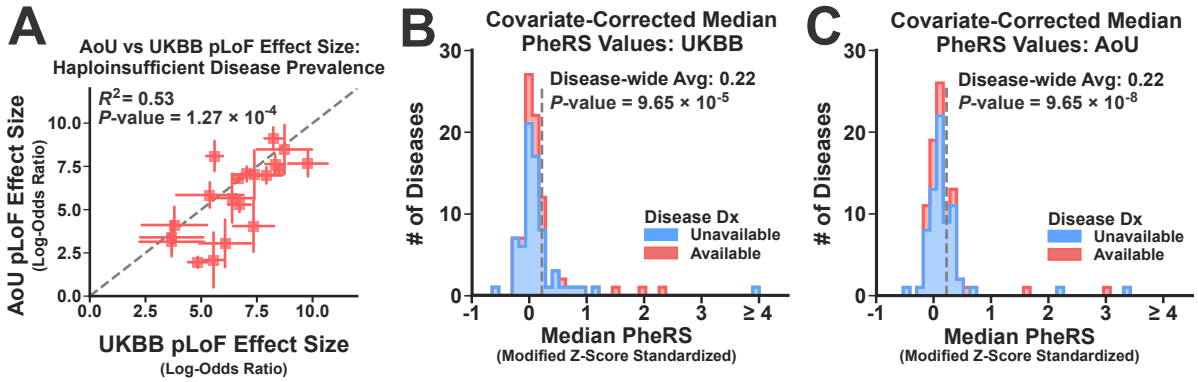


Figure 2: pLoFs have consistent average phenotypic effects in biobanks. (A): Average pLoF effects on disease prevalence are compared across the two biobanks (diseases with diagnoses available in both datasets; N=28). The datapoints represent the mean effect size (log-odds), while the bars indicate standard errors. Correlation was assessed using Pearson's correlation coefficient (R^2). (B, C): The median Phenotype Risk Score (PheRS) was computed using the pLoF carriers for each disease and standardized to the score distributions observed among non-carriers. These histograms display the statistical enrichment of positive PheRS scores among pLoF carriers in the UKBB (B; N=90) and AoU (C; N=90). Statistical significance was assessed using a Wilcoxon signed rank test (H_0: median PheRS is symmetric about $\mu < 0$). Blue color: diagnoses unavailable. Red color: diagnoses available.

in scores among the pLoF carriers was modest (Average Standardized PheRS = 0.22 in both biobanks), the PheRS estimates were systematically increased among carriers across the full set of diseases (60/90 in the UKBB, 68/90 in AoU; Wilcoxon Signed-Rank Meta-Analysis *P*-value = 2.88×10^{-11}). Moreover, many individual diseases achieved dataset-wide (14/89) or at least marginal (37/89) significance in a cross-biobank meta-analysis (see Supplemental Table 6 for full set of results). Therefore, the pLoF variants were associated with increased disease expression risk in at least a fraction of subjects for most diseases.

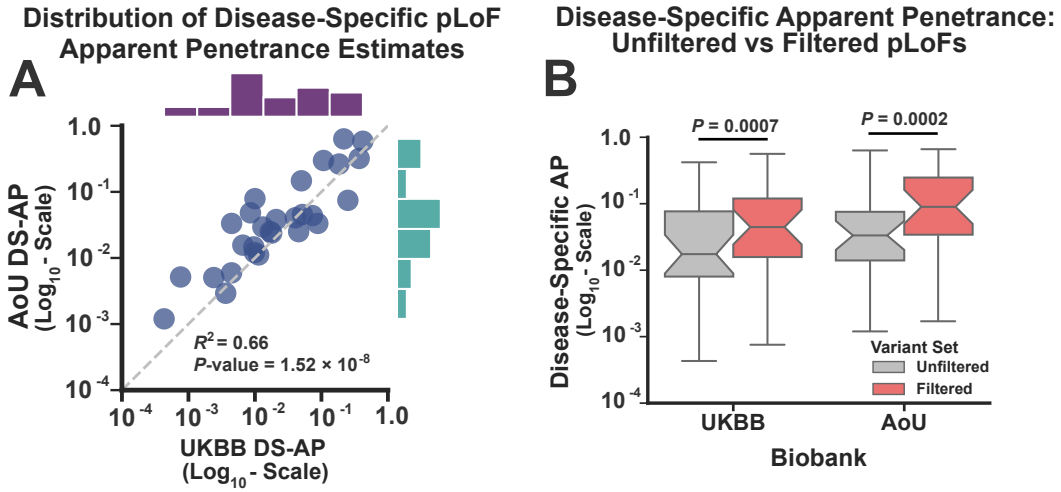


Figure 3: *Reduced pLoF Penetrance is Not Fully Explained by Annotation Artifacts. (A)*: Disease-specific apparent penetrance (DS-AP) estimates for pLoFs (DS-APs, see Methods) were computed in each biobank using diagnoses only (N=28). The estimates from the two biobanks were compared, and their correlation was assessed using Pearson's correlation coefficient (R^2). (E): Boxplots comparing the distribution of DA-AP estimates before (gray) and after (red) filtering pLoFs for possible artifacts (see main text). Statistical significance was assessed using Wilcoxon signed rank tests (one-sided test; H_0: difference in DS-AP after and before filtering is symmetric about $\mu < 0$).

Reduced pLoF Penetrance is Not Fully Explained by Annotation Artifacts

Despite their evident effects on disease expression, the apparent penetrance of the pLoFs was not necessarily high. To estimate the average pLoF penetrance for each haploinsufficient disease, we measured their phenotypic expression using disease-specific diagnostic codes, generating point estimates and 95% confidence intervals using a simple binomial model (see Methods). Figure 3A compares the disease-specific average pLoF penetrance (DS-AP) estimates across the two biobanks, again focusing on those diseases with diagnostic data available in both biobanks (*N*=28). Clearly, this is an imperfect estimate of penetrance, as it makes strong assumptions about the sensitivity of diagnoses for measuring disease expression. Nevertheless, the DS-AP estimates were correlated across biobanks (R^2 = 0.66; *P*-value = 1.52×10^{-8}), and consistent with previous analyses[43], the median DS-AP estimates across diseases were reduced (1.7% ± Median Absolute Deviation=1.6% in UKBB; 3.3% ± 2.2% in AoU). Most studies that analyze pLoFs filter these variants to remove those that likely represent annotation artifacts[53–56], which are variants that have no molecular impact but were

misannotated as LoFs by the variant prediction software. To account for these artifacts, we repeated this analysis after removing variants that impacted non-canonical transcripts (i.e. non-MANE Select[57]) and/or failed to meet a set of quality filters (assessed using the LOFTEE package[54]). Restricting the analysis to the filtered variants increased the DS-AP estimates in both biobanks (23/28 diseases in UKBB; 24/28 in AoU; Wilcoxon Signed Rank Test Meta Analysis *P*-value = 9.54×10^{-7}; see Figure 2E). However, their median values remained below 10% (4.6% ± 3.7% in UKBB; 9.0% ± 7.2% in AoU). Nevertheless, several diseases achieved apparent penetrance estimates exceeding 20% penetrance (ex: Hereditary Hemorrhagic Telangiectasia and Neurofibromatosis Type 1; see Supplemental Table 7 for details. Note: many penetrance estimates in AoU must be suppressed due to restrictions on data sharing).

Reduced Penetrance Persists after Accounting for Missing Disease Diagnoses

Thus far, diagnoses have been used to measure disease expression and estimate penetrance. This is clearly an imperfect method, as some pLoF carriers may exhibit disease expression without diagnoses[45,58]. Moreover, most haploinsufficient diseases lack disease-specific diagnostic codes that can be detected in EHR data. To overcome these issues, we developed an automated method to measure disease expression in every pLoF carrier using their recorded symptoms (see Figure 4A for illustration). Like the PheRS approach, this method computes the background symptom frequency distribution using the entire biobank (denominator in the right-hand-side of the equation in Figure 4A). To compute the probability of disease expression (conditional on being a pLoF carrier, left-hand-side of Figure 4A), the method compares the likelihood of the observed symptoms under a simple disease model (numerator in the right-hand-side) to this background distribution. This enables the method to compute a symptom-driven disease expression score for every pLoF carrier that accounts for their similarity to other affected carriers while comparing their expressed symptoms to a null background. This symptom-driven approach to measuring disease expression requires no labels and is thus unsupervised. However, it remains at risk for overfitting. Therefore, we trained the disease-specific expression models in the UKBB prior to validating them in AoU.

Evaluating the performance of the method is challenging, as no “gold-standard” disease expression dataset exists. Therefore, we used the haploinsufficient diseases with diagnostic data to validate the approach. Briefly, for those diseases with both symptom expression and diagnostic data, we used the symptom expression scores computed for each carrier to predict diagnoses, as they should be mostly concordant. Figure 4B depicts the results of this analysis using precision-recall curves. In these curves, each point on the line represents a distinct symptom-expression score. For each point, we identified those pLoF carriers with a symptom expression score that was at least as high. We then computed the fraction of LoF carriers within this set who harbored disease diagnoses (precision, y-axis) and compared this estimate to the total fraction of diagnosed LoF carriers detected within the subset (recall, x-axis). A perfect model would recover 100% of the diagnosed carriers with perfect precision (red-dotted lines). A random model could perform no better than the baseline disease diagnosis rate (gray dotted lines).

Clearly, the symptom-driven expression measurements predicted disease diagnoses significantly better than random (13-fold and 7.5-fold better in the UKBB and AoU, respectively; Randomization Test *P*-values < 0.0001, see Figure 4B). Moreover, there was not a substantial difference in model performance between the UKBB and AoU, although performance was certainly better in the UKBB (likely due to training bias). This suggests that symptom-driven expression scores are predictive of disease expression, at least according to diagnoses. To simply downstream analyses, we binarized the symptom-driven expression scores by selecting a single threshold for disease expression in each dataset. Those carriers with symptom scores above this threshold were deemed to have disease expression, while those with scores below the threshold were unexpressed. To determine this disease expression threshold, we chose the symptom scores that maximized the F_1-measure for the curves shown in Figure 4B, where the F_1-measure represents the harmonic mean of the precision and recall scores. Selecting a disease expression threshold in this manner is arbitrary without "gold-standard" data. But by using the symptom-expression score that maximized the F_1-measure, we found no strong evidence for discordance between the symptom-driven measurements and those derived using diagnoses (McNemar's Test *P*-values = 0.16 and 0.12 in the UKBB and AoU respectively).

To incorporate the symptom-driven measurements into penetrance estimates, we identified a pLoF as expressed if the carrier had a either disease diagnosis or if their symptom-driven score exceeded the F_1 thresholds from Figure 4B. For the set of 28 diseases analyzed in Figure 3, incorporating the symptom-driven expression measurements increased the median DS-AP estimates by 1.64-fold and 2.01-fold in the UKBB and AoU, respectively (Figure 4C). This is not surprising and is consistent with the hypothesis that many pLoF carriers were symptomatic but lacked formal disease diagnoses. In addition, this symptom-driven approach allowed us to analyze 51 haploinsufficient diseases that lacked diagnostic information yet had symptoms that were shared across the two biobanks. For these diseases, the purely symptom-driven DS-AP estimates were correlated across biobanks (Figure 3E; R^2 = 0.46; *P*-value = 1.22×10^{-8}). Nevertheless, overall pLoF penetrance was reduced (median DS-AP: 4.8% ± 3.5% and 9.2% ± 7.5% in the UKBB and AoU respectively), even for diseases with that had both diagnostic codes and symptom-driven expression data (5.2% ± 4.5% and 10.9% ± 9.9% respectively). The symptom-driven DS-AP estimates are provided in Supplemental Table 8.

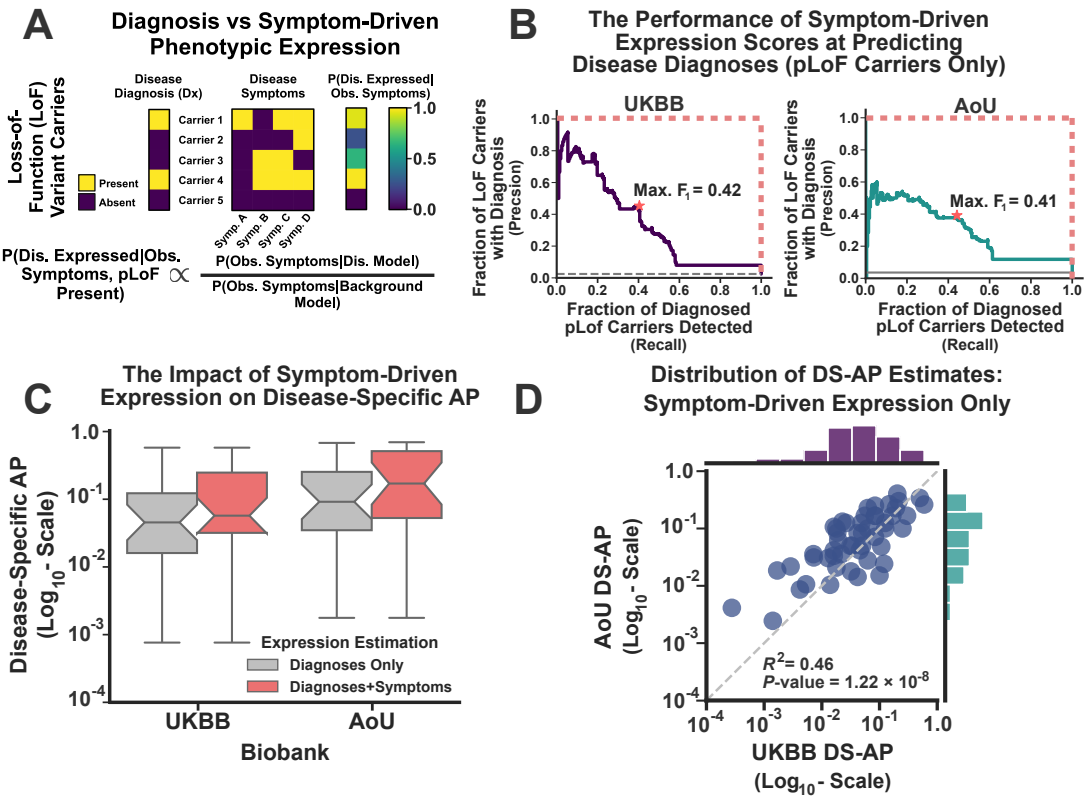
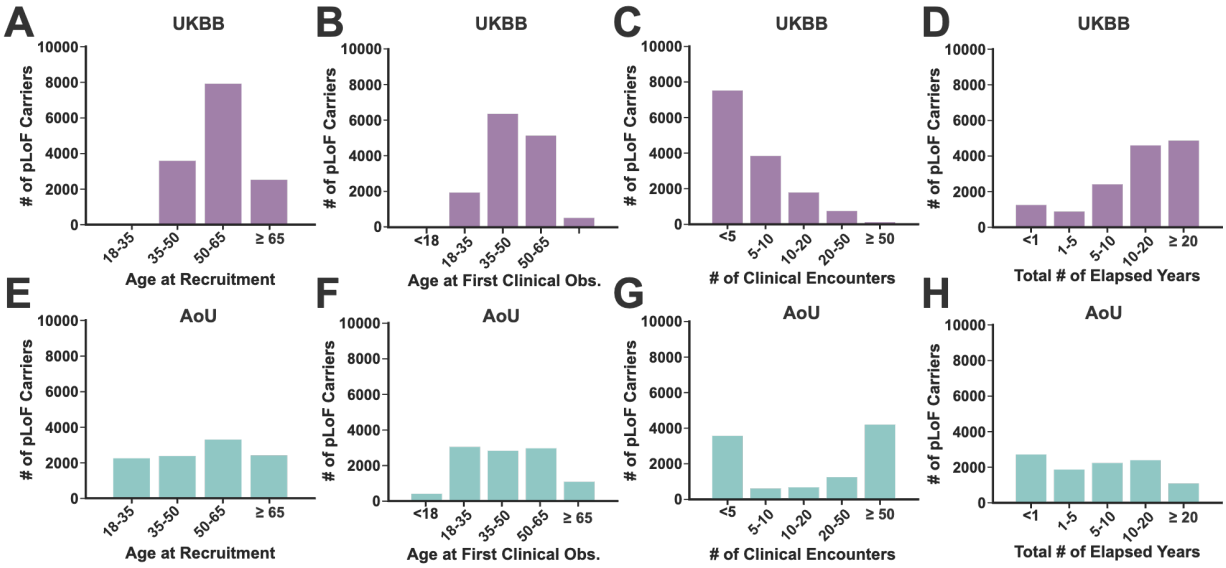


Figure 4: Reduced Penetrance Persists after Accounting for *Missing Disease Diagnoses.* (A): A simple illustration of the symptom-driven expression model. (B): Symptom-driven expression scores predict rare disease diagnoses in both datasets. These panels display the precision (i.e. fraction of LoF carriers with diagnoses; y-axis) and the recall (i.e. total fraction of pLoF carriers detected; x-axis) for a model that uses symptom-driven expression scores to identify pLoF carriers with disease diagnoses. Each point on these curves represents a different expression score threshold for identifying carriers at-risk for diagnosis. The red stars denote the expression score (≈0.97 in both datasets) that maximized the F_1-measure (harmonic mean of precision/recall) for the predictions. Gray lines indicate the performance of a random model; red-dashed lines the performance of a perfect classifier. The left panel depicts performance in the UKBB (training dataset; N = 6,129 disease-carrier pairs), while the right panel depicts performance in AoU (validation dataset; N = 8,242 disease-carrier pairs). Additional details are in the text. (C): For those diseases with diagnoses available in both biobanks (N=28), the two boxplots compare the disease-specific penetrance estimates before (gray) and after (red) including the symptom-driven scores (see Methods). (D): For those diseases without diagnostic data (N=51), the symptom-driven disease-specific APs for the pLoF variants were compared across the two biobanks. Correlation was assessed using Pearson's method (R^2).

Reduced pLoF Penetrance Persists After Accounting for Gaps in EHR Data Coverage

Another factor that could negatively impact penetrance estimation is the incomplete lifetime coverage of the EHR data in biobanks[41,45]. More specifically, these datasets capture only a fraction of their subjects' lifespans, and the data coverage for individual participants can vary widely. Therefore, extensive missing clinical data, which can occur during any lifetime interval (e.g. childhood, late adulthood, etc.), can give the false appearance of reduced penetrance. Supplemental Figure 1 shows the distributions

over the age at the time of recruitment (1A,1E), the age of the earliest clinical observation (1B, 1F), the total number of documented clinical encounters (1C,1G), and the total number of years elapsed between the first and last encounters (1D,1H) for the pLoF carriers in the UKBB and AoU, respectively. Based on these distributions, it seems likely that clinical data coverage for some pLoF carriers was too low to reliably determine their disease expression. To test this hypothesis, we first determined whether each pLoF carrier expressed *any* symptom consistent with their associated haploinsufficient disease, identifying those without symptoms as being *completely asymptomatic*. We then used the attributes from Supplemental Figure 1 as features in a set of regression models designed to predict asymptomatic carriers, independent of any variant or gene information. If the models were accurate, then prediction scores derived from these data coverage statistics could be used to remove pLoF carriers with too little clinical data to make meaningful contributions to penetrance.



Supplemental Figure 1: Clinical data coverage among pLoF carriers in the UKBB and AoU. (A, E): Distribution over the age at the time of recruitment (A: UKBB; E: AoU). (B, F): Distribution over the age at the first clinical observation (B: UKBB; F: AoU). (C, G): Distribution over the total number of clinical visits (C: UKBB; G: AoU). (D, H): Distribution over the total number of elapsed years between the first and last clinical visits (D: UKBB; H: AoU).

To perform this analysis, we separated the diseases into three groups based on their approximate onset (see Methods): childhood, young adulthood, and adulthood, as we expect that these onset classes to have different coverage requirements (e.g. childhood-onset conditions may be poorly assessed in individuals without clinical data extending below age 20). We then fit and assessed the performance of these models in both biobanks using 5-fold leave-one-out cross validation. The results are depicted in Figure 5A, where performance was assessed by comparing the model predictions to true asymptomatic status using receiver operating characteristic curves. For all three onset classes, the models were predictive of asymptomatic status, with some variability in performance across the different onset classes and biobanks.

Using the predictions from these models, we removed all subjects from our analysis who were predicted to be asymptomatic based on limited clinical data coverage

at a false positive rate of 5% (see Methods). This process filtered out 17% and 35% of the pLoF carriers in the UKBB and AoU respectively. Figure 4B depicts the increase in DS-AP for the pLoFs within both datasets after correcting for clinical data coverage. DS-AP estimates increased for most diseases in both biobanks (65/90 in the UKBB; 64/76 in AoU; Wilcoxon Signed Rank Test Meta Analysis *P*-value < 2.2×10^{-16}). Moreover, a handful of individual diseases acheived pLoF penetrance rates exceeding 50%, including Autosomal Dominant Polycystic Kidney Disease (Avg. Penetrance: 53% and 70% in the UKBB and AoU, respectively) However, this was not universally true, even for diseases that typically present during childhood (average pLoF penetrance for Tuberous Sclerosis: 6.6% and 19.2% in the UKBB and AoU respectively). Moreover, the absolute increase in DS-AP was modest, with the median penetrance estimates increasing 1.3-fold from those depicted in Figure 4 for both biobanks. The set of DS-AP estimates after filtering for clinical data coverage are provided as Supplemental Table 9.

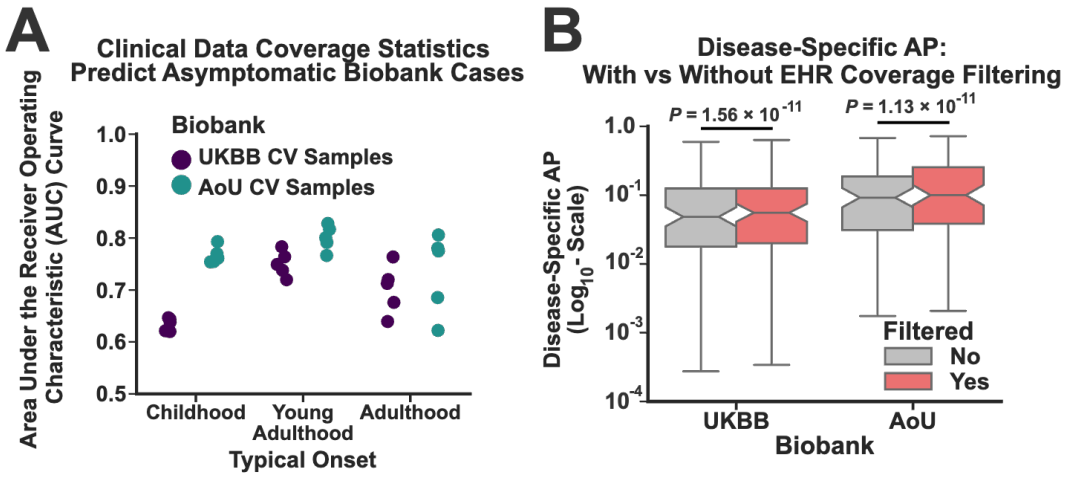


Figure 5: Correcting for clinical data coverage does not substantially increase penetrance estimates. (A): The coverage statistics displayed in Supplemental Figure 1 were used to build models to predict whether the pLoF carriers (stratified by disease onset; x-axis) would be asymptomatic for their target phenotypes. Model performance was assessed using the area under the receiver operating characteristic curve (AUC; y-axis) using leave-one-out 5-fold cross validation (CV). All models consistently performed better than random (AUC=0.5). (B): Based on the results from (A), pLoF carriers predicted to be asymptomatic at a false positive rate of 5% were removed from the analysis (see Methods). These boxplots depict the DS-AP estimates before (gray) and after (red) this filtering. Statistical significance was assessed using a Wilcoxon signed rank test (one-sided test; H_0: difference in DS-AP estimates after and before filtering are symmetric about $\mu < 0$).

Variant-Specific Genomic Features Are Predictive of pLoF Penetrance

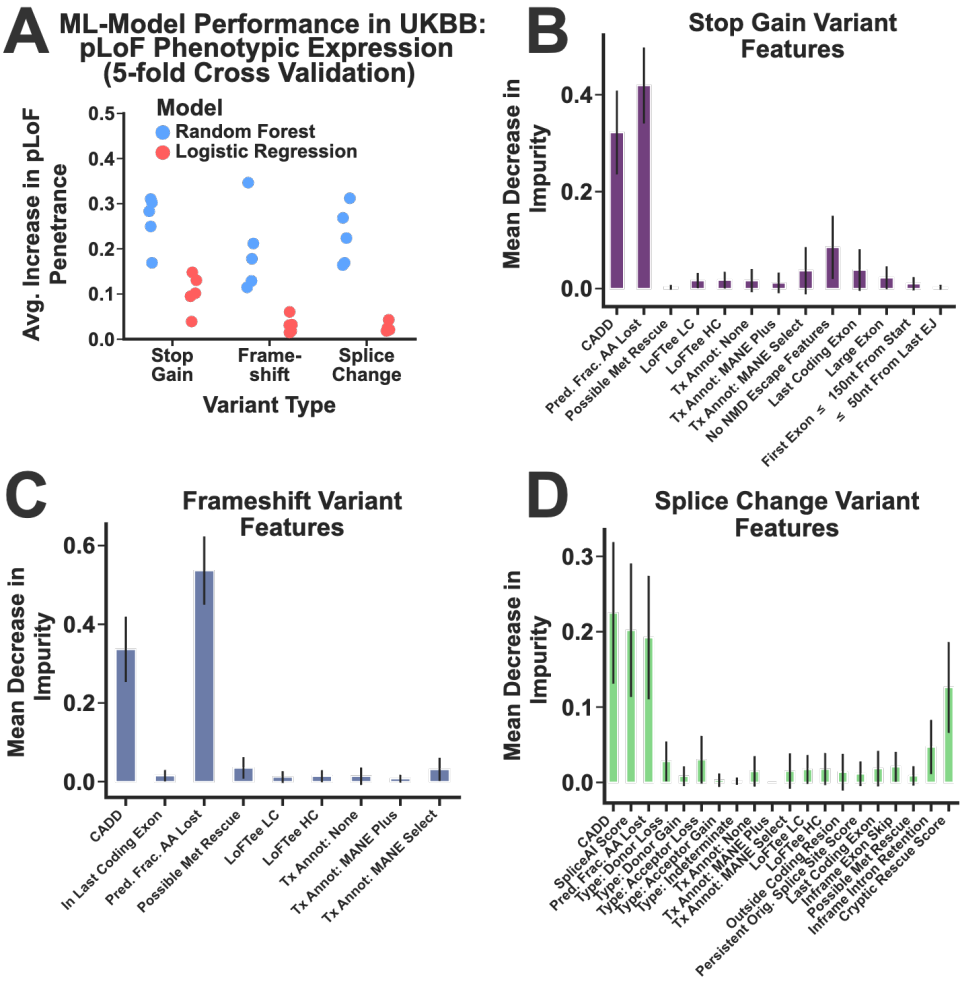
After removing annotation artifacts, imputing missed diagnoses, and correcting for incomplete clinical data coverage, pLoF penetrance estimates increased systematically. However, reduced penetrance remained common. None of the filters employed above were perfect, and residual artifacts, missed diagnoses, and incomplete

data could certainly account for some fraction of the reduced penetrance. However, it's certainly possible that some of the reduced penetrance was instead driven by pLoFs with incomplete loss-of-function. Generally, it is presumed that loss-of-function variants are approximately equivalent in terms of their molecular impact, but this may not be the case. Some may have deleterious effects but still result in some residual gene or protein activity. Such “leaky” expression could in turn drive variable penetrance. For example, incomplete non-sense mediated decay escape[59,60] could allow for the partial expression of a transcript impacted by a stop-gain variant. Detecting such an effect for an individual variant is challenging, at least using biobank data. However, we hypothesize that variant-intrinsic genomic features (ex: SpliceAI scores for splicing variants[61]), which are often used to identify annotation artifacts, may also be predictive of penetrance. If true, then variant-intrinsic genomic features should be able to identify subsets of pLoFs with high penetrance, even among variants with prior evidence for pathogenic effects according to diagnostic testing data.

To test this hypothesis, we constructed machine learning models[62] that used multiple variant-intrinsic genomic features to predict disease expression in individual pLoF carriers, where expression was measured using both disease-specific symptoms and diagnoses (see Methods for full description). Because the interpretation of different variant types relied on different features, we constructed unique models for each of the following pLoF classes: stop-gain, frameshift, and splice change (see Methods). These models were trained to predict disease expression exclusively within the UKBB (see Supplemental Figure 2 for a summary of the UKBB model training results), yielding variant expression prediction models that could be independently validated in AoU. To validate their effectiveness, we used the models to predict the expression risk for every pLoF variant in AoU. We then selected subsets of AoU pLoFs according to their predicted expression probabilities. If variant-intrinsic features were predictive of penetrance, then the average apparent penetrance of the pLoFs within the subsets should increase as the scores become more selective. At the same, increasingly restrictive expression scores will retain fewer expressed pLoFs, meaning that the total fraction of expressed pLoFs captured by the subset will decrease.

Figure 5A displays this tradeoff between average penetrance and retained fraction of expressed pLoFs over the range of machine learning (ML)-derived expression scores within the AoU validation dataset. For reference, we display this same tradeoff for the variants that pass the filter that we used in Figure 2 to remove annotations artifacts (MANE Select transcipts only[57] with a high confidence LOFTEE flag[54]). A filter that only includes variants with non-conflicting pathogenic/likely-pathogenic (P/LP) annotations in ClinVar[46] is also displayed. Clearly, the machine learning predictions can select pLoFs with progressively increasing penetrance, and on average, the model prediction scores perform better than both the simple filter (ML Model Penetrance/Recall F_1 measure = 0.26; MANE+LOFTEE F_1 measure = 0.17; Bootstrapped *P*-value < 1.0×10^{-4}) and prior P/LP annotations (ML Model Max. F_1 measure = 0.26; P/LP F_1 measure = 0.19; Bootstrapped *P*-value < 1.0×10^{-4}). In Figure 5B, the results are stratified by variant type. The models were significantly predictive for all types (Randomization Test *P*-values < 1.0×10^{-4}) but splice prediction was the most challenging (ML Model Avg. Penetrance Increase = 0.07). Figure 5C displays this same information, except now the results are stratified by disease onset. Even pLoFs in adult-

onset haploinsufficient disease genes can be filtered to maximize penetrance (ML Model Avg. Penetrance Increase from Baseline = 0.09; Randomization Test *P*-value < 1.0×10^{-4}).



Supplemental Figure 2: Training variant expression prediction models in the UKBB. (A): We evaluated the performance of two different machine learning frameworks for predicting disease expression using pLoF features: Logistic Regression (red) and Random Forests (blue). We assessed performance using the average increase in penetrance achieved by filtering pLoFs according to their expression scores. Random forest models consistently outperformed logistic regression in 5-fold leave-one-out cross validation experiments for all three variant types considered in this study. (B, C, D): In these panels, the variant-intrinsic features used by each model to predict disease expression are displayed on the x-axis. The y-axis displays an approximate measurement of feature importance (the mean decrease in impurity achieved by each pLoF feature). Error bars indicate the standard errors. A: Stop-gain; B: Frameshift; C: Splice Change.

Figure 5D displays this penetrance-recall tradeoff exclusively for variants classified as P/LP in ClinVar. The machine learning models remained strongly predictive of penetrance for this class of variants (ML Model Avg. Penetrance Increase from Baseline = 0.23; Randomization Test *P*-value < 1.0×10^{-4}), with the most stringently

filtered pLoFs approaching near complete penetrance. Based on this data, it's unlikely that missing clinical data is accounting for much of the apparent variability in penetrance, as it's hard to imagine why differences in clinical data coverage would correlate with variant-specific genomic features independently of their effects on disease expression. Instead, there are two much more likely (and non-mutually exclusive) explanations for this result[41]. First, many P/LP variants may have variable penetrance, which correlates with their genomic features. Second, the annotations in ClinVar may be incorrect. Distinguishing between these two possibilities is difficult, as most of these variants occur in only 1-2 subjects. Nevertheless, these results suggest that binary pathogenicity labels have low utility when it comes to predicting penetrance.

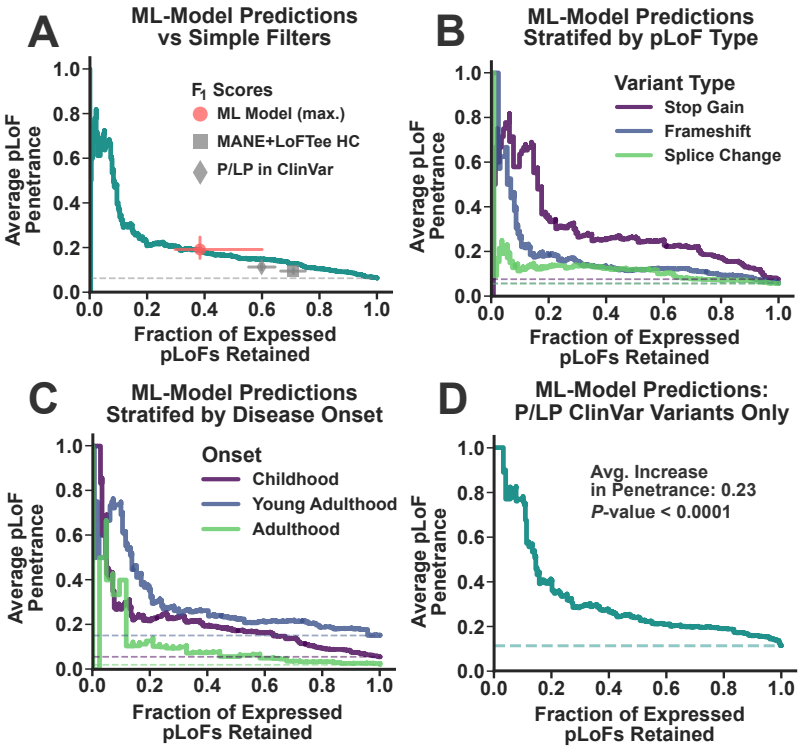


Figure 6: Genomic features are predictive of penetrance. (A): This panel displays the tradeoff between average apparent penetrance (y-axis) and the fraction of expressed pLoFs retained (x-axis) when using different model-derived expression prediction scores to filter variants. For reference, the threshold that maximized F_1 measure for the model is shown in red. The performance of a simple filter (MANE Select+High-Confidence LOFTEE flag) is shown as a gray square, while a filter that only includes variants with nonconflicting pathogenic/likely-pathogenic (P/LP) annotations in ClinVar is shown as a gray diamond. All error bars represent bootstrapped 95% confidence intervals. The dotted line indicates the performance of a random classifier. (B, C): These panels depict the same penetrance- recall results from (A), except now pLoFs are stratified by variant type (B) or typical disease onset (C). (F): This panel depicts penetrance-recall curve for those pLoFs with non-conflicting P/LP annotations in ClinVar. Significance was assessed using a randomization test.

Discussion

Broad (exome and genome) sequencing has revolutionized the field of clinical genetics[63]. More rare disease patients are being diagnosed using these technologies, and this has improved our ability to provide timely counseling and treatment to the affected individuals and their families. As a result, there is growing interest in using broad genomic testing for population screening[7–17]. Here, the goal is to identify individuals at risk for a Mendelian disease prior to symptom onset. Theoretically, this could lead to better clinical outcomes through earlier diagnosis, surveillance and management. For genomic screening to succeed, it should at least have a quantifiable positive predictive value[18–20], a statistic that directly depends on penetrance. Currently, penetrance is almost universally unknown except for a handful of unusually frequent, deleterious variants. As a result, Mendelian disease risk assessments will be imprecise for most asymptomatic carriers. This may have a limited impact on patient outcomes in many settings. However, the clinical decisions made using genomic screening will be life altering in some cases, and without penetrance information, such interventions may be unnecessarily applied to low-risk carriers.

In this study, we investigated the penetrance of one of the simplest classes of clinically relevant genetic findings: putative loss-of-function variants (pLoFs) in haploinsufficient disease genes. Consistent with prior analyses[42,43,58], we found that the apparent penetrance of these variants was reduced, with median values ranging from 5-10% depending on the biobank (Figure 3). Accounting for the extensive amount of missing clinical data in biobanks increased penetrance estimates (Figures 4 and 5), but most pLoFs remained unexpressed. In diagnostic applications, detailed criteria for variant interpretation have been developed to mitigate the risk for false positive results[27,64]. The utility of these criteria for penetrance prediction, however, is largely unknown. In this analysis, even variants with prior evidence for pathogenicity based on diagnostic testing had reduced penetrance (Figure 6A), suggesting that the utility of these annotations for penetrance prediction is limited. Importantly, machine learning models that incorporated variant-intrinsic genomic features like mutational constraint[65], splicing scores[66], and predicted non-sense mediated decay escape[59] were able to identify pLoFs with penetrance approaching 100% (Figure 6D), indicating that missing clinical data alone was unlikely to account for the reduced penetrance for many these variants.

These results suggest that screening tests for these disorders that rely on current variant interpretation guidelines will have low positive predictive values. That said, improvements seem feasible. Decades of research into variant-intrinsic features like mutation patterns, evolutionary constraint, and functional impact has led to the development of computational tools that are used to predict the pathogenicity of individual pLoFs[54,55,59,65,66], generally with the goal of eliminating annotation artifacts. However, the analyses performed in this study suggest that these tools may be effective at predicting variant penetrance, even in the absence of gene, disease, and carrier-specific information. In addition, their effectiveness at this task suggest that these tools may be capturing some degree of “leaky” or incomplete loss-of-function, which has implications for rare variant analyses beyond penetrance prediction. That said, the real-world clinical validity of these predictions remains unknown, and much work remains to

be done to ensure that penetrance predictions derived from biobanks and similar resources are replicable, calibrated, minimally biased, and broadly applicable to diverse genes, diseases, and populations.

For now, it may be possible to predict the phenotypic outcomes for some rare genotypes with near complete penetrance[67]. In addition, when variant information is combined with orthogonal data like enzymatic activity and biomarkers, the prognostic accuracy may be very high[68]. Unfortunately, such assays are only available for a tiny fraction of genetic diseases. For most, the variants themselves are the only piece of prognostic information available. Prior evidence for pathogenicity may increase the positive predictive value of a particular variant, but based on the analyses presented here, prior pathogenic annotation labels are not synonymous with high penetrance, which is not unexpected. Access to high-quality outcome data for individual genotypes will certainly help, but given their intrinsically low frequency, it will likely remain difficult to estimate penetrance and predict phenotype outcomes in individual patients for the foreseeable future. Therefore, we suggest that caution be used when returning positive genomic findings to asymptomatic patients. Even with prior evidence of pathogenicity, risk estimates remain uncertain.

Methods

Haploinsufficient Disease Curation and Annotation

We used the ClinGen[47] Database (downloaded on July 25th, 2023) to identify Mendelian disorders that have strong evidence to support haploinsufficiency as a mechanism of disease (ClinGen Dosage Haploinsufficiency Assertion Evidence Level 3). We then aligned these diseases to the Online Mendelian Inheritance in Man[48] database (downloaded on February 23, 2023) using simple string matching followed by manual curation. This yielded 91 autosomal dominant/pseudo-autosomal dominant diseases linked to 117 genes, which were manually annotated with their typical onset (Childhood, Young Adulthood, Adulthood) and general classification (Congenital Malformation, Isolated Neurodevelopmental, Complex Neurodevelopmental, Tumor Predisposition, and Other) by a board-certified clinical geneticist (author D. Blair) using clinical expertise and literature review. Afterwards, disease-specific diagnostic codes were annotated to these diseases by manually curating the terminologies[69] used by the Observational Medical Outcomes Partnership Common Data Model[70] (OMOP-CDM). Finally, the diseases were annotated with a set of Human Phenotype Ontology[71] (HPO) symptoms using the data from several ontologies, including the HPO itself (downloaded on February 21, 2023), the Disease Ontology[72] (downloaded on February 23, 2023) and OrphaNet[73] (downloaded on February 23, 2023 using the `HOOM`[74] module). The sequence and transcript information for each of the 117 genes was downloaded from the Ensembl[75] database (Release 109; GRCh38 assembly) using the `PyEnsembl`[76] package. Additional gene and transcript information (exon-intron boundaries, 5' and 3' UTRs, full coding and amino acid sequences) was downloaded using `gget`[77]. The 91 haploinsufficient diseases, along with their annotated information, are provided in Supplemental Table 1.

Aligning HPO Symptoms to the OMOP-CDM Terminology

To identify HPO[71] symptom diagnoses in the EHR data, we needed to align this ontology to the structured diagnostic data available in the electronic health records of each biobank. Because both biobanks encode their clinical data using the OMOP-CDM[70], we focused on aligning the HPO symptom terminology to the structured vocabulary used by this data model[69]. Unfortunately, aligning the HPO to other medical terminologies is largely an unsolved problem that lacks a consensus regarding best practices[78]. Therefore, we created a custom alignment by building on our previous work[24] while implementing some new techniques.

First, we created an alignment map between the HPO and SNOMED-CT[79], as the latter represents the most comprehensive medical terminology available for the dissemination of EHR data. It is also fully incorporated into the concept terminology used by the OMOP-CDM. To create an HPO-to-SNOMED map, we followed the approach of McArthur et al.[80], who created a similar map between the HPO and PheCodes[81]. First, we constructed a map linking HPO to SNOMED-CT terms if they shared a common concept in the UMLS Metathesaurus[82]. Second, we used an ontology alignment algorithm (SORTA[83]) to find all SNOMED-CT terms that mapped to an HPO

term with a similarity score of ≥ 0.8 for at least 1 of their associated string pairs (both SNOMED-CT and HPO often provide multiple strings for each term). For terms with multiple aligned string pairs, we collated all the similarity scores across the different string pairs, storing both an average and maximum score.

With an HPO-to-SNOMED map in place, the HPO terms themselves could be aligned directly to the concept terminology used by the OMOP-CDM, as a map from SNOMED-CT terms to the OMOP-CDM concepts is provided by Observational Health Data Sciences and Informatics (OHDSI) Collaborative[69,84]. However, it is important to note that one HPO term often mapped to multiple SNOMED-CT terms, which could then map to the OMOP terminology in multiple ways. Therefore, each HPO-OMOP alignment was often supported by multiple intermediary relationships. To summarize this phenomenon, we stored several pieces of information for each alignment that captured the quality of its supporting evidence. These included: the total number of intermediate relationships supporting the mapping, the fraction of these relationships that were supported by the UMLS, the fraction that achieved a SORTA string alignment similarity score ≥ 0.8, the average SORTA score across intermediaries, and maximum score achieved. In total, this process generated 35,825 unique HPO-to-OMOP alignments.

Because automated alignments like this tend to be rife with spurious results, one of the authors (D. Blair) manually reviewed 500 random mappings and annotated their medical accuracy. The accuracy was unsurprisingly variable, but overall, far better than random (average precision: 0.76). To further improve accuracy, we built a simple logistic regression classifier (implemented in `sklearn`[85]) to predict if an HPO-OMOP alignment was accurate. The model incorporated the alignment features described above as linear predictors (noting that the maximum achieved SORTA score was incorporated as interaction term with the total number of intermediate relationships). The model was trained on the 500 manually annotated alignments prior to being applied to the full dataset. In leave-one-out 5-fold cross validation experiments, the area under the receiver operator characteristic curve for the model predictions was 0.76 (standard error: 0.017), indicating that these predictions could provide a substantial improvement to alignment accuracy. Therefore, all ≈35,000 HPO-to-OMOP alignments were scored using the prediction model, and several false positive rate (FPR) thresholds were selected for downstream filtering. The complete set of HPO-to-OMOP Concept ID alignments (along with their features, manual annotations, machine learning scores, and whether they survived various FPR filtering thresholds) are provided as Supplemental Table 10. Finally, we experimented with various alignment FPR thresholds in downstream analyses. Overall, PheRS enrichment among pLoF carriers was highest when using the relationships that survived a 20% FPR threshold (data not shown). Therefore, this set of alignments was used for all the results reported in this manuscript.

Sequence Data Quality Control, Variant Annotation, and Non-Carrier Cohort Identification

This study utilized the exome sequence (ES) data from the UK Biobank (UKBB)[36] and the whole genome sequence (WGS) data from the All of Us (AoU) Research Program[40] to investigate the penetrance of putative loss-of-function (pLoF) variants in haploinsufficient disease genes. For the AoU dataset, the WGS samples undergo an

extensive quality control process, which ensures that samples meet several coverage and accuracy thresholds[40]. Therefore, all samples with WGS data that were not flagged by AoU's quality control pipeline were analyzed in this study (*N* = 245,376). For the UKBB, less sample-level quality control was performed *a priori*. Therefore, the ES data from this biobank underwent additional quality control filters consistent with those performed in previous studies[86]. Briefly, all samples that showed evidence for genetic and self-reported sex discordance (*N* = 296), sample duplication (*N* = 56), excessive SNP array-short read sequencing genotype discordance (*N* = 513, including those that lacked array data), low read coverage for the haploinsufficient genes of interest (20x coverage at less than 90% of the base pairs; *N* = 92), and excessive missing genotypes (*N* = 329, again limited to the haploinsufficient genes of interest) were excluded from the analysis (total number of samples that failed quality control: 1,156). After excluding subjects that withdrew from the UKBB study, this dataset contained a total of 468,672 subjects with both ES and EHR data.

Following sample-specific quality control filtering, variants from the exonic regions of the haploinsufficient disease genes were isolated from both datasets (performed using `bcftools`[87] in the UKBB and `hail`[88] in AoU), storing the variant genotyping calls in VCF files. Individual-level data was then stripped from these files, and the predicted molecular effect of each variant was annotated using VEP[89] (Version 110). Simultaneously, the variants were annotated with any previous interpretations documented within the ClinVar[46] database (downloaded on May 13th, 2024). Finally, all pLoFs within these datasets were identified using the LOFTEE[54] plug-in for VEP, which also provided a flag indicating the overall confidence in this assessment (high vs low confidence; HC vs LC). Using the output from VEP, each pLoF was annotated with the its most clinically significant impacted transcript[57] (MANE Select, MANE Plus Clinical, Other), and the variants were also assigned to one of three pLoF classes: stop-gain, frameshift, and splice change.

Following annotation, we returned to the VCF files that contained the individual-level genotype calls and isolated all pLoFs identified in the previous step. We then identified all carriers for each individual variant, removing those that did not meet a basic set of genotype-specific quality control filters[86]. For single nucleotide variants (SNVs), we assigned a no-call status to all carriers with a genotyping quality score < 30, sequencing depth < 7, and alternate allelic balance < 0.15. For insertion-deletion variants, we were more stringent and removed those calls with a quality score < 30, sequencing depth < 10, and alternate allelic balance < 0.2. In addition, we removed a variant from the analysis entirely if its call rate was <0.99 or if its carrier frequency was greater than 0.1% (after performing carrier-specific quality control). For the UKBB, we also *a priori* removed those variants that achieved an average read depth <10 for more than 10% of the samples in the dataset (per recommended best practices[90]). In total, this process identified 3,131 (Supplemental Table 2) and 3,889 (Supplemental Table 3) pLoFs carried by 14,010 and 11,022 subjects in the UKBB and AoU respectively. Note, some individuals harbored multiple pLoF variants within a single gene, suggesting the potential for *in cis* rescue versus sequencing artifacts. Such carriers were not excluded from basic pLoF frequency estimates (i.e. Figures 1) but were excluded from all other analyses. No further attempts were made to account *in cis* rescue events, and the extent of their impact on pLoF penetrance is a target for future work[56].

Finally, for each haploinsufficient disease, we created a unique cohort of non-carrier controls that were unlikely to be at risk for the disease of interest. To do so, we first identified all subjects in both datasets that carried any rare variant (allele frequency ≤ 0.1%, performed using `plink2`[91] or `hail`[88]) in the set of genes annotated to each disease. From the set of all possible control subjects, we then removed those that carried any rare variant in the target genes or had a no-call genotype at one of the pLoFs detected in those genes. The total number of non-carrier controls available for each disease was variable but exceeded 280,000 and 130,000 in all instances for the UKBB and AoU respectively.

Recruitment Age Analysis

We hypothesized that the ascertainment biases intrinsic to biobank recruitment would result in differences in recruitment age between pLoF carriers and non-carriers. To test this, we first identified the recruitment age for every subject. For the UKBB, recruitment age is a specific entry in the dataset (Data-Field 21022). For AoU, we estimated recruitment age using the censored birthdate provided for each subject along with the date on which their genomic biospecimen was collected. The pLoF effects on recruitment age were estimated separately for each haploinsufficient disease using an ordinary least squares (OLS) regression model applied to the cohort of pLoF carriers plus their corresponding non-carrier controls:

$$\vec{Y}_{\text{Recruitment Age}} = \mu + \vec{G} \times \beta_{\text{pLoF}} + \epsilon, \qquad (1)$$

where $\vec{G}$ denotes a binary vector indicating pLoF carrier status, β_{pLoF} indicates the disease specific pLoF effect on recruitment age, μ is an intercept term, and ϵ denotes a gaussian-distributed error term. Dataset-specific *P*-values were computed using a two-sided Student's *T*-test (using the `statsmodels`[92] package in Python), and inference results ($\hat{\beta}_{\text{pLoF}}$, $\hat{\sigma}_{\text{pLoF}}$) from the two biobanks were combined using a fixed-effects meta-analysis[93]. Disease-wide effects were quantified by taking mean of the pLoF effect estimates across diseases, and the significance of the disease-wide bias in recruitment age was assessed using a two-sided Wilcoxon Signed Rank Test. Results were combined across biobanks again using a fixed-effects meta-analysis. Supplemental Table 4 contains the complete set of results for the recruitment age analysis.

Carrier Rate Analysis

Let $C_i = 1$ indicate that the *i*th biobank subject is a carrier for at least one pLoF. Based on this definition, the pLoF carrier rate in each biobank is given by:

$$\text{Carrier Rate} = \frac{\sum_{i=1}^{N} C_i}{N}.$$

To correct carrier rates for genetic ancestry differences, we used the predicted ancestry labels provided by the All of Us Research program, which were assigned using a machine learning model trained on a set of reference samples with known ancestry[40]. We reproduced these ancestry assignments in UKBB using this same procedure. To

correct carrier rates for ancestry differences, we included these ancestry labels as covariates in a logistic regression model to predict carrier status:

$$\boldsymbol{P}\left(\vec{C}\middle|\rho, \boldsymbol{A}, \vec{a}\right) = F(\rho + \boldsymbol{A} \times \vec{a}), \tag{1}$$

where $\boldsymbol{A}$ denotes a matrix of predicted ancestry labels (for this project, predicted ancestry labels include African/African American, Admixed American, East Asian, South Asian, Middle Eastern, and European), $\vec{a}$ denotes the vector of their individual effects, and $F(X)$ denotes the logistic function. The parameter ρ represents the baseline pLoF carrier rate after adjusting for ancestry.

Haploinsufficient Disease Prevalence and Penetrance Analysis

The simplest way to measure pLoF phenotypic expression was using disease diagnoses. Let $\vec{D}$ denote a binary vector of length *N*, where in *N* is the number of pLoF carriers for some haploinsufficient disease of interest plus the number of non-carrier controls. Let $D_i = 1$ denote that the *i*th subject was diagnosed with the disease of interest at least once in their EHR data, where diagnoses were identified using a set of manually annotated OMOP-CDM concept codes (see above). Finally, let $\vec{G}$ denote a binary vector indicating the pLoF carrier status for the *N* subjects. We estimated the biobank-specific pLoF effect sizes (log-odds ratios; denoted γ_{pLoF}) using one of two approaches. For the more common diseases ($\sum_{i=1}^{N} D_i \geq 10$), we incorporated covariates into the analysis using the following log-linear model:

$$\boldsymbol{P}\left(\vec{D}\middle|\mu, \vec{G}, \gamma_{\text{pLoF}}, \boldsymbol{X}, \vec{a}\right) = F\left(\mu + \vec{G} \times \gamma_{\text{pLoF}} + \boldsymbol{X} \times \vec{a}\right), \tag{2}$$

where F denotes the logistic function, μ is an intercept term, $\boldsymbol{X}$ is a matrix of covariates, and $\vec{a}$ is a vector of covariate effect size parameters. For the current study, we incorporated the following covariates into our analysis: recruitment age, birth sex, and the first 16 principal components of the genetic relatedness matrix. Model fitting was performed using Firth-penalized maximum-likelihood estimation[94], and statistical inference was conducted using a likelihood ratio test. Even with Firth penalization, model inference returned spurious results when $\vec{D}$ was extremely sparse. Therefore, for very rare diseases ($\sum_{i=1}^{N} D_i < 10$), we constructed 2×2 contingency tables from $\vec{D}$ and $\vec{G}$ and estimated the pLoF log-odds ratio and its corresponding standard error using the `statsmodels`[92] software package in Python. *P*-values were estimated using Fisher's exact test[95]. Finally, we performed cross-biobank meta-analyses of the pLoF effect sizes using the Cochran-Mantel-Haenszel Test for stratified contingency tables (again implemented in the `statsmodels`[92] package). Supplemental Table 5 contains a summary of the results of our disease-specific prevalence association analysis.

To estimate the disease-specific pLoF apparent penetrance (DS-AP) estimates using diagnoses, we assumed a symmetric beta prior distribution over the DS-AP estimates with hyper-parameter $\theta = 0.5$. Assuming disease diagnoses among pLoF carriers follow a Bernoulli process, the posterior distribution over the DS-AP is:

$$\text{DS-AP} \sim Beta\left(\theta + \sum_{i=1}^{N} D_i \times G_i, \theta + \sum_{i=1}^{N} (1 - D_i) \times G_i\right),$$

such that the average DS-AP estimate (denoted $\overline{\text{DS-AP}}$) is simply:

$$\overline{\text{DS-AP}} = \frac{\theta + \sum_{i=1}^{N} D_i \times G_i}{2\theta + \sum_{i=1}^{N} G_i}. \quad (3)$$

In practice, $\overline{\text{DS-AP}}$ estimates were obtained in each biobank independently, allowing them to be compared across datasets (ex: Figure 3B).

Phenotype Risk Score Analysis

Most haploinsufficient diseases lack structured diagnostic codes that can be used to identify their presence or absence in EHR data. In such instances, it can be difficult to determine if a subject is in fact expressing disease symptoms. Phenotype Risk Scores[51,52] (PheRS's) were developed to address this issue. These scores measure the extent to which a subject represents an outlier in phenotype space. Their effectiveness relies on a critical assumption: Mendelian disease patients should express constellations of symptoms that are highly atypical when compared to their unaffected counterparts. Although this is sometimes true, it is not the case for all diseases. Moreover, non-Mendelian disease patients can become phenotypic outliers as well, for example, if they develop unusual complications from a common disease or multiple common diseases at once. Therefore, PheRS's are an imperfect method for assessing phenotypic expression, particularly if the goal is to separate Mendelian from non-Mendelian disease subjects based on symptom expression alone. Nevertheless, they are useful for determining if a pLoF carrier is potentially symptomatic.

Let $\vec{S}_i$ denote a binary vector of length *K*, where is *K* is the number of symptoms annotated to the Mendelian disease of interest. Let $S_{i,k} = 1$ indicate that the *k*th symptom was diagnosed at least once in the *i*th subject's EHR data. Finally, let $P(\vec{S}_i|\theta)$ denote the probability of observing the set of symptoms diagnosed in the *i*th patient, where θ represents a set of parameters that define a background symptom expression probability model. The Phenotype Risk Score for the *i*th subject (denoted PheRS_i) is given by:

$$\text{PheRS}_i = -\log\left[P(\vec{S}_i|\theta)\right]. \quad (4)$$

This formula is equivalent to the surprisal, or information content, of the diagnosed symptom set according to the model defined by θ, and it provides a measurement for how unusual or atypical this set of diagnosed symptoms is. For the approach to be effective, we must of course define the symptom expression probability model. Consistent with prior studies[51,52], we assumed that the *k*th symptom occurs independently of the others according to a Bernoulli process defined by the parameter θ_k. Therefore,

$$P\left(S_{i,k}\middle|\theta_k\right) = \theta_k^{S_{i,k}} \times (1-\theta_k)^{1-S_{i,k}}.$$

To estimate the background model parameters, we assumed that the Mendelian disease cases were sufficiently rare in the general population such that their risk of biasing the symptom-specific parameter estimates (denoted $\hat{\theta}_k$) was negligibly low. Therefore, we estimated each symptom expression parameter independently using the maximum likelihood estimator for a Bernoulli process:

$$\hat{\theta}_k = \frac{\sum_{i=1}^{N} S_{i,k}}{N}, \tag{5}$$

where *N* denotes the total number of subjects in the biobank. After estimating this expression model, the PheRS_i score for each subject becomes:

$$\mathrm{PheRS}_i = \sum_{k=1}^{K} -\log\left(\hat{\theta}_k\right) \times S_{i,k} - \log\left(1-\hat{\theta}_k\right) \times \left(1-S_{i,k}\right). \tag{6}$$

In practice, we further adjusted the raw PheRS's for confounding covariates (recruitment age, birth sex, and the first 16 components of the genetic relatedness matrix) using ordinary least squares regression.

After computing the covariate-adjusted disease-specific PheRS's for every subject in both biobanks, we then compared the distribution of these scores between pLoF carriers and their non-carrier controls. To assign statistical significance, we used a one-sided Brunner-Munzel Non-Parametric Hypothesis Test (implemented in `scipy`[95]), which evaluated the null hypothesis that the PheRS's observed in the pLoF carriers were stochastically less than those observed in controls. A fixed effects meta-analysis was performed using the effect size and standard error estimates produced by the Brunner-Munzel Tests performed in each biobank. Finally, for the histograms in Figures 2B and 2C, the median PheRS's in the pLoF carriers were converted to modified Z-scores using the medians and median absolute deviations estimated within non-carrier controls. The complete set of PheRS results for all diseases are given in Supplemental Table 6.

Estimating Symptom-Driven Disease Expression Scores

PheRS's can suggest that a subject is a phenotypic outlier, but these scores do not necessarily provide an accurate assessment of whether a Mendelian disease is being expressed or not. For example, consider autosomal dominant polycystic kidney disease (ADPKD). Clearly, a pLoF carrier who has bilateral renal cysts complicated by chronic kidney disease is expressing the phenotype, but what if a carrier only experiences proteinuria? Proteinuria is certainly a symptom of ADPKD, so this carrier's PheRS score will be greater than 0. But proteinuria is an incredibly common symptom in the general population. Therefore, just because an ADPKD pLoF carrier experiences proteinuria at some point in their life doesn't mean that they are expressing ADPKD.

To overcome this issue, we formulated the following symptom-driven disease expression model. As before, let $P(\vec{S}_i|\theta)$ denote the probability that a set of symptoms $\vec{S}_i$ is being expressed according to some general background distribution. In addition, let $P(\vec{S}_i|\delta)$ denote the probability that this symptom set is instead expressed within an individual affected by a Mendelian disease (where the parameter set δ defines the expression model). Finally, let $E_i = 1$ indicate that the disease of interest is being expressed in the *i*th carrier. Consistent with the diagram in Figure 4A, the probability of disease expression in the *i*th pLoF carrier is given by:

$$P(E_i = 1|\vec{S}_i, \delta, \theta, \pi) = F\left[\log\left(\frac{\pi \times P(\vec{S}_i|\delta)}{(1-\pi) \times P(\vec{S}_i|\theta)}\right)\right], \tag{7}$$

where π is the prior probability of disease expression among all carriers and F is the logistic function.

For this symptom-driven expression model to be effective, both the disease-specific expression model (i.e. $P(\vec{S}_i|\delta)$) and the expression prior probability (i.e. π) must be either known *a priori* or estimated from the data. Estimating π from the data is relatively straightforward, but the disease-specific expression model may be incredibly complex and is largely unknown. Moreover, the independence assumption invoked for PheRS estimation is unlikely to hold for Mendelian diseases, as it is the co-occurrence of multiple unusual symptoms that typically defines a Mendelian disease.

To overcome these issues, we assumed that $P(\vec{S}_i|\delta)$ follows a completely arbitrary distribution over symptoms sets. More specifically, let δ define a multinomial distribution over all possible expressed symptom sets (i.e. all possible sets except the empty set), such that the distribution is defined by a parameter set with cardinality $2^K - 1$, where *K* denotes the total number of symptoms annotated to some disease of interest. Clearly, even for modest values of *K*, the dimensionality of the model becomes unwieldly, so we made the simplifying assumption that the possible set of symptoms is much smaller than $2^K - 1$ (i.e. many of the multinomial distribution parameters are equal to 0). Practically, we assumed that only $M_{\text{Obs}} + 1$ symptom sets were possible, where M_{Obs} denotes the number of unique symptoms sets observed across all biobank participants. The $+1$ term allows for the addition of a generic symptom set that accounts for any non-empty set that was not observed in the biobank, which enables the model to be trained in one biobank yet still be applicable to another. Note, the total complement of observed symptom sets in either biobank was very sparse compared to the cardinality of all possible sets, typically numbering in the 10s or 100s.

With this assumption in place, the Mendelian disease symptom expression model was defined as:

$$P(\vec{S}_i = \vec{s}|\delta) = \sum_{m=1}^{M_{\text{Obs}}+1} \delta_m \times \mathbf{1}(\vec{s} \equiv \mathcal{S}_m)$$

where $\mathbf{1}(\vec{s} \equiv \mathcal{S}_m)$ is an indicator function that returns 1 if and only if the observed symptom set $\vec{s}$ is identical to the symptom set whose expression probability is defined

by δ_m (denoted $\mathcal{S}_m$ in the equation above). The symptom expression model defined in Eqn. 7 can then be used to specify the following likelihood for the observed symptom data:

$$P(\boldsymbol{S}|\mathrm{E},\theta,\delta,\pi) = \prod_{i=1}^{V} \sum_{E_i \in 0,1} \left[\pi \times P\left(\vec{S}_i \middle| \delta\right)\right]^{E_i} \times \left[(1-\pi) \times P\left(\vec{S}_i \middle| \theta\right)\right]^{1-E_i}, \quad (8)$$

where $\boldsymbol{S}$ denotes the matrix of diagnosed symptom sets across the V pLoF carriers. By estimating the model parameters (denoted $\hat{\theta}$, $\hat{\delta}$, and $\hat{\pi}$) through likelihood maximization, the posterior probability over disease expression (defined in Eqn. 7) can be estimated.

For many diseases, the total number of pLoF carriers was small, making it difficult to simultaneously estimate both the disease-specific and background expression models simultaneously. Therefore, the background expression models used for PheRS estimation were used to define $P\left(\vec{S}_i \middle| \hat{\theta}\right)$ (see Eqn. 5 for estimation procedure). As a result, only the disease expression prior π and the parameters defining the disease-specific expression model (denoted δ) needed to be estimated. To regularize these estimates in the face of sparse data, we assumed that these parameters were drawn from uniform Beta and Dirichlet distributions respectively. The model specified in Eqn. 8 was then fit by maximizing a lower-bound on the marginal likelihood using variational Bayesian inference[96]. The posterior distributions over the individual expression probabilities, denoted $P\left(E_i = 1 \middle| \vec{S}_i, \hat{\delta}, \hat{\theta}, \hat{\pi}\right)$ for the *i*th carrier, were generated automatically during inference. In practice, we fit the disease-specific expression models (i.e. $P\left(\vec{S}_i \middle| \delta\right)$) only within the UKBB (given its larger sample size), as such models remain at risk for overfitting even though inference is technically unsupervised. After fitting in the UKBB, the parameters estimated in this biobank were used to predict expression probabilities in AoU. Note, this procedure eliminated 15 diseases from our analysis, as these disorders did not share diagnosed symptoms across bioanks.

The previously described model generated expression probabilities for every eligible pLoF carrier. However, it did not ensure that these probabilities were calibrated to disease expression risk. In other words, if $P\left(E_i = 1 \middle| \vec{S}_i, \hat{\delta}, \hat{\theta}, \hat{\pi}\right) = 0.5$, it was hard to determine what this meant from a disease expression perspective. To place these probabilities on a coherent scale, we turned to the set of diseases that have both diagnostic and symptom data available in both biobanks. If the symptom-driven model produced coherent expression probabilities, then these scores should be predictive of which pLoF carriers harbor Mendelian disease diagnoses.

To test this hypothesis, we used the symptom-driven expression probabilities to predict Mendelian disease diagnoses among pLoF carriers, computing the precision and recall scores across all possible symptom-driven expression probability thresholds. The results of this analysis are displayed in Figure 4B. Clearly, the symptom-driven expression probabilities performed better than random in both datasets. In addition, the performance of the expression probabilities was relatively consistent across the two biobanks, although performance was better in the UKBB. To a select a symptom expression threshold for downstream analyses, we identified the expression probability

score that maximized the F_1 measure (harmonic mean of the precision and recall) for the predictions shown in Figure 4B. This threshold was nearly identical across the two biobanks (0.975 and 0.972 in the UKBB and AoU respectively). Importantly, after defining disease expression according to this threshold, the symptom-driven expression predictions were statistically indistinguishable from the disease diagnoses themselves (McNemar's Test for paired nominal data, implemented in `statsmodels`[92]).

In all downstream analyses, we treated disease expression as a binary outcome. More specifically, we considered a pLoF to be expressed if the carrier harbored a Mendelian disease diagnosis (i.e. $D_i = 1$, assuming diagnostic data was available) or if their symptom-driven expression probability (denoted $P\left(E_i = 1 \middle| \vec{S}_i, \hat{\delta}, \hat{\theta}, \hat{\pi}\right)$) exceeded the F_1 thresholds described above. Treating disease expression as binary enabled us to estimate DS-AP's using the same methods that were used for simple diagnoses (see Eqn. 3 for details). It also greatly simplified the machine learning analyses, as models for binary prediction are well-established. Additional work is needed to effectively incorporate the uncertainty that is inherent to measuring Mendelian disease expression into the types of analyses performed in this study.

Strategy for Removing Samples with Incomplete Clinical Data Coverage

Biobanks are rife with incomplete clinical data, as the EHR is an imperfect representation of a patient's phenotype. Moreover, biobank subjects are enrolled into these studies well into adulthood, and there is no guarantee that the records captured by the study represent their complete clinical history. Supplemental Figure 1 illustrates the extensive variability in data coverage that was observed within the UKBB and AoU. Given the limited data available for many of these subjects, phenotypic imputation was unrealistic. Therefore, we devised a method to flag and remove subjects from our analysis that had unacceptably low clinical data coverage.

Let $\vec{A}$ denote a binary vector of asymptomatic indicators, where $A_i = 1$ indicates that the *i*th pLoF carrier had no evidence for disease expression based on disease-specific diagnostic code(s) and/or documented symptoms (i.e. had no disease-relevant diagnoses). Moreover, let $\boldsymbol{W}$ denote a matrix of clinical data coverage statistics. For this analysis, we used the following four statistics to define data coverage: Age at First Clinical Encounter, Age at Recruitment, Total Number of Documented Clinical Encounters, and Age at Last Clinical Encounter. These four statistics provided a basic summary of a patient's interaction with the healthcare system, at least according to the information in the biobanks. Finally, let $\vec{b}$ denote a vector of coverage statistic effect size parameters. We modeled the probability of phenotypic *non*-expression (i.e. asymptomatic status) conditional on clinical data coverage using the following log-linear model:

$$\boldsymbol{P}\left(\vec{\mathbf{A}} \middle| \mu, \boldsymbol{W}, \vec{b}\right) = F\left(\mu + \boldsymbol{W} \times \vec{b}\right)$$

where F denotes the logistic function and μ is the intercept term. Because different diseases will have different coverage requirements (depending on their onset, pathophysiology, etc), we fit three versions of this model in both biobanks by grouping diseases together based on their typical onset.

More specifically, each version of the model was repeatedly fit in both biobanks using leave-one-out 5-fold cross validation (model fitting was performed using the `LogisticRegression` function available in `sklearn`[85] using the default hyperparameters). For each iteration, 80% of the onset-grouped pLoF carriers were used to estimate the parameters for the disease non-expression model. The remaining 20% were used for validation. Model performance was assessed using the area under the receiver operating characteristic curve. All models performed better than random, but there was considerable variability in their performance across typical onset and biobanks. To flag pLoF carriers with insufficient clinical data, we identified the 5% false positive rate threshold in each validation subset. If a subject in a validation subset had a non-expression probability that exceeded this threshold, they were flagged for removal from downstream analyses. As discussed in the main text, this *ad hoc* procedure removed a substantial fraction of pLoF carriers from both datasets (17% and 35% in the UKBB and AoU respectively). Moreover, the average pLoF penetrance estimates increased systematically after filtering. Nevertheless, the absolute increase in phenotypic expression that occurred because of this filtering was low.

Predicting pLoF Phenotypic Expression using Variant-Specific Features

Let $\vec{V}_i$ denote a vector of genomic features that characterize the pLoF variant carried by the *i*th subject. Examples of such features include its relative position within the amino acid sequence[97] or its deleteriousness based on computational prediction tools[65,66]. The goal of this analysis is to predict the probability of phenotypic expression directly from the set of features that are unique to the pLoF carried by the *i*th subject:

$$P\left(E_i = 1 \middle| \vec{V}_i, \theta\right) = \mathcal{F}\left(\vec{V}_i; \theta\right)$$

where $\mathcal{F}$ is some function that maps the vector $\vec{V}_i$ onto disease expression probability space via a parameter set θ. Practically, different models can accomplish this goal. For this study, we constructed $\mathcal{F}$ using the random forest algorithm[62] implemented in the `sklearn`[85], which builds predictive models via an ensemble of individual decision trees. Model fitting was performed by minimizing the logarithmic loss function of the prediction model when applied to a cohort of training carriers (training algorithm hyperparameters: `min_samples_leaf`=5, `min_samples_split`=10, `n_estimators`=500). Note, we also considered simpler methods for constructing $\mathcal{F}$ (i.e. penalized logistic regression) but found that they performed systematically worse than this ensemble learning approach (see Supplemental Figure 2), likely due to the latter's ability to capture non-linear effects.

Any predictive model built using machine learning is at risk for overfitting, particularly models with many free parameters like random forests. To minimize the risk for overfitting, machine learning model inference was performed exclusively in the UKBB, after which the models were independently evaluated in AoU. In addition, only completely asymptomatic pLoF carriers were included as negative cases in the UKBB training dataset to avoid confounding the model with carriers who were weakly symptomatic but did not reach the severity threshold required to designate them as phenotypically expressed. Given that the two biobanks were recruited from the

populations of two different countries, the risk that the validation dataset was contaminated with subjects from the training dataset was very low.

Finally, different types of pLoF variants have distinct features that likely impact their risk for expression[55]. Therefore, distinct phenotypic expression models were constructed for each of the three variant types analyzed in this study (stop gain, frameshift, and splice change). The remainder of this section describes the variant-specific features that were used to build phenotype expression prediction models for each class of pLoFs. These features rely heavily on the ideas from prior studies[55,97,98].

Variant Class Agnostic Features:

- CADD Score[65]: The Combined Annotation-Dependent Depletion (CADD) score predicts the deleteriousness of individual variants using a single numerical score derived from a wide range of variant-specific features, including but not limited to evolutionary conservation, DNA sequence motifs, and predicted impact on biochemical activity. Uniquely, CADD does not build these scores by training on a set of variants known to cause human disease. Instead, the scores are inferred by fitting a machine learning model to a set of evolutionarily neutral variants (proxy-negative cases) and a set of simulated mutations, which may or may not be deleterious (proxy-positive cases). This makes CADD well-suited for the analysis conducted in this study, as the score should not be polluted with information from prior ClinVar annotations.
- LOFTEE Confidence Flag[54]: The LOFTEE plug-in for VEP[89] not only identifies putative loss-of-function variants but also assigns them a confidence flag (low or high) based on several variant-specific features (e.g. distance from end of transcript, ancestral alleles, etc.; see https://github.com/konradjk/LOFTEE for details)
- Transcript Type[57]: All variants were assigned to one of three transcript types (MANE Select, MANE Plus Clinical, Other) based on the most clinically relevant transcript that was predicted to be impacted.

Stop-Gain Variant Features:

- Predicted Non-sense Mediate Decay (NMD) Escape[60]: It is well known that some stop-gain variants escape non-sense mediated decay, enabling the expression of a potentially functional but truncated transcript. To predict NMD escape, we used the decision tree developed in Lindeboom et al[59]. Note, we did not encode predicted NMD Escape using a binary annotation (Present, Absent) but instead included the reason for the predicted escape into the model (No NMD Escape Present, Last Exon, First Exon ≤ 150nt from Start, Large Exon, ≤50nt from Last Exon-Exon Junction).
- Predicted Fraction of Amino Acids Lost[97]: If a variant escapes NMD, this feature computes the fraction of the amino acid sequence predicted to be lost. For stop-gain variants, this is simply it's relative distance from the N-terminus (according to the MANE Select[57] transcript).
- Possible Methionine Rescue (Translation Re-initiation)[55]: If a stop-gain variant occurs early enough in the amino acid sequence, then translation can potentially

be rescued by another methionine residue that occurs just downstream. The exact criteria needed to be met for this to occur are unknown and may be variable across proteins. For this analysis, a stop gain variant had to meet the following criteria to flag for possible methionine rescue: 1) located in the first exon and 2) have a downstream methionine for alternate translation initiation that truncated <10% of the total protein length.

Frameshift Variant Features

- Last Coding Exon[55]: This is a simple binary feature that indicates if the frameshift variant occurred in the last exon.
- Predicted Fraction of Amino Acids Impacted[97]: This feature computes the relative fraction of amino acids predicted to be lost by an expressed frameshift. Like stop-gain variants, this feature captures the relative distance from N-terminus of the protein for the last normal amino acid.
- Possible Methionine Rescue[55]: This feature is computed in the same fashion for frameshift and stop gain variants.
- Note, NMD escape is certainly possible for frameshift variants with the added complexity that the escape is occurring on a frameshifted sequence. It's possible that additional features based on NMD escape would improve frameshift penetrance prediction, but additional work is needed to determine when these rules may apply.

Splice Change Variant Features

- SpliceAI Score[66]: SpliceAI is a deep learning model that predicts changes in the splicing probabilities at different sites induced by a genetic variant relative to the splicing probabilities for the reference sequence (assuming some specific transcript model). For the current analysis, we re-computed SpliceAI scores using the Ensembl transcripts for each gene (Release 109), allowing for a maximum 500bp between the variant and impacted site. For expression prediction, the maximum SpliceAI score (maximum difference in splicing probability between the reference and mutated transcript) was included as feature. Note, several additional features were derived for splice variants based on the SpliceAI output. These are outlined in detail below.
- Predicted Fraction of Amino Acids Lost[97]: Like the other variant classes, we computed the fraction of amino acids that would theoretically be lost based on the splice site location assuming that it was expressed rather than undergoing NMD. Determining the exact location of the last normal amino acid for splice variants can be challenging. Therefore, we set the last normal amino acid to be the residue just proximal to the impacted splice site in the transcript model. This could clearly be improved (ex: by considering in-frame splice rescue events, exon skipping, etc), but this will be the focus of future work.
- Splice Mutation Type: Pathogenic splice mutations can impact transcript structure in complex ways, sometimes inducing multiple changes simultaneously. For the sake of simplicity, we used the SpliceAI output to assign each mutation to one of five classes based on the highest SpliceAI score observed for the variant: Donor

Gain, Donor Loss, Acceptor Gain, Acceptor Loss and Indeterminate (i.e. maximum SpliceAI score = 0.0 or NaN).

- Outside Coding Region: Some splice sites occur in exons that lie outside the coding region. Although they could result in loss-of-function, many of these may be tolerated. Therefore, we included a binary feature that flagged splice variants predicted to impact only non-coding exons.
- Last Coding Exon[55]: This feature indicates whether a splice mutation is predicted to impact the last coding exon. Like the other variant classes, such mutations should be more likely to be tolerated.
- Persistent Original Splice Site Score: Sometimes, SpliceAI predicts that the original splice site remains intact with some non-zero probability, which may be indicative of leaky wild type expression. Therefore, we computed the difference between the SpliceAI score for the original and derived sites. Generally, this is simply equivalent to the global SpliceAI score, but other times, a variant increases the splicing probability for the wildtype splice site along with the derived site. This feature accounts for this phenomenon.
- In-frame Exon Rescue[55]: If the exon impacted by a splice change has a nucleotide length that is a multiple of 3, then it can theoretically be skipped without disrupting the reading frame. This phenomenon was accounted for in the model using a binary feature (Present, Absent).
- Possible Methionine Rescue[55]: For splice variants, this is a less likely rescue mechanism. Nevertheless, given that a variant impacted the first exon, we allowed for possible methionine rescue assuming that there was a methionine residue in the second exon that truncated less than 10% of the amino acid sequence.
- In-frame Intron Retention[55]: If the intron to be spliced out has a nucleotide length that is a multiple of 3, then it can potentially be retained without impacting the transcript reading frame. This phenomenon is accounted for in the model using a binary feature (Present, Absent).
- Cryptic Rescue Score[55,98]: Many times, when SpliceAI predicts a primary splice site change, a secondary change is predicted to occur simultaneously that could negate the impact of the primary. More specifically, if a genetic variant is predicted cause a donor (acceptor) loss event in a transcript, there can be a complementary donor (acceptor) gain event just upstream/downstream of the predicted loss site but with a lower SpliceAI score. If this event remains in-frame with the original transcript, then the impact of the mutation may be minimal, as this complementary site could compensate for the loss. Alternatively, many donor (acceptor) gain events occur in-frame with the original donor (acceptor) site. So as long the downstream acceptor (upstream donor) site remains intact, then the impact of the variant may be insignificant. This Cryptic Rescue Score summarizes both possible rescue events using the output from SpliceAI. For primary splice site loss events (donor or acceptor), the Cryptic Rescue Score is simply the SpliceAI score for the in-frame gain event (assigned 0.0 if no in-frame gain is predicted). For primary in-frame gain events (donor or acceptor), the Cryptic Splice Score is harder to define. For this analysis, we used the corresponding splice site loss score (ex: a loss score of 1.0 should indicate that

this predicted in-frame gain is preferentially being used) but acknowledge that this very much imperfectly captures the phenomenon. Clearly, more work is needed to effectively capture the complexity of splice mutation rescue events.

Additional Statistical Methods

Unless otherwise noted, the statistical analyses described in the main text and/or figure legends were performed using the implementations (sometimes with slight modification) available in the following Python packages: `scipy`[95], `sklearn`[85], `statsmodels`[92], and `pandas`[99]. Bootstrapped hypothesis testing was performed by generating empirical distributions for the target parameter estimates using re-sampling with replacement (10,000 re-samples for all tests). Randomization tests were performed similarly. All meta-analyses were performed using a fixed-effects model based on the standard normal distribution[93].

Data Availability

The genomic and electronic health data used for this analysis are publicly available but have strict data use agreements. The process for obtaining access to these biobanks can be found on their respective websites: https://www.researchallofus.org/register/ (All of Us Research Program) and https://www.ukbiobank.ac.uk/enable-your-research/register (UK Biobank). Haploinsufficient disease annotations are provided in Supplemental Table 1. The custom HPO-to-OMOP concept alignments generated in this study are provided as Supplemental Table 10. All other databases used in this analysis are freely available in the public domain.

Software Availability

All software packages used to conduct this study are open source and available in the public domain. Any custom software packages or scripts created for the purpose of this study will be made available prior to formal publication.

Acknowledgements

The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276. In addition, the All of Us Research Program would not be possible without the partnership of its participants.

This research was conducted using the UK Biobank Resource under Application Number 99922, which uses data provided by patients and collected by the NHS as part of their care and support. We are extremely grateful to the participants of the UK Biobank, without whom this research would not have been possible.

This work was supported by grants from the National, Heart, Lung and Blood Institute (K38HL164956) and the George Banks and Sarah Ellen Huntington Memorial Fund.

References

1. Srivastava S, Love-Nichols JA, Dies KA, Ledbetter DH, Martin CL, Chung WK, Firth HV, Frazier T, Hansen RL, Prock L, Brunner H, Hoang N, Scherer SW, Sahin M, Miller DT. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. Genet Med. 2019;21(11):2413–2421. PMCID: PMC6831729

2. Scocchia A, Wigby KM, Masser-Frye D, Del Campo M, Galarreta CI, Thorpe E, McEachern J, Robinson K, Gross A, Ajay SS, Rajan V, Perry DL, Belmont JW, Bentley DR, Jones MC, Taft RJ. Clinical whole genome sequencing as a first-tier test at a resource-limited dysmorphology clinic in Mexico. npj Genomic Med. Nature Publishing Group; 2019 Feb 14;4(1):1–12.

3. Manickam K, McClain MR, Demmer LA, Biswas S, Kearney HM, Malinowski J, Massingham LJ, Miller D, Yu TW, Hisama FM. Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: an evidence-based clinical guideline of the American College of Medical Genetics and Genomics (ACMG). Genetics in Medicine. 2021 Nov 1;23(11):2029–2037.

4. Yaron Y, Ofen Glassner V, Mory A, Zunz Henig N, Kurolap A, Bar Shira A, Brabbing Goldstein D, Marom D, Ben Sira L, Baris Feldman H, Malinger G, Krajden Haratz K, Reches A. Exome sequencing as first-tier test for fetuses with severe central nervous system structural anomalies. Ultrasound in Obstetrics & Gynecology. 2022;60(1):59–67.

5. van der Sanden BPGH, Schobers G, Corominas Galbany J, Koolen DA, Sinnema M, van Reeuwijk J, Stumpel CTRM, Kleefstra T, de Vries BBA, Ruiterkamp-Versteeg M, Leijsten N, Kwint M, Derks R, Swinkels H, den Ouden A, Pfundt R, Rinne T, de Leeuw N, Stegmann AP, Stevens SJ, van den Wijngaard A, Brunner HG, Yntema HG, Gilissen C, Nelen MR, Vissers LELM. The performance of genome sequencing as a first-tier test for neurodevelopmental disorders. Eur J Hum Genet. 2023 Jan;31(1):81–88. PMCID: PMC9822884

6. Cirillo L, Becherucci F. The evolving role of first-tier exome sequencing in medical diagnostics. Nephrol Dial Transplant. 2024 Mar 27;39(4):560–563. PMID: 37858299

7. Bodian DL, Klein E, Iyer RK, Wong WSW, Kothiyal P, Stauffer D, Huddleston KC, Gaither AD, Remsburg I, Khromykh A, Baker RL, Maxwell GL, Vockley JG, Niederhuber JE, Solomon BD. Utility of whole-genome sequencing for detection of newborn screening disorders in a population cohort of 1,696 neonates. Genet Med. 2016 Mar;18(3):221–230. PMID: 26334177

8. Bailey DB, Gehtland LM, Lewis MA, Peay H, Raspa M, Shone SM, Taylor JL, Wheeler AC, Cotten M, King NMP, Powell CM, Biesecker B, Bishop CE, Boyea BL, Duparc M, Harper BA, Kemper AR, Lee SN, Moultrie R, Okoniewski KC, Paquin RS, Pettit D, Porter KA, Zimmerman SJ. Early Check: translational science at the intersection of public health and newborn screening. BMC Pediatr. 2019 Jul 17;19(1):238. PMCID: PMC6636013

9. Foss KS, O'Daniel JM, Berg JS, Powell SN, Cadigan RJ, Kuczynski KJ, Milko LV, Saylor KW, Roberts M, Weck K, Henderson GE. The Rise of Population Genomic Screening: Characteristics of Current Programs and the Need for Evidence Regarding Optimal Implementation. Journal of Personalized Medicine. Multidisciplinary Digital Publishing Institute; 2022 May;12(5):692.

10. Buchanan AH, Lester Kirchner H, Schwartz MLB, Kelly MA, Schmidlen T, Jones LK, Hallquist MLG, Rocha H, Betts M, Schwiter R, Butry L, Lazzeri AL, Frisbie LR, Rahm AK, Hao J, Willard HF, Martin CL, Ledbetter DH, Williams MS, Sturm AC. Clinical outcomes of a genomic screening program for actionable genetic conditions. Genet Med. 2020 Nov;22(11):1874–1882. PMCID: PMC7605431

11. Casalino S, Frangione E, Chung M, MacDonald G, Chowdhary S, Mighton C, Faghfoury H, Bombard Y, Strug L, Pugh TJ, Simpson J, Arnoldo S, Aujla N, Bearss E, Binnie A, Borgundvaag B, Chertkow H, Clausen M, Dagher M, Devine L, Di Iorio D, Friedman SM, Fung CYJ, Gingras AC, Goneau LW, Kaushik D, Khan Z, Lapadula E, Lu T, Mazzulli T, McGeer A, McLeod SL, Morgan G, Richardson D, Singh H, Stern S, Taher A, Wong I, Zarei N, Greenfeld E, Hao L, Lebo M, Lane W, Noor A, Taher J, Lerner-Ellis J. Genome screening, reporting, and genetic counseling for healthy populations. Hum Genet. 2023 Feb;142(2):181–192. PMCID: PMC9638226

12. Chen T, Fan C, Huang Y, Feng J, Zhang Y, Miao J, Wang X, Li Y, Huang C, Jin W, Tang C, Feng L, Yin Y, Zhu B, Sun M, Liu X, Xiang J, Tan M, Jia L, Chen L, Huang H, Peng H, Sun X, Gu X, Peng Z, Zhu B, Zou H, Han L. Genomic Sequencing as a First-Tier Screening Test and Outcomes of Newborn Screening. JAMA Netw Open. 2023 Sep 5;6(9):e2331162. PMCID: PMC10474521

13. Green RC, Shah N, Genetti CA, Yu T, Zettler B, Uveges MK, Ceyhan-Birsoy O, Lebo MS, Pereira S, Agrawal PB, Parad RB, McGuire AL, Christensen KD, Schwartz TS, Rehm HL, Holm IA, Beggs AH, BabySeq Project Team. Actionability of unanticipated monogenic disease risks in newborn genomic screening: Findings from the BabySeq Project. Am J Hum Genet. 2023 Jul 6;110(7):1034–1045. PMCID: PMC10357495

14. Stone K. The Generation Study — Knowledge Hub [Internet]. GeNotes. [cited 2024 Aug 20]. Available from: https://www.genomicseducation.hee.nhs.uk/genotes/knowledge-hub/the-generation-study/

15. Stark Z, Scott RH. Genomic newborn screening for rare diseases. Nat Rev Genet. 2023 Nov;24(11):755–766. PMID: 37386126

16. Chung WK, Kanne SM, Hu Z. An Opportunity to Fill a Gap for Newborn Screening of Neurodevelopmental Disorders. Int J Neonatal Screen. 2024 Apr 16;10(2):33. PMCID: PMC11036277

17. Baple EL, Scott RH, Banka S, Buchanan J, Fish L, Wynn S, Wilkinson D, Ellard S, MacArthur DG, Stark Z. Exploring the benefits, harms and costs of genomic newborn screening for rare diseases. Nat Med. 2024 Jul;30(7):1823–1825. PMID: 38898121

18. Woerner AC, Gallagher RC, Vockley J, Adhikari AN. The Use of Whole Genome and Exome Sequencing for Newborn Screening: Challenges and Opportunities for Population Health. Front Pediatr. 2021;9:663752. PMCID: PMC8326411

19. Horton R, Wright CF, Firth HV, Turnbull C, Lachmann R, Houlston RS, Lucassen A. Challenges of using whole genome sequencing in population newborn screening. BMJ. British Medical Journal Publishing Group; 2024 Mar 5;384:e077060. PMID: 38443063

20. Turnbull C, Firth HV, Wilkie AOM, Newman W, Raymond FL, Tomlinson I, Lachmann R, Wright CF, Wordsworth S, George A, McCartney M, Lucassen A. Population screening requires robust evidence—genomics is no exception. The Lancet. Elsevier; 2024 Feb 10;403(10426):583–586. PMID: 38070525

21. Zhao S, Agafonov O, Azab A, Stokowy T, Hovig E. Accuracy and efficiency of germline variant calling pipelines for human genome data. Sci Rep. Nature Publishing Group; 2020 Nov 19;10(1):20222.

22. Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, Johanson E, Boja E, Maier EJ, Serang O, Jáspez D, Lorenzo-Salazar JM, Muñoz-Barrera A, Rubio-Rodríguez LA, Flores C, Kyriakidis K, Malousi A, Shafin K, Pesout T, Jain M, Paten B, Chang PC, Kolesnikov A, Nattestad M, Baid G, Goel S, Yang H, Carroll A, Eveleigh R, Bourgey M, Bourque G, Li G, Ma C, Tang L, Du Y, Zhang S, Morata J, Tonda R, Parra G, Trotta JR, Brueffer C, Demirkaya-Budak S, Kabakci-Zorlu D, Turgut D, Kalay Ö, Budak G, Narcı K, Arslan E, Brown R, Johnson IJ, Dolgoborodov A, Semenyuk V, Jain A, Tetikol HS, Jain V, Ruehle M, Lajoie B, Roddey C, Catreux S, Mehio R, Ahsan MU, Liu Q, Wang K, Ebrahim Sahraeian SM, Fang LT, Mohiyuddin M, Hung C, Jain C, Feng H, Li Z, Chen L, Sedlazeck FJ, Zook JM. PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. Cell Genom. 2022 Apr 27;2(5):100129. PMCID: PMC9205427

23. Fahed AC, Wang M, Homburger JR, Patel AP, Bick AG, Neben CL, Lai C, Brockman D, Philippakis A, Ellinor PT, Cassa CA, Lebo M, Ng K, Lander ES, Zhou AY, Kathiresan S, Khera AV. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. Nat Commun. 2020 Aug 20;11(1):3635.

24. Blair DR, Hoffmann TJ, Shieh JT. Common genetic variation associated with Mendelian disease severity revealed through cryptic phenotype analysis [Internet]. 2021 Aug p. 2021.08.26.21262300. Available from: https://www.medrxiv.org/content/10.1101/2021.08.26.21262300v1

25. Kingdom R, Beaumont RN, Wood AR, Weedon MN, Wright CF. Genetic modifiers of rare variants in monogenic developmental disorder loci. Nat Genet. 2024 May;56(5):861–868. PMCID: PMC11096126

26. Tukker AM, Royal CD, Bowman AB, McAllister KA. The Impact of Environmental Factors on Monogenic Mendelian Diseases. Toxicol Sci. 2021 Mar 2;181(1):3–12. PMCID: PMC8599782

27. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL. Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015 May;17(5):405–424. PMCID: PMC4544753

28. Kingdom R, Wright CF. Incomplete Penetrance and Variable Expressivity: From Clinical Studies to Population Cohorts. Front Genet. 2022;13:920390. PMCID: PMC9380816

29. Risch NJ, Bressman SB, Senthil G, Ozelius LJ. Intragenic Cis and Trans Modification of Genetic Susceptibility in DYT1 Torsion Dystonia. Am J Hum Genet. 2007 Jun;80(6):1188–1193. PMCID: PMC1867106

30. Chen S, Parmigiani G. Meta-Analysis of BRCA1 and BRCA2 Penetrance. J Clin Oncol. 2007 Apr 10;25(11):1329–1333. PMCID: PMC2267287

31. Hoffmann TJ, Sakoda LC, Shen L, Jorgenson E, Habel LA, Liu J, Kvale MN, Asgari MM, Banda Y, Corley D, Kushi LH, Quesenberry CP, Schaefer C, Van Den Eeden SK, Risch N, Witte JS. Imputation of the Rare HOXB13 G84E Mutation and Cancer Risk in a Large Population-Based Cohort. PLoS Genet. 2015 Jan 28;11(1):e1004930. PMCID: PMC4309593

32. Menozzi E, Schapira AHV. Exploring the Genotype–Phenotype Correlation in GBA-Parkinson Disease: Clinical Aspects, Biomarkers, and Potential Modifiers. Front Neurol [Internet]. Frontiers; 2021 Jun 24 [cited 2024 Aug 28];12. Available from: https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2021.694764/full

33. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008 Sep;84(3):362–369. PMCID: PMC3763939

34. Kvale MN, Hesselson S, Hoffmann TJ, Cao Y, Chan D, Connell S, Croen LA, Dispensa BP, Eshragh J, Finn A, Gollub J, Iribarren C, Jorgenson E, Kushi LH, Lao R, Lu Y, Ludwig D, Mathauda GK, McGuire WB, Mei G, Miles S, Mittman M, Patil M, Quesenberry CP Jr, Ranatunga D, Rowell S, Sadler M, Sakoda LC, Shapero M, Shen L, Shenoy T, Smethurst D, Somkin CP, Van Den Eeden SK, Walter L, Wan E, Webster T, Whitmer RA, Wong S, Zau C, Zhan Y, Schaefer C, Kwok PY, Risch N. Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. Genetics. 2015 Aug 1;200(4):1051–1060.

35. Carey DJ, Fetterolf SN, Davis FD, Faucett WA, Kirchner HL, Mirshahi U, Murray MF, Smelser DT, Gerhard GS, Ledbetter DH. The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet Med. 2016;18(9):906–913. PMCID: PMC4981567

36. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J. The UK Biobank resource with deep phenotyping and genomic data. Nature. Nature Publishing Group; 2018 Oct;562(7726):203–209.

37. Belbin GM, Cullina S, Wenric S, Soper ER, Glicksberg BS, Torre D, Moscati A, Wojcik GL, Shemirani R, Beckmann ND, Cohain A, Sorokin EP, Park DS, Ambite JL, Ellis S, Auton A, Bottinger EP, Cho JH, Loos RJF, Abul-Husn NS, Zaitlen NA, Gignoux CR, Kenny EE. Toward a fine-scale population health monitoring system. Cell. Elsevier; 2021 Apr 15;184(8):2068-2083.e11. PMID: 33861964

38. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, Reeve MP, Laivuori H, Aavikko M, Kaunisto MA, Loukola A, Lahtela E, Mattsson H, Laiho P, Della Briotta Parolo P, Lehisto AA, Kanai M, Mars N, Rämö J, Kiiskinen T, Heyne HO, Veerapen K, Rüeger S, Lemmelä S, Zhou W, Ruotsalainen S, Pärn K, Hiekkalinna T, Koskelainen S, Paajanen T, Llorens V, Gracia-Tabuenca J, Siirtola H, Reis K, Elnahas AG, Sun B, Foley CN, Aalto-Setälä K, Alasoo K, Arvas M, Auro K, Biswas S, Bizaki-Vallaskangas A, Carpen O, Chen CY, Dada OA, Ding Z, Ehm MG, Eklund K, Färkkilä M, Finucane H, Ganna A, Ghazal A, Graham RR, Green EM, Hakanen A, Hautalahti M, Hedman ÅK, Hiltunen M, Hinttala R, Hovatta I, Hu X, Huertas-Vazquez A, Huilaja L, Hunkapiller J, Jacob H, Jensen JN, Joensuu H, John S, Julkunen V, Jung M, Junttila J, Kaarniranta K, Kähönen M, Kajanne R, Kallio L, Kälviäinen R, Kaprio J, Kerimov N, Kettunen J, Kilpeläinen E, Kilpi T, Klinger K, Kosma VM, Kuopio T, Kurra V, Laisk T, Laukkanen J, Lawless N, Liu A, Longerich S, Mägi R, Mäkelä J, Mäkitie A, Malarstig A, Mannermaa A, Maranville J, Matakidou A, Meretoja T, Mozaffari SV, Niemi MEK, Niemi M, Niiranen T, O´Donnell

CJ, Obeidat M, Okafo G, Ollila HM, Palomäki A, Palotie T, Partanen J, Paul DS, Pelkonen M, Pendergrass RK, Petrovski S, Pitkäranta A, Platt A, Pulford D, Punkka E, Pussinen P, Raghavan N, Rahimov F, Rajpal D, Renaud NA, Riley-Gillis B, Rodosthenous R, Saarentaus E, Salminen A, Salminen E, Salomaa V, Schleutker J, Serpi R, Shen H yi, Siegel R, Silander K, Siltanen S, Soini S, Soininen H, Sul JH, Tachmazidou I, Tasanen K, Tienari P, Toppila-Salmi S, Tukiainen T, Tuomi T, Turunen JA, Ulirsch JC, Vaura F, Virolainen P, Waring J, Waterworth D, Yang R, Nelis M, Reigo A, Metspalu A, Milani L, Esko T, Fox C, Havulinna AS, Perola M, Ripatti S, Jalanko A, Laitinen T, Mäkelä TP, Plenge R, McCarthy M, Runz H, Daly MJ, Palotie A. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature. 2023;613(7944):508–518. PMCID: PMC9849126

39. Johnson R, Ding Y, Bhattacharya A, Knyazev S, Chiu A, Lajonchere C, Geschwind DH, Pasaniuc B. The UCLA ATLAS Community Health Initiative: Promoting precision health research in a diverse biobank. Cell Genom. 2023 Jan 11;3(1):100243. PMCID: PMC9903668

40. All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. Nature. 2024 Mar;627(8003):340–346. PMCID: PMC10937371

41. Wright CF, Sharp LN, Jackson L, Murray A, Ware JS, MacArthur DG, Rehm HL, Patel KA, Weedon MN. Guidance for estimating penetrance of monogenic disease-causing variants in population cohorts. Nat Genet. 2024 Jul 29; PMID: 39075210

42. Wright CF, West B, Tuke M, Jones SE, Patel K, Laver TW, Beaumont RN, Tyrrell J, Wood AR, Frayling TM, Hattersley AT, Weedon MN. Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting. Am J Hum Genet. 2019 Feb 7;104(2):275–286. PMCID: PMC6369448

43. Forrest IS, Chaudhary K, Vy HMT, Petrazzini BO, Bafna S, Jordan DM, Rocheleau G, Loos RJF, Nadkarni GN, Cho JH, Do R. Population-Based Penetrance of Deleterious Clinical Variants. JAMA. 2022 Jan 25;327(4):350–359. PMCID: PMC8790667

44. Mirshahi UL, Colclough K, Wright CF, Wood AR, Beaumont RN, Tyrrell J, Laver TW, Stahl R, Golden A, Goehringer JM, Frayling TF, Hattersley AT, Carey DJ, Weedon MN, Patel KA. Reduced penetrance of MODY-associated HNF1A/HNF4A variants but not GCK variants in clinically unselected cohorts. Am J Hum Genet. 2022 Nov 3;109(11):2018–2028. PMCID: PMC9674944

45. Bastarache L, Peterson JF. Penetrance of Deleterious Clinical Variants. JAMA. 2022 May 17;327(19):1926–1927. PMCID: PMC9350877

46. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipatla Z, Malheiro A, McDaniel K, Ovetsky M, Riley G, Zhou G, Holmes JB, Kattman BL, Maglott DR. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan 4;46(D1):D1062–D1067. PMCID: PMC5753237

47. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, Plon SE, Ramos EM, Sherry ST, Watson MS, ClinGen. ClinGen--the Clinical Genome Resource. N Engl J Med. 2015 Jun 4;372(23):2235–2242. PMCID: PMC4474187

48. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D514-517. PMCID: PMC539987

49. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 2015 Mar 31;12(3):e1001779. PMCID: PMC4380465

50. The “All of Us” Research Program. New England Journal of Medicine. Massachusetts Medical Society; 2019 Aug 15;381(7):668–676. PMID: 31412182

51. Bastarache L, Hughey JJ, Hebbring S, Marlo J, Zhao W, Ho WT, Van Driest SL, McGregor TL, Mosley JD, Wells QS, Temple M, Ramirez AH, Carroll R, Osterman T, Edwards T, Ruderfer D, Velez Edwards DR, Hamid R, Cogan J, Glazer A, Wei WQ, Feng Q, Brilliant M, Zhao ZJ, Cox NJ, Roden DM, Denny JC. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science. 2018 16;359(6381):1233–1239. PMCID: PMC5959723

52. Bastarache L, Hughey JJ, Goldstein JA, Bastraache JA, Das S, Zaki NC, Zeng C, Tang LA, Roden DM, Denny JC. Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. Journal of the American Medical Informatics Association. 2019 Dec 1;26(12):1437–1447.

53. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang Z, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld J, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, Barnes IHA, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012 Feb 17;335(6070):823–828. PMCID: PMC3299548

54. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O’Donnell-Luria AH,

Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Neale BM, Daly MJ, MacArthur DG. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434–443.

55. Singer-Berk M, Gudmundsson S, Baxter S, Seaby EG, England E, Wood JC, Son RG, Watts NA, Karczewski KJ, Harrison SM, MacArthur DG, Rehm HL, O'Donnell-Luria A. Advanced variant classification framework reduces the false positive rate of predicted loss-of-function variants in population sequencing data. Am J Hum Genet. 2023 Sep 7;110(9):1496–1508. PMCID: PMC10502856

56. Gudmundsson S, Singer-Berk M, Stenton SL, Goodrich JK, Wilson MW, Einson J, Watts NA, Lappalainen T, Rehm HL, MacArthur DG, O'Donnell-Luria A. Exploring penetrance of clinically relevant variants in over 800,000 humans from the Genome Aggregation Database. bioRxiv. 2024 Jun 13;2024.06.12.593113. PMCID: PMC11195293

57. Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, Cox E, Davidson C, Ermolaeva O, Farrell CM, Fatima R, Gil L, Goldfarb T, Gonzalez JM, Haddad D, Hardy M, Hunt T, Jackson J, Joardar VS, Kay M, Kodali VK, McGarvey KM, McMahon A, Mudge JM, Murphy DN, Murphy MR, Rajput B, Rangwala SH, Riddick LD, Thibaud-Nissen F, Threadgold G, Vatsan AR, Wallin C, Webb D, Flicek P, Birney E, Pruitt KD, Frankish A, Cunningham F, Murphy TD. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022 Apr;604(7905):310–315. PMCID: PMC9007741

58. Forrest IS, Duffy Á, Park JK, Vy HMT, Pasquale LR, Nadkarni GN, Cho JH, Do R. Genome-first evaluation with exome sequence and clinical data uncovers underdiagnosed genetic disorders in a large healthcare system. Cell Rep Med. 2024 Apr 19;5(5):101518. PMCID: PMC11148562

59. Lindeboom RGH, Vermeulen M, Lehner B, Supek F. The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. Nat Genet. 2019 Nov;51(11):1645–1651. PMCID: PMC6858879

60. Dyle MC, Kolakada D, Cortazar MA, Jagannathan S. How to get away with nonsense: Mechanisms and consequences of escape from nonsense-mediated RNA decay. Wiley Interdiscip Rev RNA. 2020 Jan;11(1):e1560. PMCID: PMC10685860

61. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, Chow ED, Kanterakis E, Gao H, Kia A, Batzoglou S, Sanders SJ, Farh KKH. Predicting

Splicing from Primary Sequence with Deep Learning. Cell. 2019 Jan 24;176(3):535-548.e24. PMID: 30661751

62. Breiman L. Random Forests. Machine Learning. 2001 Oct 1;45(1):5–32.

63. Chung CCY, Hue SPY, Ng NYT, Doong PHL, Hong Kong Genome Project, Chu ATW, Chung BHY. Meta-analysis of the diagnostic and clinical utility of exome and genome sequencing in pediatric and adult patients with rare diseases across diverse populations. Genet Med. 2023 Sep;25(9):100896. PMID: 37191093

64. Amendola LM, Jarvik GP, Leo MC, McLaughlin HM, Akkari Y, Amaral MD, Berg JS, Biswas S, Bowling KM, Conlin LK, Cooper GM, Dorschner MO, Dulik MC, Ghazani AA, Ghosh R, Green RC, Hart R, Horton C, Johnston JJ, Lebo MS, Milosavljevic A, Ou J, Pak CM, Patel RY, Punj S, Richards CS, Salama J, Strande NT, Yang Y, Plon SE, Biesecker LG, Rehm HL. Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. Am J Hum Genet. 2016 Jun 2;98(6):1067–1076. PMCID: PMC4908185

65. Schubach M, Maass T, Nazaretyan L, Röner S, Kircher M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. Nucleic Acids Research. 2024 Jan 5;52(D1):D1143–D1154.

66. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, Chow ED, Kanterakis E, Gao H, Kia A, Batzoglou S, Sanders SJ, Farh KKH. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019 Jan 24;176(3):535-548.e24. PMID: 30661751

67. Kingsmore SF, Smith LD, Kunard CM, Bainbridge M, Batalov S, Benson W, Blincow E, Caylor S, Chambers C, Del Angel G, Dimmock DP, Ding Y, Ellsworth K, Feigenbaum A, Frise E, Green RC, Guidugli L, Hall KP, Hansen C, Hobbs CA, Kahn SD, Kiel M, Van Der Kraan L, Krilow C, Kwon YH, Madhavrao L, Le J, Lefebvre S, Mardach R, Mowrey WR, Oh D, Owen MJ, Powley G, Scharer G, Shelnutt S, Tokita M, Mehtalia SS, Oriol A, Papadopoulos S, Perry J, Rosales E, Sanford E, Schwartz S, Tran D, Reese MG, Wright M, Veeraraghavan N, Wigby K, Willis MJ, Wolen AR, Defay. T. A genome sequencing system for universal newborn screening, diagnosis, and precision medicine for severe genetic diseases. Am J Hum Genet. 2022 Sep 1;109(9):1605–1619. PMCID: PMC9502059

68. Adhikari AN, Gallagher RC, Wang Y, Currier RJ, Amatuni G, Bassaganyas L, Chen F, Kundu K, Kvale M, Mooney SD, Nussbaum RL, Randi SS, Sanford J, Shieh JT, Srinivasan R, Sunderam U, Tang H, Vaka D, Zou Y, Koenig BA, Kwok PY, Risch N, Puck JM, Brenner SE. The Role of Exome Sequencing in Newborn Screening for Inborn Errors of Metabolism. Nat Med. 2020 Sep;26(9):1392–1397. PMCID: PMC8800147

69. Reich C, Ostropolets A, Ryan P, Rijnbeek P, Schuemie M, Davydov A, Dymshyts D, Hripcsak G. OHDSI Standardized Vocabularies-a large-scale centralized reference ontology for international data harmonization. J Am Med Inform Assoc. 2024 Feb 16;31(3):583–590. PMCID: PMC10873827

70. OMOP Common Data Model [Internet]. [cited 2024 Aug 20]. Available from: https://ohdsi.github.io/CommonDataModel/

71. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, Danis D, Balagura G, Baynam G, Brower AM, Callahan TJ, Chute CG, Est JL, Galer PD, Ganesan S, Griese M, Haimel M, Pazmandi J, Hanauer M, Harris NL, Hartnett MJ, Hastreiter M, Hauck F, He Y, Jeske T, Kearney H, Kindle G, Klein C, Knoflach K, Krause R, Lagorce D, McMurry JA, Miller JA, Munoz-Torres MC, Peters RL, Rapp CK, Rath AM, Rind SA, Rosenberg AZ, Segal MM, Seidel MG, Smedley D, Talmy T, Thomas Y, Wiafe SA, Xian J, Yüksel Z, Helbig I, Mungall CJ, Haendel MA, Robinson PN. The Human Phenotype Ontology in 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D1207–D1217. PMCID: PMC7778952

72. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, Felix V, Jeng L, Bearer C, Lichenstein R, Bisordi K, Campion N, Hyman B, Kurland D, Oates CP, Kibbey S, Sreekumar P, Le C, Giglio M, Greene C. Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res. 2019 08;47(D1):D955–D962. PMCID: PMC6323977

73. Orphanet [Internet]. [cited 2024 Aug 20]. Available from: https://www.orpha.net/

74. Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine JP, Gargano M, Harris NL, Matentzoglu N, McMurry JA, Osumi-Sutherland D, Cipriani V, Balhoff JP, Conlin T, Blau H, Baynam G, Palmer R, Gratian D, Dawkins H, Segal M, Jansen AC, Muaz A, Chang WH, Bergerson J, Lauledderkind SJF, Yüksel Z, Beltran S, Freeman AF, Sergouniotis PI, Durkin D, Storm AL, Hanauer M, Brudno M, Bello SM, Sincan M, Rageth K, Wheeler MT, Oegema R, Lourghi H, Della Rocca MG, Thompson R, Castellanos F, Priest J, Cunningham-Rundles C, Hegde A, Lovering RC, Hajek C, Olry A, Notarangelo L, Similuk M, Zhang XA, Gómez-Andrés D, Lochmüller H, Dollfus H, Rosenzweig S, Marwaha S, Rath A, Sullivan K, Smith C, Milner JD, Leroux D, Boerkoel CF, Klion A, Carter MC, Groza T, Smedley D, Haendel MA, Mungall C, Robinson PN. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. 2019 Jan 8;47(Database issue):D1018–D1027. PMCID: PMC6324074

75. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, Becker A, Bennett R, Berry A, Bhai J, Bhurji SK, Bignell A, Boddu S, Branco Lins PR, Brooks L, Ramaraju SB, Charkhchi M, Cockburn A, Da Rin Fiorretto L, Davidson C, Dodiya K, Donaldson S, El Houdaigui B, El Naboulsi T, Fatima R, Giron CG, Genez T, Ghattaoraya GS, Martinez JG, Guijarro C, Hardy M, Hollis Z, Hourlier T, Hunt T, Kay M, Kaykala V, Le T, Lemos D, Marques-Coelho D, Marugán JC, Merino GA, Mirabueno LP, Mushtaq A, Hossain SN, Ogeh DN, Sakthivel MP, Parker A, Perry

M, Piližota I, Prosovetskaia I, Pérez-Silva JG, Salam AIA, Saraiva-Agostinho N, Schuilenburg H, Sheppard D, Sinha S, Sipos B, Stark W, Steed E, Sukumaran R, Sumathipala D, Suner MM, Surapaneni L, Sutinen K, Szpak M, Tricomi FF, Urbina-Gómez D, Veidenberg A, Walsh TA, Walts B, Wass E, Willhoft N, Allen J, Alvarez-Jarreta J, Chakiachvili M, Flint B, Giorgetti S, Haggerty L, Ilsley GR, Loveland JE, Moore B, Mudge JM, Tate J, Thybert D, Trevanion SJ, Winterbottom A, Frankish A, Hunt SE, Ruffier M, Cunningham F, Dyer S, Finn RD, Howe KL, Harrison PW, Yates AD, Flicek P. Ensembl 2023. Nucleic Acids Res. 2023 Jan 6;51(D1):D933–D941. PMCID: PMC9825606

76. pyensembl package — pyensembl 0.8.10 documentation [Internet]. [cited 2024 Aug 20]. Available from: https://pyensembl.readthedocs.io/en/latest/pyensembl.html

77. Luebbert L, Pachter L. Efficient querying of genomic reference databases with gget. Bioinformatics. 2023 Jan 1;39(1):btac836. PMCID: PMC9835474

78. Tan AL, Gonçalves RS, Yuan W, Brat GA, EHR TC for CC of C 19 by, Gentleman R, Kohane IS. Implications of mappings between ICD clinical diagnosis codes and Human Phenotype Ontology terms [Internet]. arXiv; 2024 [cited 2024 Aug 16]. Available from: http://arxiv.org/abs/2407.08874

79. SNOMED CT [Internet]. U.S. National Library of Medicine; [cited 2020 Jul 10]. Available from: https://www.nlm.nih.gov/healthit/snomedct/index.html

80. McArthur E, Bastarache L, Capra JA. Linking rare and common disease vocabularies by mapping between the human phenotype ontology and phecodes. JAMIA Open. 2023 Apr;6(1):ooad007. PMCID: PMC9976874

81. Bastarache L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. Annu Rev Biomed Data Sci. 2021 Jul 20;4:1–19. PMCID: PMC9307256

82. Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. Bull Med Libr Assoc. 1993 Apr;81(2):217–222. PMCID: PMC225764

83. Pang C, Sollie A, Sijtsma A, Hendriksen D, Charbon B, de Haan M, de Boer T, Kelpin F, Jetten J, van der Velde JK, Smidt N, Sijmons R, Hillege H, Swertz MA. SORTA: a system for ontology-based re-coding and technical annotation of biomedical phenotype data. Database (Oxford). 2015;2015:bav089. PMCID: PMC4574036

84. Athena [Internet]. [cited 2024 Aug 20]. Available from: https://athena.ohdsi.org/search-terms/start

85. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D,

Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(85):2825–2830.

86. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, Benner C, Liu D, Locke AE, Balasubramanian S, Yadav A, Banerjee N, Gillies CE, Damask A, Liu S, Bai X, Hawes A, Maxwell E, Gurski L, Watanabe K, Kosmicki JA, Rajagopal V, Mighty J, Jones M, Mitnaul L, Stahl E, Coppola G, Jorgenson E, Habegger L, Salerno WJ, Shuldiner AR, Lotta LA, Overton JD, Cantor MN, Reid JG, Yancopoulos G, Kang HM, Marchini J, Baras A, Abecasis GR, Ferreira MAR. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature. 2021;599(7886):628–634. PMCID: PMC8596853

87. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. GigaScience. 2021 Feb 1;10(2):giab008.

88. Hail Team. Hail 0.2. Available from: https://github.com/hail-is/hail

89. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome Biology. 2016 Jun 6;17(1):122.

90. UK Biobank Whole Exome Sequencing 300k Release: Analysis Best Practices [Internet]. Available from: https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/UKB_WES_AnalysisBestPractices.pdf

91. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015 Feb 25;4:7. PMCID: PMC4342193

92. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. Proceedings of the 9th Python in Science Conference. 2010 Jan 1;2010.

93. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods. 2010 Apr;1(2):97–111. PMID: 26061376

94. Wang X. Firth logistic regression for rare variant association tests. Front Genet [Internet]. Frontiers; 2014 [cited 2021 Aug 6];0. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2014.00187/full

95. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P.

SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. Nature Publishing Group; 2020 Mar;17(3):261–272.

96. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An Introduction to Variational Methods for Graphical Models. Machine Learning. 1999 Nov 1;37(2):183–233.

97. Beaumont RN, Hawkes G, Gunning AC, Wright CF. Clustering of predicted loss-of-function variants in genes linked with monogenic disease can explain incomplete penetrance. Genome Medicine. 2024 Apr 26;16(1):64.

98. de Sainte Agathe JM, Filser M, Isidor B, Besnard T, Gueguen P, Perrin A, Van Goethem C, Verebi C, Masingue M, Rendu J, Cossée M, Bergougnoux A, Frobert L, Buratti J, Lejeune É, Le Guern É, Pasquier F, Clot F, Kalatzis V, Roux AF, Cogné B, Baux D. SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation. Human Genomics. 2023 Feb 10;17(1):7.

99. pandas: powerful Python data analysis toolkit [Internet]. pandas; 2022 [cited 2022 Apr 18]. Available from: https://github.com/pandas-dev/pandas