Common genetic variation associated with Mendelian disease severity revealed through cryptic phenotype analysis

Authors:
David R Blair[1*], Thomas J Hoffmann[2,3], Joseph T Shieh[1,2*]

Affiliations:
[1]Division of Medical Genetics, Department of Pediatrics, Benioff Children's Hospital
[2]Institute for Human Genetics
[3]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA

Corresponding Authors:
[*]David.Blair@ucsf.edu; Joseph.Shieh2@ucsf.edu

Abstract

Clinical heterogeneity is common in Mendelian disease, but small sample sizes make it difficult to identify specific contributing factors. However, if a rare disease represents the severely affected extreme of a spectrum of phenotypic variation, then modifier effects may be apparent within a larger subset of the population. Analyses that take advantage of this full spectrum could have substantially increased power. To test this, we developed cryptic phenotype analysis (CPA), a model-based approach that uses symptom data to infer latent quantitative traits that capture disease-related phenotypic variability. By applying this approach to 50 Mendelian diseases in two large cohorts of patients, we found that these quantitative traits reliably captured disease severity. We then conducted genome-wide association analyses for five of the inferred cryptic phenotypes, uncovering common variation that was predictive of Mendelian disease-related diagnoses and outcomes. Overall, this study highlights the utility of computationally derived phenotypes and biobank-scale cohorts for investigating the complex genetic architecture of Mendelian diseases.

Advances in sequencing technology, cohort generation, and data dissemination have enabled the rapid identification of thousands of rare genetic variants associated with Mendelian diseases[1,2]. A great deal of this success can be attributed to their relatively simple genetic architectures: they are predominantly caused by deleterious alleles clustered within a limited number of genomic loci. Nevertheless, clinical heterogeneity is commonly observed among cases[1,3,4]. For example, Marfan Syndrome, an autosomal dominant disorder caused by mutations in the *FBN1* gene, is associated with cardiovascular, ocular, skeletal and even pulmonary abnormalities. Individuals with pathogenic *FBN1* alleles rarely manifest all of the associated symptoms[5], and even individuals within the same family can display disparate phenotypes[6]. Some of the clinical variability observed among Mendelian disease cases is attributable to allelic heterogeneity[1,3], but multiple lines of evidence also suggest a role for environmental and genetic background effects[4,7–11].

The identification of specific factors that modify Mendelian disease severity is inherently limited by the low prevalence of these disorders. Generally, it is difficult (but not impossible[12,13]) to construct cohorts of affected cases that are large enough to identify genetic and environmental modifiers, especially if they have relatively modest effect sizes. Given this limitation, many studies that investigate modifier effects have instead relied on model organisms[14,15] or the integration of orthogonal analyses[16,17]. As an alternative approach, we and others hypothesize that some Mendelian disorders may represent the severely affected extreme of a spectrum of pathologic variation. For conditions like familial hypercholesterolemia[18], hereditary breast cancer[19], and long QT syndrome[20], this relationship is well documented. As a result, the interplay between rare and common genetic variation has

been systematically investigated[21–24]. In each of these examples, however, the analyses were possible because the condition of interest mapped to a univariate (often quantitative) phenotype. For Mendelian disorders that instead map to complex arrays of disparate symptoms, investigating the interplay between common and rare genetic variation becomes substantially more difficult.

With this in mind, we developed a probabilistic, model-based approach that infers latent quantitative traits that capture Mendelian disease severity using their diagnosed symptoms (Cryptic Phenotype Analysis). We then systematically tested the method on 50 different Mendelian disorders in two independent patient cohorts (UCSF Clinical Data Warehouse [UCSF][25], UK Biobank [UKBB][26]), uncovering multiple traits that captured disease severity. To validate the latent phenotype model, we used exome sequencing data to demonstrate that pathogenic variation in known disease genes was associated with the inferred traits. Finally, we performed genome-wide association studies (GWAS) to identify common variation (in the form of polygenic scores) that is associated with cryptic phenotype severity and Mendelian-disease related outcomes. This approach replicated the known architecture of a well-characterized genetic condition (α-1-Antitrypsin Deficiency [A1ATD]) while also identifying common variant modifiers for two Mendelian kidney diseases: Alport Syndrome (AS) and Autosomal Dominant Polycystic Kidney Disease (ADPKD). Overall, this study suggests that phenotype-driven approaches applied to biobank-scale data represent a powerful method for investigating the complex genetic architecture of rare diseases.

Results

Overview of a phenotype-driven approach for identifying common variant Mendelian disease modifiers

Figure 1 outlines the approach taken to identify common-variant modifiers of Mendelian disease severity. It assumes that the Mendelian disorder of interest maps to the severely affected extreme of a spectrum of phenotypic variation (**Figure 1A, upper left**). Furthermore, this trait is not limited to the Mendelian disease cases but is present throughout a larger subset of the population. Critically, this spectrum of variation cannot be measured directly. Instead, the trait is analyzed implicitly by a clinician, who translates their observations into a set of symptoms (**Figure 1A, upper left**). These symptoms are then documented in the medical record, typically as a combination of structured and unstructured data. Building upon previous work[9,27,28], we aligned structured electronic medical record (EMR) data (i.e. ICD10 diagnostic codes[29], see **Supplemental Figure 1** and **Methods**) to the symptoms annotated within the Human Phenotype Ontology[30]. This enabled us to construct a symptom matrix that encodes the severity of a specific Mendelian disease (**Figure 1A, right**), which can then be used to recover the cryptic, quantitative trait of interest (**Figure 1B**).

The process of decoding an observed symptom matrix into an underlying cryptic phenotype is equivalent to a form of matrix decomposition (**Figure 1B**). In this scenario, the symptom matrix is decomposed into a risk function (**Figure 1B, upper right**) and collection of one or more latent phenotypes (**Figure 1B, lower right**). There are numerous ways to perform matrix decomposition[31]. Using methods developed for machine learning[32,33], we designed a simple probability model for the observed symptom matrix that preserved its binary nature and

enabled accurate, scalable inference of the desired latent phenotypes (see **Methods**). Note, the recovery of these phenotypes is inherently limited by the loss of information that occurs when translating a quantitative trait into collection of symptoms. Therefore, inferred cryptic phenotypes will be inherently noisy (see **Figure 1B, lower right** for example) unless the matrix contains hundreds of distinct observations, which is unrealistic for most rare diseases.

There is no guarantee that cryptic phenotypes inferred using this approach capture the severity of the intended Mendelian diseases, as the method is unsupervised. Therefore, we performed multiple analyses to ensure that the inferred traits reliably modeled the phenotypic variability of interest (see **Figure 1C**, study overview). We hypothesized that genetic factors associated with this variability are consistent across the full spectrum of phenotype severity. As a result, genetic modifiers identified in more mildly affected individuals should be predictive of outcomes in Mendelian disease cases. To test this, we used GWAS to identify common variation associated with each cryptic phenotype (**Figure 1C**, bottom). Using a withheld sample of unrelated control and rare-disease affected individuals (as determined by genetic data), we confirmed that the identified common variant effects were indeed associated with cryptic phenotype severity, disease-related laboratory measurements, and symptom onset/progression.

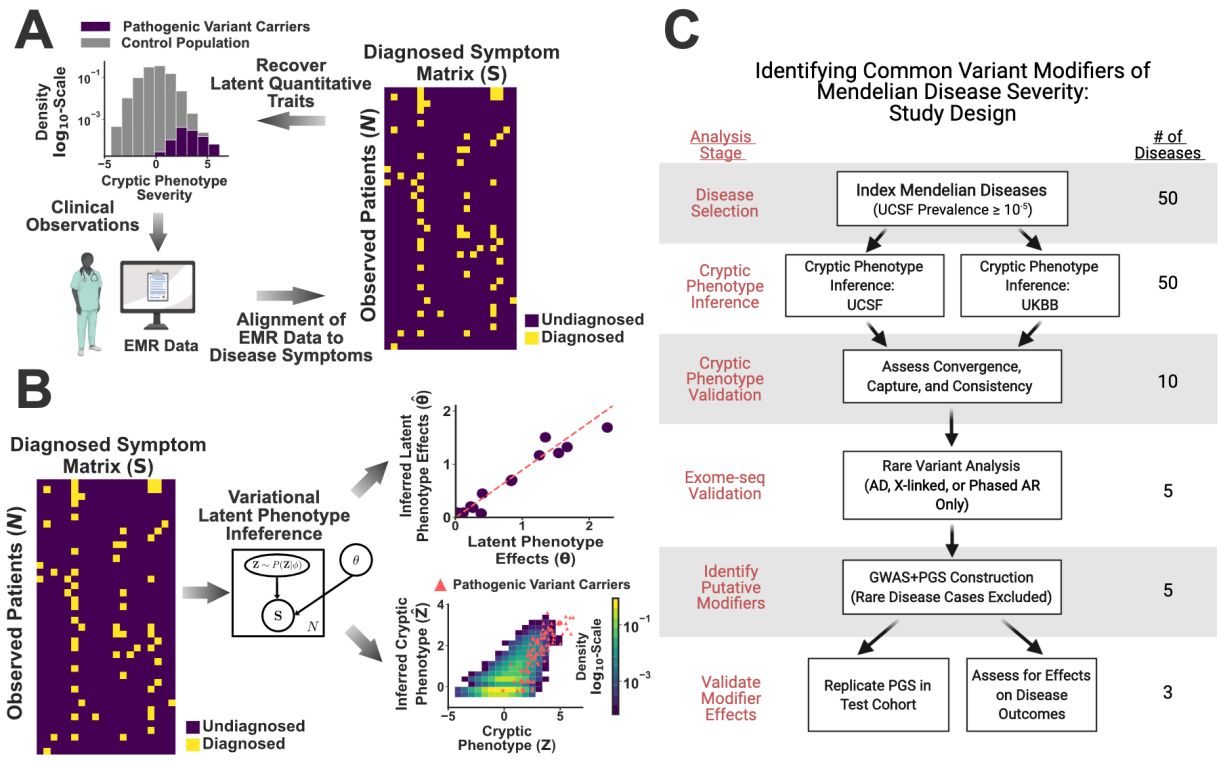

Figure 1: A phenotype-driven approach to identifying common variant modifiers. (A): Schematic illustrating the assumptions underlying cryptic phenotypes and the proposed workflow. (B): Illustration of the model-based approach to symptom matrix decomposition and cryptic phenotype recovery. (C): Flow diagram describing the approach to inferring and validating cryptic phenotypes, which were subsequently used to identify common variant modifiers. UCSF: UCSF Clinical Data Warehouse; UKBB: UK Biobank. Figure 1C was created using Biorender.com.

Quantifying disease-related severity through Cryptic Phenotype Analysis (CPA)

The approach to cryptic phenotype inference relies on fitting a generative probability model to the observed disease symptom matrix. Due to the unsupervised nature of this inference, the latent phenotypes inferred with this approach may or may not capture the severity of the desired Mendelian disease. To circumvent this issue, we performed cryptic phenotype inference only for those diseases that: 1) mapped to specific diagnoses available in structured EMR data and 2) had prevalence of at least 10^{-5} in the UCSF dataset (to ensure adequate sample size for validation, see **Supplemental Data File 1** for complete list). Generative probability models were fit to the symptom matrices for each of the 50 Mendelian

disorders meeting these criteria within both the UCSF (*N*≈1.2 million) and UKBB (N≈500,000) datasets. Consistent models were recovered for 38 of the 50 disorders (**Methods**), with the remainder suffering from convergence issues in at least one of the two datasets (**Figure 1C**).

To ensure that the inferred cryptic phenotypes captured the variability of the intended Mendelian disease, we assessed whether the trait was systematically elevated among already diagnosed cases using withheld testing data (**Figure 2A**, exemplar Mendelian disease HHT). For 31 of the 38 disorders, the cryptic phenotypes were significantly increased among the diagnosed cases in the UCSF dataset (Bonferroni-corrected bootstrapped *P*-value < 0.05, **Supplemental Data File 6**). To verify that the cryptic phenotypes were not dataset dependent, symptom matrix probability models were independently inferred using the UKBB, a population with different ascertainment, demographics, and healthcare infrastructure[26]. For 18 of the 31 disorders, the model inferred with the UKBB dataset reproduced the elevated cryptic phenotypes among withheld cases (Bonferroni-corrected bootstrapped *P*-value < 0.05, **Supplemental Data File 5**).

Although the cryptic phenotype models replicated in both datasets for nearly 40% of the original 50 conditions, their performance (with respect to increased severity among diagnosed cases) in the UKBB was systematically worse (**Figure 2A** and **Figure 2B** for HHT; **Figure 2D** for global comparison; unpaired T-test *P*-value=0.003). The source of this decreased performance is likely multifactorial. For example, the ICD10 encoding within the UKBB is less granular (see **Methods**). This in turn decreases the number of symptoms available for model inference, which can lead to decreased performance. Consistent with this hypothesis, we note that much of the difference in dataset performance disappears when models inferred within the UKBB are

applied to the UCSF data (**Figures 2A, 2B** and **2C** for HHT; see **Figures 2D** and **2E** for a global comparison; unpaired T-test *P*-value=0.17). That said, there are many differences between the clinical datasets in general (sample sizes, population demographics, data provenance, etc.), and it is difficult to disentangle all potential factors. Ideally, cryptic phenotypes would be jointly inferred in both datasets simultaneously, allowing their unique information to be shared systematically. However, because our follow up genetic analyses can only be performed in the UKBB (the UCSF dataset lacks genetic information), all subsequent analyses were performed using models inferred in the UKBB.

Note, even with stringent filtering, the models that replicated within both datasets had variable consistency (as assessed using the R^2 among their inferred cryptic phenotypes, **Figure 2F**). Ultimately, 10 of the 18 disorders resulted in phenotype models that generated R^2 values among the inferred traits ≥ 0.2. From this set of 10 conditions, five had a known genetic mechanism that could be directly ascertained within UKBB data (autosomal dominant, X-linked, or phased autosomal recessive); these were selected for follow up rare and common variant genetic analyses (**Table 1**). Among this group of five, there was still variability in cryptic phenotype consistency (MFS R^2=0.21 vs. A1ATD R^2=0.89; **Table 1**), which may have affected the performance of downstream analyses.

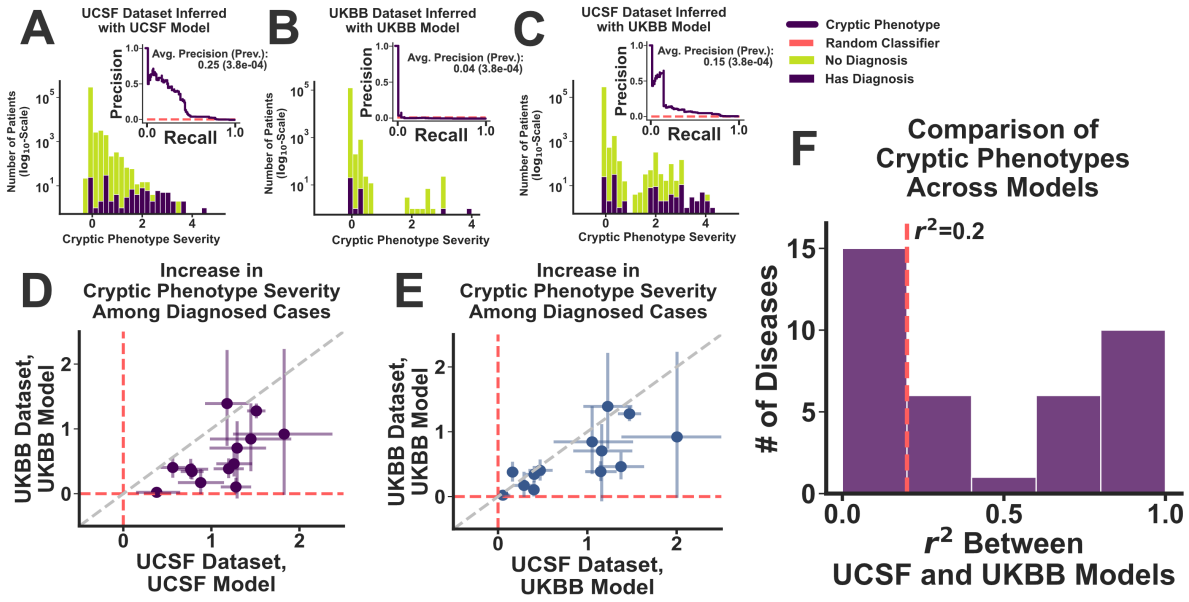

Figure 2: Cryptic phenotype inference in the UCSF and UKBB datasets. (A): Distribution of HHT cryptic phenotype severities among the subjects in the UCSF testing dataset, stratified by their HHT diagnostic status (green: controls; purple: HHT cases). (A, inset): Precision-recall curve for the prediction of HHT diagnoses using the cryptic phenotype. The approximate performance of a random classifier is shown in red. Panel (B) displays the same information for the UKBB dataset, which was generated using an independently inferred phenotype model. Panel (C) displays the same information as (A), except that the UKBB phenotype model is used to generate the cryptic phenotypes in the UCSF dataset. (D, E): The increase in cryptic phenotype severities among diagnosed cases is displayed jointly for both datasets. Panel (D) compares the results of the UCSF model with those generated by UKBB model directly. Panel (E) instead compares the UKBB results with those obtained in the UCSF dataset using the UKBB model. Panel (F): The coefficient of determination (R^2) between the cryptic phenotypes generated by the UCSF and UKBB models was computed for each replicating disease in the UCSF dataset. The resulting distribution over this statistic is displayed.

Table 1: Mendelian diseases selected for cryptic phenotype validation and analyses.

Disease Name	Abbreviation	Cryptic Phenotype Consistency (R^2)	Causal Genes	Variants Analyzed
Alpha-1-antitrypsin Deficiency	A1ATD	0.89	*SERPINA1*	Missense (E342K)
Hereditary Hemorrhagic Telangiectasia	HHT	0.23	*ACVRL1*; *ENG*; *SMAD4*	LP/P ClinVar Variants; Novel LoF
Marfan Syndrome	MFS	0.21	*FBN1*	LP/P ClinVar Variants; Novel LoF
Alport Syndrome	AS	0.45	*COL4A3*; *COL4A4*; *COL4A5*	LP/P ClinVar Variants; Novel LoF
Autosomal Dominant Polycystic Kidney Disease	ADPKD	0.88	*PKD1*; *PKD2*	LP/P ClinVar Variants; Novel LoF

Validating the inferred cryptic phenotypes through exome sequencing

To further validate the cryptic phenotypes inferred for the conditions in **Table 1**, we conducted rare variant association analyses to ensure that these traits could replicate known mechanisms of disease. Because these analyses were conducted for validation rather than discovery, we focused on rare variants that were either: 1) annotated as pathogenic/likely-pathogenic (P/LP) in ClinVar[2] or 2) predicted[34] to be loss-of-function (LoF) alleles (referred to as P/LP variants subsequently, **Supplemental Data File 7** for full list). Linear modeling was performed to assess whether variants were significantly associated with the corresponding cryptic phenotype (**Methods**). For all five disorders, the disease-associated variants had large effect sizes and were significantly associated with their corresponding cryptic phenotypes (**Figures 3A-C** and **Supplemental Figure 6A, 6B**). However, there was significant phenotypic variability seen among P/LP variant carriers, and many of these subjects had few if any apparent symptoms (i.e. cryptic phenotypes=0).

There are multiple factors that may contribute to the phenotypic variability seen among P/LP carriers. First, EMR data is an imperfect proxy for an individual's true symptoms, and it is certainly possible that missing information accounts for significant fraction of this variability. Second, some of the P/LP variants may be misclassified. Consistent with this hypothesis, we note that variants that were flagged due to annotation issues (see **Methods**) tend to have a systematically reduced effect size (see **Figures 3A-C**). Third, confirmation bias could result in the diagnosis of symptoms that would otherwise be left out of the EMR (ex: epistaxis in a known case of HHT), resulting in inflated cryptic phenotypes among diagnosed cases (see **Figures 3A-C, inset**).

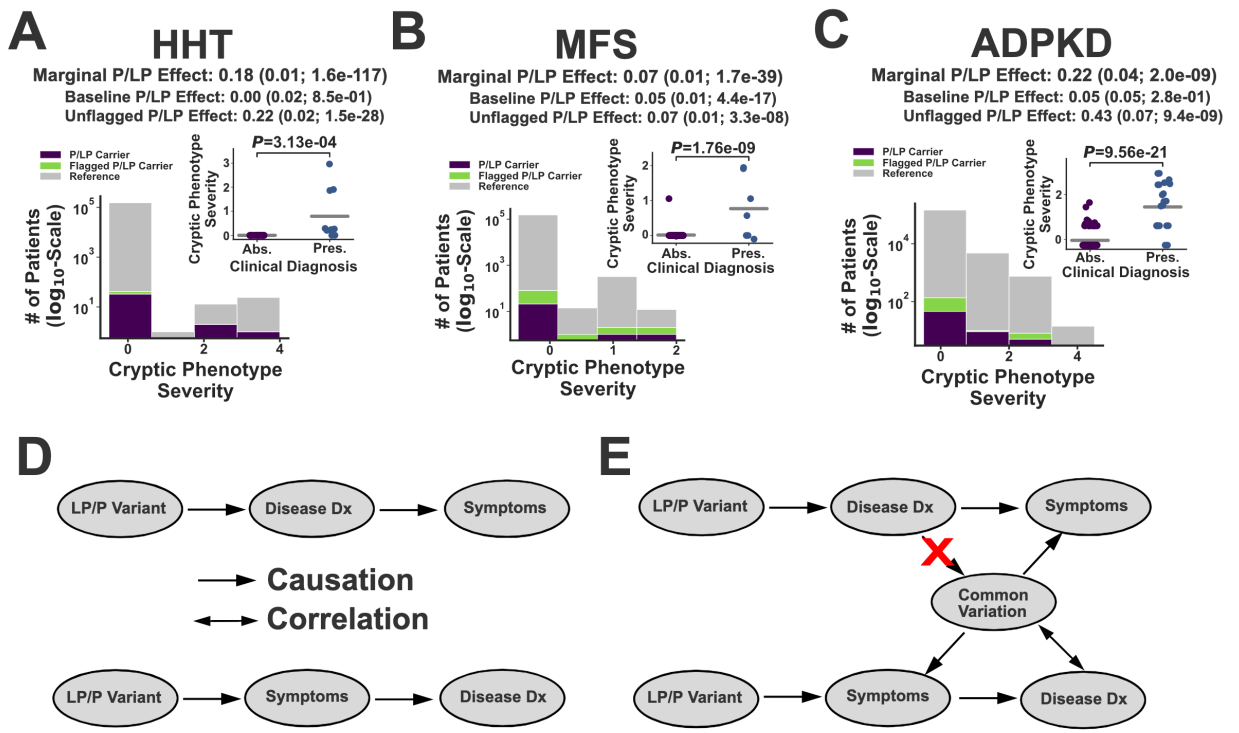

Figure 3: Exome sequencing validation of the inferred cryptic phenotypes. Panels A-C: Distribution of cryptic phenotype severity across three different genotypes: Reference (grey), P/LP Carrier (purple), and Flagged P/LP Carrier (green). The marginal (flagged and unflagged variants) P/LP cryptic phenotype effect size is shown at the top of each panel. The baseline and unflagged variant effects are displayed below the marginal effects. Parentheses contain the standard errors and *P*-values for effects, which were estimated using ordinary least squares. The insets display the cryptic phenotype severity estimates among the P/LP carriers for each condition, stratified by whether the rare disease diagnosis is absent (Abs.) or present (Pres.) in the EMR. Gray bars represent average values, and *P*-values were computed using least squares. (A): Hereditary Hemorrhagic Telangiectasia (HHT). (B) Marfan Syndrome (MFS). (C) Autosomal Dominant Polycystic Kidney Disease (ADPKD). (D): Illustration of the two biases that could lead to increased cryptic phenotypes among diagnosed carriers. Top: post-diagnosis confirmation bias. Bottom: pre-diagnosis ascertainment bias. (E): Common variant modifiers could be used to distinguish between these competing models, as common variation would only be correlated with disease diagnosis under the ascertainment bias scenario.

That said, the inflated cryptic phenotypes would also be observed if only the most severely affected individuals receive a rare disease diagnosis. In other words, the inflation of cryptic phenotypes seen among diagnosed pathogenic variant carriers could instead be driven by ascertainment bias at the level of disease diagnosis (**Figure 3D**). The identification of specific genetic modifiers could help differentiate between these two models (**Figure 3E**). Since it is impossible for a disease diagnosis to alter an individual's genotype, an association between common variation that modifies disease expressivity and the diagnosis itself is only consistent

with a model in which symptom severity affects disease ascertainment (**Figure 3E, bottom**). Therefore, investigating a role for common variation in cryptic phenotype severity may serve two purposes. It can identify background genetic variation that may modify symptom severity, but it can also help distinguish different types of bias that may be present within EMR data.

Genome-wide association analyses uncover common variation associated with cryptic phenotype variability

To identify potential common variant cryptic phenotypic modifiers, we first generated two datasets using the UKBB for each disease-trait pair, which we refer to as the training and target cohorts. The training cohort included a set of unrelated subjects with similar ancestry (Caucasian, **Methods**); the corresponding P/LP carriers, diagnosed rare disease cases, and all their 3rd degree or closer relatives were specifically excluded from this dataset (see **Methods**). This training cohort was used for genome-wide association analyses and polygenic prediction model inference (N=294,133-308,381). Alternatively, the target cohort contained a subset of unaffected and unrelated subjects of similar ancestry (to provide power for replication; N=32,682-34,265) plus all the individuals affected by the monogenic disease of interest (after removing any 3rd degree or closer relatives among this subset; N=166-17,163).

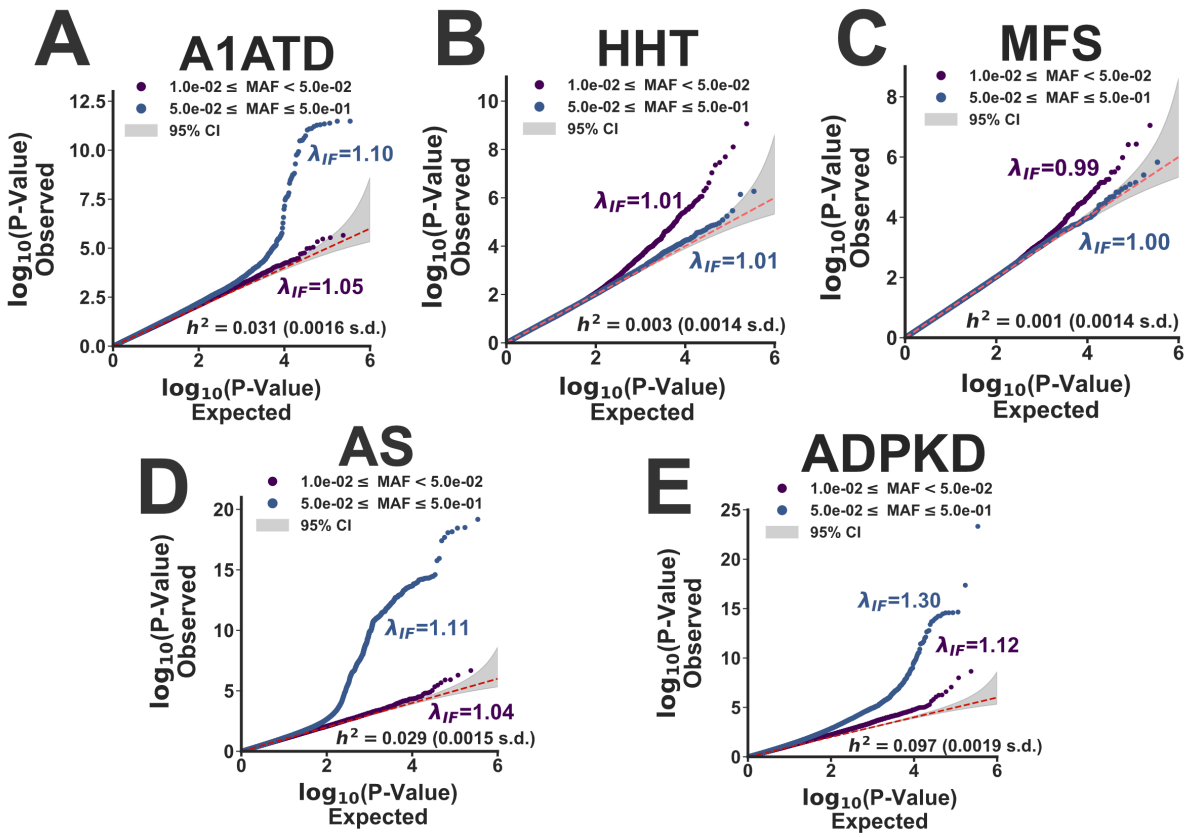

Figure 4: Common variation associated with cryptic phenotype severity. (A-E): Each panel displays the observed versus expected *P*-value quantiles for the cryptic phenotype genome-wide association statistics, stratified by allele frequency (purple: 0.01≤AF≤0.05; blue 0.05≤AF≤0.50). Genomic inflation factors (λ_{IF}) are provided in addition to common variant heritability estimates ($h^2 \pm$ std errors, **Methods**). (A): α-1-Antitrypsin Deficiency. (B) Hereditary Hemorrhagic Telangiectasia. (C): Marfan Syndrome. (D) Alport Syndrome. (E) Autosomal-Dominant Polycystic Kidney Disease.

The results of the genome-wide analyses conducted on the training cohort for the five cryptic phenotypes are summarized using Quantile-Quantile plots in **Figure 4**. For three of the five disorders (A1ATD, AS, and ADPKD, **Figure 4A**, **4D**, and **4E**), the common variant heritability was significantly increased from zero, consistent with a role for genetic background effects in phenotypic variability. For two disorders (HHT and MFS), the heritability was indistinguishable from zero, even though there was evidence for test-statistic inflation at low minor allele frequencies (potentially secondary to the non-Gaussian nature of the cryptic phenotype distribution). These results do not exclude a role for common genetic background effects in the phenotypic variability of these traits. The cryptic phenotype models for both these conditions

showed reduced consistency across datasets (R^2=0.23 and R^2=0.21 for HHT and MFS), suggesting that improved modeling may be able to infer cryptic phenotypes with better performance. Nonetheless, based on these heritability results, polygenic prediction models were inferred (using individual level data[35], see **Methods**) for the cryptic phenotypes belonging to A1ATD, AS, and ADPKD (models provided in **Supplemental Data Files 8-10**). These models were then used to impute polygenic scores (PGS) into the target cohorts so that the detected genetic effects could be replicated and validated (see below).

Cryptic phenotype-associated genetic variation modifies alpha-1 antitrypsin deficiency (A1ATD) severity.

Alpha-1-antitrypsin deficiency (A1ATD) is a relatively common genetic disorder that leads to early-onset emphysema, liver disease, and auto-inflammatory conditions[36]. The Pi*Z allele (*rs28929474*) in *SERPINA1* is the most common cause of severe A1ATD, although the penetrance of this variant is incomplete. The clinical manifestations associated with the Pi*Z allele are known to depend heavily on environmental background effects (smoking, alcohol use, etc.)[37], and common variant modifiers likely also play a significant role[38]. Using the cryptic phenotype approach, we aimed to further investigate the potential effects of background genetic variation on A1ATD severity.

The GWAS conducted on the A1ATD cryptic phenotype (Figure 5A, Manhattan plot) detected three genome-wide significant loci. Not surprisingly, they have all been previously been linked to chronic pulmonary disease, lung function, and smoking[39] (**Supplemental Table 2**). Such results are consistent with the strong effects that smoking is known to have on A1ATD

severity[37]. To further investigate, we examined the interaction between smoking history (measured as reported pack-years; UKBB Data Field: 20161) and the Pi*Z allele using the inferred cryptic phenotype. Symptom severity was substantially elevated among heavy smokers, both within and across the pathogenic genotypes (**Figure 5B** and **5B inset** for Pi*MZ and Pi*ZZ genotypes respectively). The cryptic phenotype polygenic score (PGS) was strongly associated with smoking history (**Figure 5C**), and after regressing out the effects of smoking, the PGS remained associated with phenotypic severity (β_{PGS}=0.02; *P*-value=1.6×10^{-11}). This suggests that the PGS may capture background effects that are independent of smoking history. However, it is important to note that the relationship among smoking history, *SERPINA1* genotype, and polygenic load is likely complex. For example, **Figure 5E** depicts the PGS effects on cryptic phenotype severity among pathogenic variant carriers, stratified by smoking history and genotype. Notably, the PGS effect varies considerably depending on whether an individual has ever smoked, particularly among Pi*ZZ carriers ($\beta_{PGSxPi*ZZ}$=0.41 among smokers vs. $\beta_{PGSxPi*ZZ}$=-0.13 among non-smokers; LR test for smoking-by-PGS interaction effects: *P*-value=2.1×10^{-9}). The source of this variability is uncertain, but we hypothesize that it may be partially driven by smoking cessation/abstinence among more severely affected pathogenic variant carriers (**Supplemental Figure 7B**).

To further validate the inferred PGS, we tested whether polygenic load was significantly associated with A1ATD diagnoses. Unfortunately, structured diagnostic data for A1ATD is not available in the UKBB medical records, but A1ATD diagnoses (as provided by a physician) were ascertained as part of a survey that was conducted among the study participants (UKBB Data Field: 22152). Consistent with ascertainment bias at the level of disease diagnosis (see **Figures**

3D and **3E**), we note the cryptic phenotype PGS was significantly associated with the risk for A1ATD diagnosis (Firth-corrected logistic regression LR test; β_{PGS} =0.50; *P*-value=0.01), which was compounded by the large (and expected) effects of the pathogenic genotypes themselves (β_{PiZZ}=8.98; *P*-value=1.5×10^{-26}; β_{PiMZ}=4.75; *P*-value=2.3×10^{-15}).

To determine if increased polygenic load translated to other outcomes, we examined the variability in age-of-onset for chronic obstructive pulmonary disease (COPD; Data Field: 42016) among the different genotypes within our target cohort. Consistent with prior knowledge, both the Pi*MZ and Pi*ZZ genotypes resulted in more frequent and earlier onset COPD (β_{PiZZ}=2.8$\pm$0.2, *P*-value=2.2×10^{-25}; β_{PiMZ}=0.16$\pm$0.06, *P*-value=0.02; Cox-Proportional Hazard regression, **Methods**). Furthermore, smoking history (in pack-years) had a profound effect on COPD onset (β_{Smoke}=0.47$\pm$0.01; *P*-value=6.8×10^{-239}), which included significant smoking-by-genotype interaction effects (LR Test *P*-value=1.9×10^{-5}). Finally, after correcting for smoking history, the cryptic phenotype PGS had a significant, additive effect (β=0.20$\pm$0.03; *P*-value=2.5×10^{-15}, **Figure 5E**), which persisted even when limiting the analyses to only those individuals that carry the Pi*MZ/Pi*ZZ genotypes (β_{PGS}=0.19$\pm$0.04; *P*-value=4.7×10^{-6}; see **Figure 5F**). This additive PGS effect also replicated in spirometry measurements (**Supplemental Figure 7C**). Note, we performed this analysis using only those subjects with the Pi*ZZ genotype, but the sample size (N=102) was likely too small to detect a significant effect (β_{PGS}=0.17$\pm$0.24; *P*-value=0.46, see **Supplemental Figure 7D**). In total, these results indicate that the cryptic phenotype for A1ATD replicates much of the known architecture for A1ATD[37] while also identifying common genetic variation that modifies symptom expression and severity.

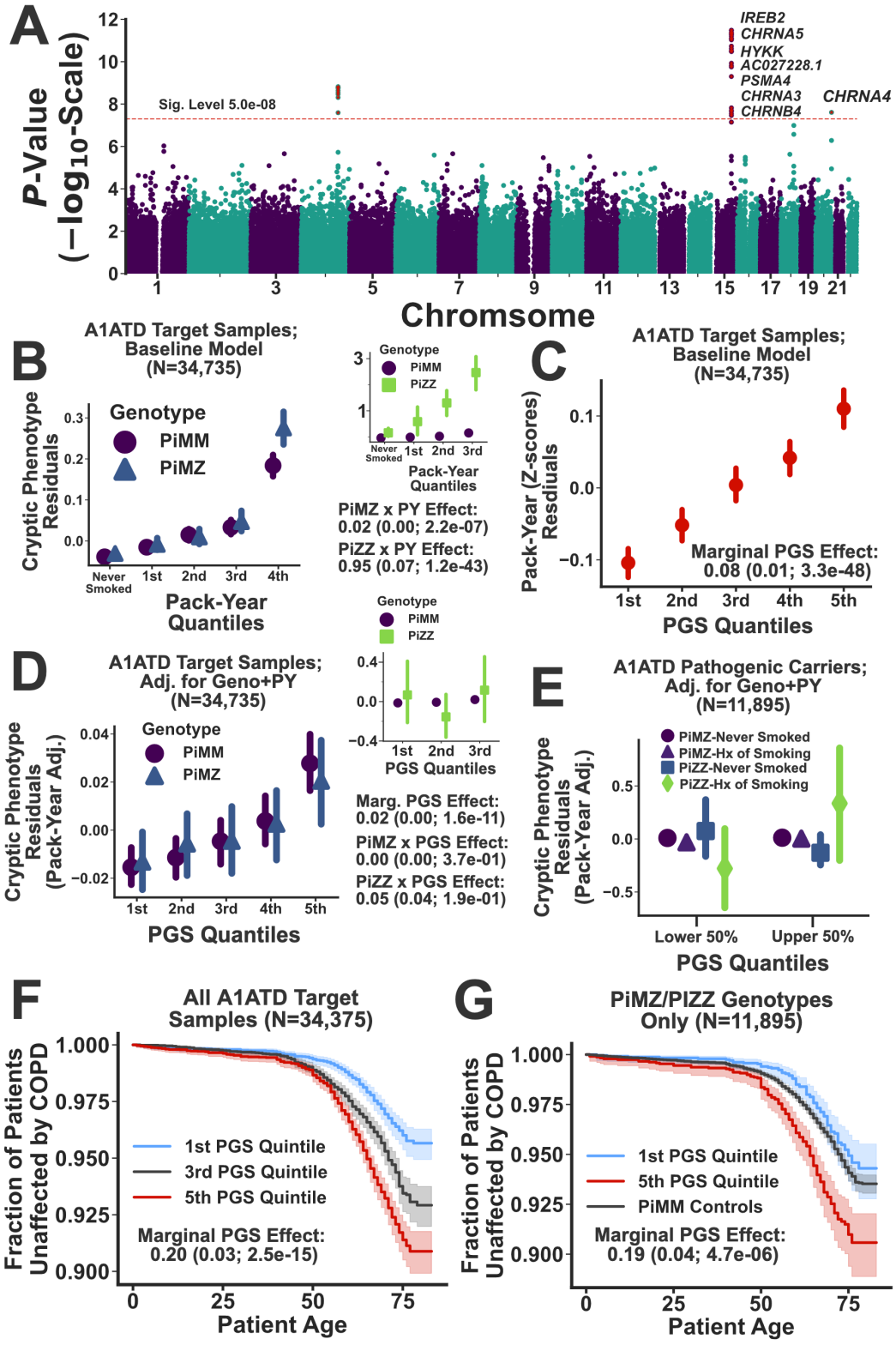

Figure 5: Cryptic phenotype-associated genetic variation modifies A1ATD severity. (A): Manhattan plot displaying the genome-wide association statistics as a function of chromosomal position. Genes were assigned to loci using FUMA[40]. The 5×10^{-8} significance threshold is displayed as dashed red line, and significant variants (along with those SNPs in linkage disequilibrium) are highlighted with red stars. (B): Cryptic phenotype residuals are stratified by the Pi*MZ/Pi*MM genotypes

and plotted as function of pack-year quantiles. Inset: Cryptic phenotype residuals plotted against pack-year quantiles, now stratified by the Pi*ZZ/Pi*MM genotypes. The association statistics for the genotype x smoking interaction terms are included below the inset. (C): Smoking history (expressed as pack-years after adjusting for baseline covariates) is plotted against PGS quantiles. (D): Cryptic phenotype residuals, after adjusting for baseline covariates, genotype, and smoking history, are plotted against PGS quantiles and stratified by the Pi*MZ/Pi*MM genotypes. The inset displays the same information but now stratified by the Pi*ZZ/Pi*MM genotypes. The association statistics for the PGS effects are included below the inset. (E) Cryptic phenotype residuals within the upper and lower 50th percentiles of the PGS distribution are stratified by both genotype and smoking history. (F): Kaplan-Meier curves for COPD onset after stratifying the target cohort according PGS quintiles. (G): Same as in (F), except subjects were restricted to the Pi*MZ/Pi*ZZ genotypes only. The PGS effect size and association statistics (computed using a Cox PH model) are provided for the subjects depicted in both (F) and (G).

Cryptic phenotype analysis reveals common genetic variation associated with monogenic kidney disease severity and outcomes.

Alport Syndrome and Autosomal-Dominant Polycystic Kidney Disease represent two of the most common forms of hereditary kidney disease[41,42], although their underlying molecular pathophysiology is distinct. Alport Syndrome is a genetically heterogenous Type IV collagenopathy linked to the *COL4A3*, *COL4A4*, and *COL4A5* genes. The collagen isoforms produced by these genes play an integral role in maintaining basement membrane integrity within the glomerulus[43], cochlea[44] and eye[45]. In its mildest form (often referred Thin Basement Membrane Nephropathy[46]), the disorder is associated with persistent hematuria that uncommonly progresses to chronic kidney disease. In such cases, the disease is typically caused by heterozygous pathogenic variants located within any of the three causative genes. In the severe form, the disease is characterized by end-stage renal disease, hearing loss, and vision abnormalities. Such individuals typically harbor hemizygous variants in *COL4A5* (X-linked) or biallelic pathogenic variants in *COL4A3*/*COL4A4*[41].

Alternatively, ADPKD is linked to the *PKD1* and *PKD2* genes, which encode two integral membrane proteins that play critical but complex roles in Ca^{2+} regulation and ciliary functioning[42]. Phenotypically, ADPKD leads to chronic kidney disease more consistently,

although there is again a great deal of variability in age of onset and rate of progression[47]. Moreover, extra-renal manifestations are present in a significant fraction of ADPKD patients, and such symptoms include other organ cysts, vascular aneurysms, hernias, and bronchiectasis[48].

To investigate a role for common genetic variation in AS and ADPKD variability, we conducted GWAS on their respective cryptic phenotypes. The results are displayed in **Figures 6A** and **6D**. For Alport Syndrome, three loci reached genome-wide significance (see **Supplemental Table 3**). Interestingly, the locus on chromosome 19 has previously been linked to hematuria[48], and the locus on chromosome 13 is located within the intron of another Type IV collagen isoform (*COL4A2*). This locus has also been linked to neurovascular phenotypes[39]. The third locus is proximal to the MHC region on chromosome 6, and due to complex linkage disequilibrium, it has been associated with many disparate phenotypes[39]. The GWAS for ADPKD uncovered 30 independently associated loci (see **Figure 6D** and **Supplemental Table 4**), most of which have been previously linked to kidney disease and blood pressure regulation[39].

After performing the genome-wide association analyses, prediction models were constructed to capture the global effects of polygenic load on cryptic phenotype severity. With respect to Alport Syndrome, the inferred PGS had a significant marginal effect in the withheld target cohort (β_{PGS}=0.03±0.00; *P*-value=9.8×10^{-14}), which was more pronounced among the P/LP carriers ($\beta_{PGSxP/LP}$=0.09±0.03; *P*-value=0.002). Diagnostic data for Alport Syndrome is not available within the UKBB, so we instead focused on two critical outcomes related to the disease: Recurrent and Persistent Hematuria (UKBB Data Field: 132002) and End-Stage Renal Disease (ESRD; UKBB Data Field: 42026). P/LP variants in AS genes were significantly associated

with both outcomes (Persistent Hematuria: $\beta_{P/LP}$=1.06, *P*-value=0.04; ESRD: $\beta_{P/LP}$=1.70, *P*-value=3.7×10^{-4}; Firth-corrected logistic regression), although these effects were less apparent within the age-of-onset data (see **Supplemental Figure 8D** and **8E**). The cryptic phenotype PGS was marginally associated with Persistent Hematuria (β_{PGS}=0.31; *P*-value=0.007; Firth-corrected logistic regression), an effect that was also apparent when modeling the age-of-onset (β_{PGS}=0.31$\pm$0.12; *P*-value=0.03 see **Figure 6C**). Unfortunately, there were too few Persistent Hematuria cases to determine if there was a significant interaction effect between the polygenic background and P/LP variants ($\beta_{PGSxP/LP}$=-0.00$\pm$0.62; *P*-value=0.86). Note, there was no evidence that the cryptic phenotype PGS for Alport Syndrome was associated with ESRD (β_{PGS}=0.03; *P*-value=0.82; Firth-corrected logistic regression). However, it was significantly predictive of urine microalbuminuria (β_{PGS}=3.10$\pm$1.36; *P*-value=0.023; **Supplemental Figure 8C**), suggesting that the PGS correlates with glomerular dysfunction.

As was the case for AS, the PGS constructed using the cryptic phenotype for ADPKD was again strongly associated with the trait in the target cohort (β_{PGS}=0.06$\pm$0.00; *P*-value=3.5×10^{-134}), and like before, the effect was more pronounced among the P/LP carriers ($\beta_{PGSxP/LP}$=0.15$\pm$0.05; *P*-value=0.003; see **Figure 6E**). In contrast to AS, diagnostic data for ADPKD is available within the UKBB. As expected, P/LP carrier status was strongly associated with Mendelian disease diagnoses (Firth-corrected logistic regression; $\beta_{P/LP}$=4.47; *P*-value=3.5×10^{-31}), but the inferred PGS had no discernable marginal (β_{PGS}=0.02; *P*-value=0.67) or interaction ($\beta_{PGSxP/LP}$=0.37; *P*-value=0.13) effects. A substantial fraction of P/LP carriers in the UKBB were diagnosed with ADPKD (specifically, 35% with polycystic kidney disease and 48% with cystic kidney disease in general), so it is possible that these diagnoses lacked the variability

needed to detect interaction effects. Therefore, we also examined if the cryptic phenotype PGS affected ADPKD onset and rate-of-progression.

Cystic Kidney Disease onset (UKBB Data Field: 132533) was modeled as a function of both P/LP carrier status and polygenic load. As expected, ADPKD P/LP carriers were at high risk for early-onset cystic kidney disease (Cox PH model; $\beta_{P/LP}$=3.73±0.29; *P*-value=3.2×10^{-36}), consistent with the known pathophysiology of disorder. Interestingly, there was a significant interaction effect between the cryptic phenotype PGS and P/LP carrier status ($\beta_{PGSxP/LP}$=0.50 ±0.16; *P*-value=0.002; see **Figure 6F**), consistent with a model in which polygenic load modulates ADPKD severity. Because ESRD is the downstream effect of severe cystic kidney disease, we used the onset of this phenotype as a proxy for ADPKD progression. Once again, P/LP carrier status had a profound effect on ESRD onset ($\beta_{P/LP}$=3.91±0.37; *P*-value=7.0×10^{-19}, see **Supplemental Figure 9D**), and there was again a significant interaction effect between P/LP carrier status and polygenic load ($\beta_{PGSxP/LP}$=0.50 ±0.21; *P*-value=0.02; see **Supplemental Figure 9E**). To further verify this effect, we estimated[49] the glomerular filtration rate (eGFR) within our target cohort. P/LP carriers had significantly lower eGFR values ($\beta_{P/LP}$=-15.2 ±2.0; *P*-value=7.2×10^{-15}), and there was again a significant interaction effect between carrier status and the inferred PGS ($\beta_{PGSxP/LP}$=-3.8±1.3; *P*-value=0.005; see **Supplemental Figure 9C**). Overall, these results suggest that polygenic burden is associated with worse outcomes among P/LP carriers, consistent with a role for common variant effects in modifying ADPKD disease severity.

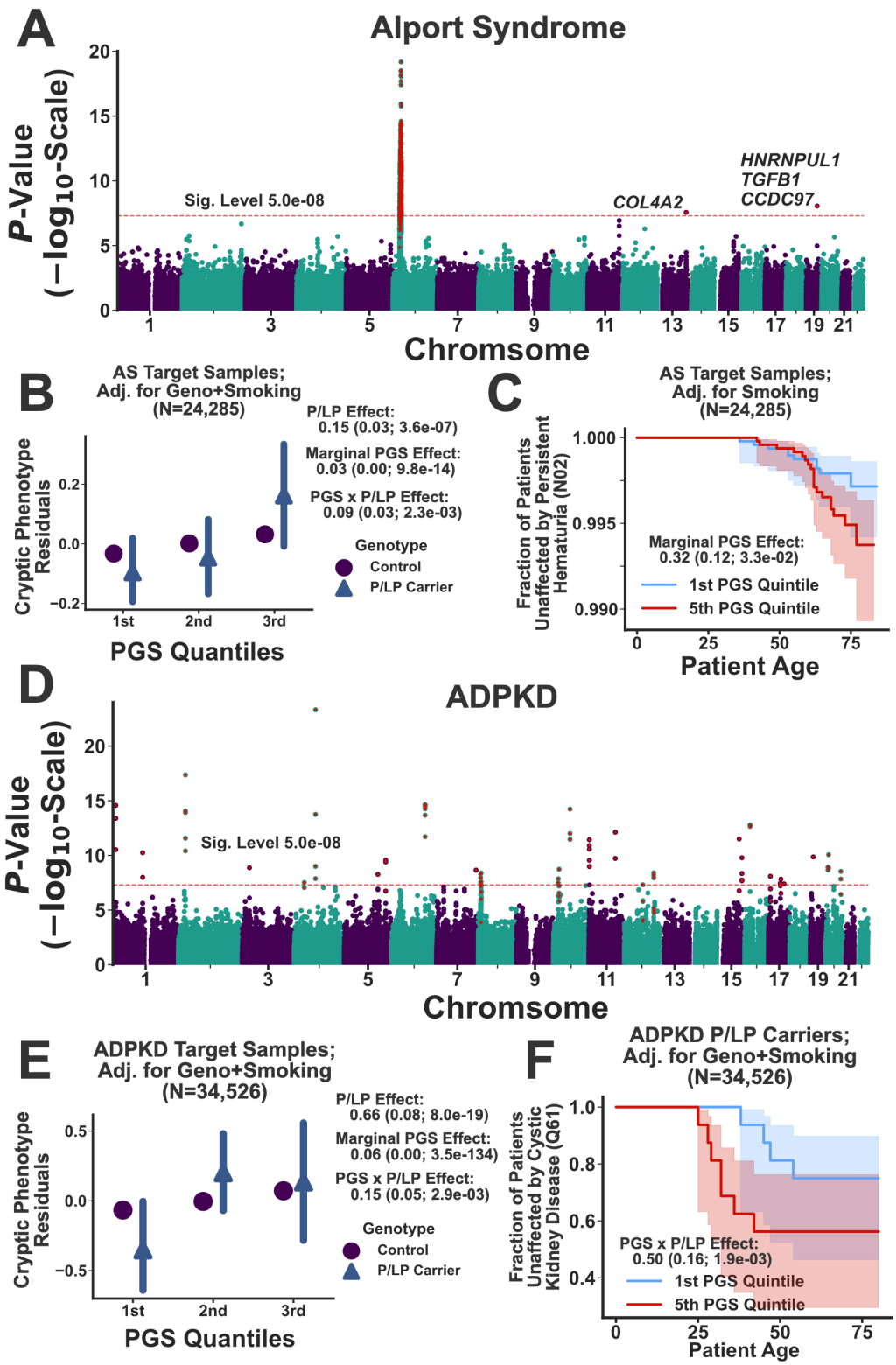

Figure 6: Common variant modifiers of monogenic kidney disease revealed through cryptic phenotype analysis. (A): Manhattan plot displaying the genome-wide association statistics for the AS cryptic phenotype as a function of chromosomal position. Genes were assigned to loci using FUMA[40]. The 5×10^{-8} significance threshold is displayed as dashed red line, and significant variants (along with those SNPs in linkage disequilibrium) are highlighted with red stars. (B): AS cryptic phenotype

residuals, after adjusting for baseline covariates, P/LP genotype, and pack-years, are plotted against PGS quantiles and stratified by the P/LP carrier status. The association statistics for the P/LP variants, the PGS, and their interaction effects are included to the right. (C): Kaplan-Meier curve for Persistent Hematuria (**Methods**) is stratified by PGS quintile. The marginal PGS effect was estimated using a Cox PH model. (D): Manhattan plot displaying the genome-wide association statistics for the ADPKD cryptic phenotype as a function of chromosomal position. (E): ADPKD cryptic phenotype residuals, after adjusting for baseline covariates, P/LP genotype, smoking status (**Methods**), are plotted against PGS quantiles and stratified by the P/LP carrier status. (F): Kaplan-Meier curve for Cystic Kidney Disease onset stratified by PGS quintile. Note, the P/LP carriers are depicted; however, the PGSxP/LP interaction effects were computed using the complete target cohort.

Discussion

Cryptic Phenotype Analysis (CPA) provides a model-based approach for inferring quantitative traits that capture rare disease variability. In the current study, we used these traits to identify common variants putatively associated with the clinical heterogeneity observed among Mendelian disease cases. This approach relies on two assumptions. First, the condition of interest must represent the extreme of a spectrum of pathologic variation. Second, shared genetic factors need to drive phenotypic variability among the mild and severely affected individuals. If true, then the effects of genetic and potentially even environmental modifiers should be detectable within a subset of the population that extends beyond the rare disease cases. Consistent with this hypothesis, CPA enabled us to identify common variation that was putatively associated with Mendelian disease severity. This variation was predictive of disease-related symptoms, laboratory tests, and outcomes in withheld cases. Overall, our study suggests that CPA represents a powerful new method for investigating the genetic architecture of rare disease-associated traits, although there are multiple avenues for further investigation.

Cryptic Phenotype Analysis (CPA) has several attractive properties. First, it performs latent phenotype inference using an unsupervised generative model, thereby directly estimating a quantitative trait that captures symptom variability. Second, its model-based

nature allows cryptic phenotypes to be directly imputed into new datasets, albeit only if the observed data is encoded using the same format. Third, the model is modular (i.e. composed of multiple conditionally independent components), so it could easily be extended in multiple ways (direct incorporation of disease labels, the inclusion of laboratory data, joint modeling across multiple independent datasets, etc.). This could potentially result in more accurate latent phenotypes and increased power for downstream analyses. Alternatively, simpler heuristic[9] and discriminative[50] approaches also exist for quantifying rare disease variability, and additional work is needed to determine if and/or when such indirect approaches can be used to perform the types of genetic analyses illustrated here.

The results from the common variant association analyses demonstrate that polygenic load likely plays a role in Mendelian disease variability. These polygenic effects were detected at the level of the cryptic phenotypes themselves (**Figures 5D**, **5E**, **6B**, and **6E**), but they were also apparent when examining outcomes known to be associated Mendelian disease severity (spirometry measurements, glomerular filtration rate, symptom age-of-onset, etc.). Although the results replicated across Mendelian diseases (i.e. polygenic load was consistently associated with more severe outcomes), it is difficult to replicate the results across datasets, as a unique combination of information (extensive EMR data, genome-wide common variation, exome sequencing data) is required. However, biobanks with linked medical and genetic data are becoming increasingly common[51–55], and as a result, we suspect that phenotype-driven approaches that incorporate the full-spectrum of genetic variation will be more widely applied.

In summary, Cryptic Phenotype Analysis enables the systematic estimation of quantitative traits that capture spectrums of pathologic variation. Although the focus here was

on Mendelian disorders, the approach could theoretically be applied to diseases with even more complex genetic architectures provided that they are associated with a diverse array of symptoms (e.g. systemic lupus erythematosus). Moreover, we inferred these distributions to perform common variant genetic analyses. However, they likely have other applications as well. For example, the cryptic morbidity distributions could be used in conjunction with other data to assist with rare variant annotation and interpretation (see **Figure 3A-C** for examples), particularly in cases where there is limited availability of legacy sequencing data (ex: populations that are under-represented in variant databases). The work described here builds upon a growing number of studies[9,56,57] that demonstrate the utility of applying statistical models of human phenotypes to large-scale medical record and genetic datasets. Ongoing developments in this field will continue to shed light onto the genetic complexity of human diseases.

Methods

Clinical datasets

Phenotypic analyses were conducted using the University of California San Francisco De-Identified Clinical Data Warehouse (UCSF-CDW)[25], a database of structured health information that is made available to UCSF researchers free-of-charge. The data was captured for use on May 31st, 2019 and includes roughly 8 years of clinic visits and inpatient hospitalizations (see Supplemental Methods). Following capture, patient demographic data was aligned to the International Classification of Disease, Tenth Revision, Clinical Modification (ICD10-CM)[29] diagnostic codes available within the medical encounters. The individual diagnostic codes were simplified by collapsing multiple appearances of each code into a single value (at-least-one binarization), enabling the full set of diagnostic codes specific to each patient to be stored as a sparse, binary array. The ICD10-CM codes were filtered according to multiple criteria, which are described in the Supplemental Methods. This generated a dataset containing the diagnostic status of 10,483 ICD10-CM codes aligned to 1,204,212 patients. This is subsequently referred to as the *UCSF-ICD10-CM* dataset.

The *UCSF-ICD10-CM* was further processed in two ways. First, the ICD10-CM codes were transformed into Human Phenotype Ontology (HPO)[30] terms using a customized mapping, the construction of which is outlined below and in the Supplemental Methods (resulting dataset denoted *UCSF-HPO*). This alignment resulted in a global diagnostic matrix encoding 1,674 HPO symptoms. Second, we translated the ICD10-CM codes into the ICD10 terminology utilized by the UK Biobank (ICD10-UKBB)[58], taking advantage of the fact that the UKBB encoding is a less granular subset of the ICD10-CM (details regarding the precise translation can be found within

our vLPI software package available on Github). This processed dataset is subsequently referred to as *UCSF-ICD10-UKBB*. The *UCSF-ICD10-UKBB* dataset was also translated into HPO terms (denoted *UCSF-HPO-UKBB*). These less granular datasets contained 4,933 and 1,423 diagnostic terms respectively.

The UK Biobank (UKBB) is a collection of ≈500,000 middle-aged British adults who have received extensive genotyping and phenotyping[26]. The bulk UKBB dataset was downloaded on Jan 22nd, 2020 using the software provided by the organization. Following download, the raw data file was parsed, isolating demographic variables of interest and collapsing main/secondary inpatient summary diagnoses into a single data value (using at-least-one binarization). The resulting diagnostic codes were filtered according to multiple criteria (see Supplementary Methods), resulting in a 1:1 correspondence between the diagnostic codes available within the UKBB and the *UCSF-ICD10-UKBB* datasets. These ICD10 codes were then translated into HPO terms. The full UKBB dataset (after removing withdrawn subjects; N=502,488) was used for cryptic phenotype inference, but the subjects were also filtered according to recommended best practices for genetic analyses[26,59]. Filtering resulted in the following two subsets: 1) 485,014 subjects (with exome data, N= 199,234) that remained after removing individuals whose genetic data is likely to be confounded by artefact (*UKBB-Full*), and 2) 342,796 unrelated subjects (with exome data, N=153,182) of likely Western European (Caucasian) ancestry (*UKBB-Unrelated*). Further details regarding the processing can be found in the Supplemental Methods.

Because the UCSF-CDW and UKBB were both used for phenotype model inference and evaluation, the datasets were *a priori* divided each into training and testing subsets. To ensure

that the testing datasets contained positive cases for each rare disease included in our analysis, distinct training and testing subsets were generated for every disorder. The subsets were constructed by randomly subsampling 75% of the data for training and 25% for testing while maintaining an equal ratio of diagnosed rare disease cases in each (see below). All model inference and preliminary analyses were performed using the training datasets, while the testing datasets were only used for the final evaluation of cryptic phenotypes (**Figures 2D** and **2E**).

Aligning rare diseases to structured medical data

Based on previous work[60–62,9,27], we integrated multiple biomedical ontologies and terminologies to map rare diseases and their symptoms to structured medical data (i.e. diagnostic billing codes). To generate a set of rare diseases for analysis, we first used the Human Disease Ontology[63] to obtain mappings between the Online Mendelian Inheritance in Man (OMIM) database[64] and the ICD10-CM terminologies. Building on previous work[61], we curated the OMIM-to-ICD10-CM alignments, selecting and grouping ICD10-CM codes that reliably mapped to a single or homogenous set of OMIM diseases, ensuring that the disorders were also annotated within the Human Phenotype Ontology[30]. This resulted in 166 rare, Mendelian conditions that were aligned to both the HPO and ICD10 terminologies (**Supplementary Figure 1**). The 166 diseases were sorted according to their diagnostic prevalence in the UCSF-CDW and the number of aligned HPO terms (**Supplemental Methods**); 50 disorders were selected for follow up testing (listed in **Supplemental Data File 1)**.

The HPO symptoms themselves were aligned to the ICD10-CM terminology in an automated fashion by integrating the information contained within multiple biomedical ontologies[65–68]. Details regarding the alignment are provided in the Supplemental Methods. This resulted in 1,674 unique alignments between HPO terms and ICD10-CM codes (**Supplemental Data File 2)**. We assessed their performance by using them as features in a rare disease diagnosis prediction task (**Supplemental Figure 2**). We found that prediction models constructed from the annotated[69], ICD10-CM-aligned HPO terms had performances that were similar to models constructed using the complete ICD10-CM codebook (see **Supplemental Table 1**).

Cryptic phenotype analysis

Cryptic phenotype analysis (CPA) refers to the process by which a set of symptoms is used to infer a univariate, latent trait that captures the clinical heterogeneity observed within a disease of interest. The quantitative but cryptic phenotype can be used to assess clinical variability in both the diagnosed cases and the more general population, enabling the types of analyses that are described above. CPA consists of two stages. In the first, the symptoms annotated to a particular disease are decomposed into a low-dimensional set of quantitative, latent phenotypes. In the second stage, the trait that best captures disease morbidity (i.e. its symptom expressivity) is identified, since multiple latent traits are often recovered from a single symptom matrix. Below, we briefly outline the two stages of CPA. A more detailed description is provided in the Supplementary Methods.

Latent Phenotype Inference

Consider the set of K symptoms that are associated with some rare disease of interest, and furthermore, assume that these symptoms are binary (present/absent) and permanent (i.e. once diagnosed, they do not resolve). Let $S_{i,j}$ denote the status of the jth symptom in the ith subject such that $S_{i,j} = 1$ indicates that the patient has been diagnosed with this symptom. Furthermore, let $\boldsymbol{S}$ denote an $N \times K$-dimensional matrix of symptom diagnoses such that the ith row of the matrix (denoted S_i) contains the diagnoses for subject i. Finally, let $\boldsymbol{Z}$ denote an $N \times L$-dimensional matrix of latent phenotypes, where each column represents the magnitude (i.e. severity) of an independent latent phenotype. We modeled the joint likelihood of the disease symptoms and latent phenotypes according to:

$$\mathrm{P}(\mathbf{S}, \mathbf{Z}|\theta) = \mathrm{f}(\mathbf{Z}; \theta) \times P(\mathbf{Z}),$$

where $f(\mathbf{Z}; \theta)$ is the symptom risk function (defined by the parameter set θ) that maps the latent phenotypes onto the matrix of symptom probabilities (i.e. $f(\boldsymbol{Z}; \theta) \in [0,1]^{N \times K} \equiv P(\boldsymbol{S}|\boldsymbol{Z}, \theta)$) and $P(\mathbf{Z})$ is a generative model for the latent phenotypes themselves. Additional details regarding $f(\boldsymbol{Z}; \theta)$ and $P(\mathbf{Z})$ are provided in the Supplemental Methods.

Given an observed symptom matrix (denoted $\boldsymbol{S} = \boldsymbol{s}$), we obtained estimates for the symptom risk function parameters (denoted $\hat{\theta}$) by optimizing a lower bound approximation to the model marginal likelihood (i.e. $P(\boldsymbol{s}|\theta) = \int P(\boldsymbol{s}, \boldsymbol{Z}|\theta) d\boldsymbol{Z}$) using an amortized, variational inference algorithm[32,33]. Model inference was conducted using the training subsets only. Estimates for the latent phenotypes of interest (denoted $\widehat{\boldsymbol{Z}}$) were obtained as a direct by-product of this optimization process (see **Supplemental Methods**). In practice, the observed symptom matrices for each rare disease were constructed from the *UCSF-HPO*, the *UKBB-HPO*,

and *the UCSF-HPO-UKBB* datasets using the annotations available on the HPO website (see **Supplemental Data File 3** for the complete disease-to-symptom mappings)[69]. However, some of the aligned symptoms were manually curated in attempt to resolve convergence issues (see Supplemental Methods); **Supplemental Data File 4** contains the final disease-to-symptom mappings used to infer the cryptic phenotypes for the 10 diseases that passed all our filters (see below). Additional details concerning our model inference and evaluation procedures are provided in the Supplemental Methods.

Cryptic Phenotype Identification and Evaluation

Following inference, we assigned each rare disease a single cryptic phenotype, which we define as the latent trait that best captures the symptom frequency intrinsic to the rare disease of interest (i.e. its morbidity). By default, all our models were initialized with a total of 10 possible latent phenotype components, as multiple pathologic processes can contribute to the correlation structure observed among some set of symptoms (see **Supplemental Methods** for more information). Although this meant that many of our models were initially overdetermined, we found that our inference algorithm was able to automatically remove unnecessary components by zeroing out their parameters in the symptom risk function. The number of latent components that remained following model inference was termed the model's *effective rank* (L_{eff}, see Supplementary Methods for precise definition), which was typically much less than the number of components used to initialize the model (**Supplementary Figure 4**). When $L_{\text{eff}} = 1$, then this single component was automatically selected to be the disease's cryptic phenotype. When $L_{\text{eff}} > 1$, then each inferred latent

phenotype was used separately as a classifier to predict rare disease diagnoses in the training dataset, noting that the component that best captures the expressivity of a disease should be most predictive of diagnostic status (see **Figures 3A-C, inset** for examples). This top-performing latent component (assessed using the average precision score[70]) was then selected as the disease's cryptic phenotype.

Note, the model fitting described above was completed in both the UCSF and UKBB datasets, with the caveat that not all the Mendelian diseases in **Supplemental Data File 1** map to specific ICD10 diagnostic codes in the UKBB dataset (the encoding for this dataset is more limited, see above). Therefore, cryptic phenotype models inferred in the UKBB dataset were replicated in the UCSF dataset (using *UCSF-HPO-UKBB*, see **Figure 2D** and **2E** for results). To ensure the assigned cryptic phenotypes were in fact capturing Mendelian-disease related morbidity, we compared the average cryptic phenotype severity among diagnosed cases to their undiagnosed controls (using the test datasets only). For a cryptic phenotype to replicate, the average symptom severity among Mendelian disease cases had to be significantly higher in both the UCSF and UKBB datasets (significance assessed through bootstrapped re-sampling[71] after performing Bonferroni corrections, see **Figure 2D** and **2E**). If Mendelian disease diagnostic codes were not available in the UKBB, then this increase in cryptic phenotype severity only needed to replicate in the UCSF dataset.

Beyond replication, we also wanted to ensure that models inferred within the two independent datasets were consistent, meaning that they generated similar results when applied to the same dataset. Therefore, the phenotype models inferred within the UKBB were directly applied to the *UCSF-HPO-UKBB* dataset. Consistency was then assessed in three ways.

First, the same latent component had to be assigned as the cryptic phenotype in both datasets (see above). Second, the UKBB model had to reproduce the increase in phenotype severity observed among the Mendelian disease cases within this new dataset. Third, the cryptic phenotypes produced by the UCSF and UKBB models needed to be highly correlated (as assessed through the coefficient of determination, R^2). Using an R^2 cutoff of 0.2, ten of the original fifty Mendelian disorders survived our replication and consistency filters. However, it is entirely plausible that replicable and consistent cryptic phenotypes could have been inferred for the other disorders through careful curation of annotated symptoms, larger sample sizes, and more focused adjustment of inference algorithm parameters (see **Supplemental Methods**).

Validation of the cryptic phenotypes using exome-sequencing data

The cryptic phenotypes for the five diseases listed in **Table 1** were further validated through rare variant association studies. This required identifying pathogenic variant carriers within the UKBB. For A1ATD, the causal Pi*Z allele (*rs28929474*) was directly ascertained through array-based genotyping, so carriers of the Pi*MZ and Pi*ZZ genotypes were identified in the call/imputation files (see UKBB Data Category 263). For the remaining diseases, we downloaded the VCF files that contained the known causal genes (see **Table 1**). We then used the ClinVar database VCF (available at https://ftp.ncbi.nlm.nih.gov/pub/clinvar/) to isolate all variants annotated as pathogenic/likely pathogenic (accomplished using `bcftools`[72]). Because heterozygous loss-of-function (LoF) is an established molecular mechanism for each of diseases in **Table 1** (except for A1ATD), we also identified LoF variants that were not listed in ClinVar, which were annotated using the LOFTEE plugin[34] for the Ensembl Variant Effect

Predictor[73]. Not all the variants isolated in this manner have equivalent levels of evidence for pathogenicity. Therefore, we added a flag to each variant to indicate if: 1) it had conflicting annotations, 2) it was annotated by a single submitter, or 3) it was located within a non-canonical transcript (LoF variants only). **Supplemental Data File 7** contains a complete list of the P/LP variants analyzed in this study.

Using the VCF and genotype call files, we then identified all carriers of the P/LP variants described above. To assess whether the variants were associated with cryptic phenotype severity, we estimated their average genetic effects using the following linear model:

$$CP_i = \beta_0 + \beta_{P/LP} \times G_i + \vec{a} \times \vec{X}^T,$$

where CP_i denotes the cryptic phenotype of the *i*th subject, β_0 is an intercept parameter, $\beta_{P/LP}$ is the average effect parameter for the P/LP variants, G_i is the carrier status of the *i*th patient, and $\vec{X}^T$ denotes a vector of covariates (with their corresponding parameter vector given by $\vec{a}$). Sex, age (inverse rank-transformed to remove skew), UKBB array platform, and the first 10 principal components of the genetic relatedness matrix were used as covariates. The analysis was limited to unrelated individuals of similar ancestry (Caucasian) to reduce the risk for population structure confounding (N=153,182). Estimates for the parameters were produced using ordinary least squares, and per-parameter significance was assessed using a two-sided T-test. To account for effects related to variant annotation, we also fit the following linear model:

$$CP_i = \beta_0 + \beta_{P/LP} \times G_i + \beta_{Unflagged\ P/LP} \times G_i \times U_i + \vec{a} \times \vec{X}^T,$$

where U_i is a binary variable that indicates if the *i*th patient carries a variant without any annotation flags (see above). This enabled us to decompose the phenotypic contributions of

P/LP variants into baseline ($\beta_{P/LP}$) and unflagged ($\beta_{Unflagged\ P/LP}$) effects, which are displayed at the top of the panels in **Figure 3A-C** and **Supplemental Figure 6B**. Note, the AS phenotype is known to be more severe among hemizygous male carriers of *COL4A5* pathogenic variants, consistent with X-linked inheritance. Therefore, we included an interaction term between sex and *COL4A5* carrier status during our molecular validation of the AS cryptic phenotype. This interaction effect did not reach statistical significance ($\beta_{COL4A5xSex}$=0.09$\pm$0.15; *P*-value=0.56), likely due to the small number of male P/LP *COL4A5* carriers in the dataset (N=12 in *UKBB-Unrelated*). As a result, sex-specific interaction effects were not included in downstream analyses.

Common variant genome-wide association analyses for the inferred cryptic phenotypes.

Genome-wide association studies were performed to identify common genetic variants associated with the cryptic phenotypes assigned to the diseases in **Table 1.** To reduce the risk of confounding, the association analyses were conducted using a subset of patients isolated from *UKBB-Unrelated* ($N = 342{,}796$) that met the following criteria: 1) did not possess a P/LP variant in a gene linked to the disease of interest, 2) were never diagnosed with this disease, and 3) were not a 3rd degree or closer relative of any of these subjects. From this training cohort, a random subset of 10% were removed and added to the Mendelian disease P/LP carriers. This second dataset is called the target cohort, and it was used to perform polygenic score replication and validation. All SNPs meeting the following criteria were in included into the analyses: directly genotyped by the UKBB, minor allele frequency (MAF) $\geq$ 1%, missing genotype fraction $\leq$ 5%, and Hardy-Weinberg Equilibrium (HWE) P-value $\geq 10^{-12}$. Note, a

relatively limited number of genetic markers (579,429 SNPs) met these criteria, but this smaller set of features enabled us construct individual-level prediction models for polygenic score inference (see below).

Genome-wide association studies (GWAS) were conducted by fitting the following linear model to each cryptic phenotype:

$$CP_i = \beta_0 + \beta_j^{SNP} \times G_{i,j} + \vec{a} \times \vec{X}^T,$$

where CP_i indicates the cryptic phenotype in the *i*th patient, β_j^{SNP} represents the average effect of the *j*th SNP, $G_{i,j}$ encodes the minor allele count ($G_i \in \{0,1,2\}$; additive model), and $\vec{X}^T/\vec{a}$ denote covariates/effect parameters respectively. Sex, age (rank-normalized), UKBB array platform, and the first 10 principal components of the genetic relatedness matrix were used as covariates. Association statistics were estimated using the `Plink2`[75] software package (`--glm` command). Lead SNPs and their corresponding annotations were generated using the FUMA[40] platform. The loci identified for the three diseases with genome-wide significant effects are provided as **Supplemental Tables 2**, **3**, and **4** (A1ATD, AS, and ADPKD respectively)**.**

SumHer (available within the LDAK toolkit)[76] was used to produce estimates for the fraction of the additive variance explained by the genotyped SNPs (narrow-sense heritability, denoted h^2). This required the specification of an underlying heritability model[76]. Based on recommended best-practices, we used the LDAK-Thin model given its simplicity and portability to individual-level prediction. This required computing a tagging file, which was constructed using a random subset (N=10,000) of *UKBB-Unrelated*. First, duplicate SNPs were identified using the `ldak --thin` command with the following options: `--window-prune .98 --window-kb 100`. Next, the tagging file itself was constructed using the `ldak --calc-`

`tagging` command (with options `--power -.25 --window-cm 1 --save-matrix YES`). Finally, narrow-sense heritability estimates were produced from the GWAS summary statistics using the `ldak --sum-hers` command (while also storing the per-SNP heritability estimates for downstream analyses).

Polygenic scores summarizing the common variant association statistics were computed for the three diseases in **Table 1** that had cryptic phenotype h^2 estimates significantly greater than 0 (A1ATD, AS, and ADPKD). These scores were estimated using a prediction model that was inferred from the individual-level genotype data available within each training cohort. More specifically, we used LDAK-Bolt-Predict[35] (`ldak --bolt` command) to estimate effect sizes for every SNP included in the cryptic phenotype association analyses (while conditioning on the covariates included in the initial linear model, see above). This required access to the per-SNP heritability estimates, which were produced by the SumHer model (see above). Note, 10% of the training data was withheld during model inference (using the `--cv-proportion .1` flag) to estimate prior parameters. After model fitting was complete, polygenic scores were imputed into the target cohort using the `--calc-scores` command (with `--power` flag set to 0). The per-SNP effect size estimates produced by the predictor models are included as **Supplemental Data Files 8**, **9**, and **10** (A1ATD, AS, and ADPKD respectively).

Estimating the effects of polygenic load on Mendelian disease severity and outcomes.

Polygenic scores (PGS) were imputed into the target cohorts for each rare disease in order to: 1) replicate the PGS-cryptic phenotype relationships, 2) assess for interaction effects

between the PGS and P/LP variants, and 3) determine if high polygenic load was associated with established Mendelian disease outcomes.

The first two analyses were accomplished by fitting the following linear model within the target cohort of each cryptic phenotype:

$$CP_i = \beta_0 + \beta_{PGS} \times \xi_i + \beta_{P\backslash LP} \times G_i + \beta_{PGSxP\backslash LP} \times G_i \times \xi_i + \vec{a} \times \vec{X}^T,$$

where ξ_i represents the PGS for the *i*th patient, β_{PGS} represents its average phenotypic effect, and $\beta_{PGSxP\backslash LP} \times G_i \times \xi_i$ models the interaction between the PGS and the P/LP variants. In the case of A1ATD, the two pathogenic genotypes (Pi*ZZ and Pi*MZ) were modeled as separate genetic effects, each with their own PGS interaction terms. For AS, both flagged and unflagged P/LP variants were included into the analysis, as they were both shown to influence cryptic phenotype severity (see **Supplemental Figure 6B**). For ADPKD, only the unflagged variants were included, as there was no detectable phenotypic effect for the flagged variants, suggesting that they most likely represent annotation noise (see **Figure 3C**). The previous model was fit using ordinary least squares, and association statistics were computed using a two-sided T-test.

Regarding covariates (i.e. $\vec{a} \times \vec{X}^T$), sex, age, UKBB array platform, and the first 10 principal components of the genetic relatedness matrix were included in every model. Smoking history was also included into each model, although its incorporation varied across diseases. For A1ATD and AS, self-reported pack-years (defined by Data Field: 20161) were used to quantify smoking history. Note, there was a significant interaction effect between the Pi*Z allele and smoking history (as expected), so interaction terms between pack-years and the pathogenic genotypes were included into the cryptic phenotype model for this disease (see **Figure 5B**). There was no significant interaction effect between pack-years and P/LP carrier

status for AS ($\beta_{\text{Pack-years x P/LP}}$=0.03±0.05; *P*-value=0.55), so smoking interaction terms were not included for this disorder.

Regarding ADPKD, smoking had a strong protective effect on cryptic phenotype severity such that those P/LP carriers with a history of ever-smoking (provided by Data Field: 20160) had systematically lower cryptic phenotype scores ($\beta_{\text{Smoke x P/LP}}$=-0.41 ±0.11; *P*-value=$1.2 \times 10^{-4}$). This result is clearly at odds with the known pathophysiology of smoking and renal disease, and it likely stems from the fact that subjects with moderate-to-severe ADPKD are often diagnosed at a young age, prior to when smoking behavior is established (see **Supplemental Figure 9D** for Kaplan-Meier curve of ESRD among P/LP carriers). Consistent with this hypothesis, significantly fewer P/LP carriers reported ever-smoking when compared to the general population (see **Supplemental Figure 9B**). Based on these results, the relationship between smoking history and ADPKD severity is likely to be confounded by multiple unmeasured factors (specifically, medical intervention and counseling). Given our inability to adequately adjust for such complex confounding, smoking history in ADPKD was modeled using a simple binary variable (UKBB Data Field: 20160), which was included along with a P/LP interaction term. Note, similar confounding likely plays a role in the interaction effects between smoking and genotype for the other disorders (see **Figure 5E** and **Supplemental Figure 7B** for examples), but it was only significant enough to reverse the established morbidity relationship for ADPKD.

To confirm a role for polygenic load on Mendelian disease outcomes, we examined its effect on quantitative measurements that capture established pathophysiology but are distinct from symptoms used to construct the cryptic phenotype. For A1ATD, we used the FEV1/FVC ratio (UKBB Data Field: 20258), a measurement derived from spirometry that quantifies the

severity of obstructive lung disease (see **Supplemental Figure 7C**). For AS, we examined urine microalbumin level (UKBB Data Field: 30500), which correlates with renal health and glomerular barrier function (see **Supplemental Figure 8C**). Finally, for ADPKD, we computed an estimate[49] of the glomerular filtration rate (eGFR) from serum creatinine level (UKBB Data Field: 30700), which is often used as a proxy for overall renal function (see **Supplemental Figure 8C**)**.** The models themselves incorporated the same genetic and covariate effects that were used for the cryptic phenotype models, and they were again fit using ordinary least squares with association statistics computed from a two-sided T-test.

Finally, the effect of polygenic load on Mendelian disease severity was assessed by estimating its association with: 1) the rare disease diagnosis itself (when available) and 2) the onset of clinically important outcomes. The effect of the PGS on Mendelian disease diagnostic risk was modeled using logistic regression according to:

$$\text{Log-Odds}(D_i) = \beta_0 + \beta_{PGS} \times \xi_i + \beta_{P\backslash \text{LP}} \times G_i + \beta_{PG\text{SxP}\backslash \text{LP}} \times G_i \times \xi_i + \vec{a} \times \vec{X}^T,$$

where D_i is in a binary variable indicating whether a disease diagnosis is present or absent. The covariates included were sex, age, UKBB array platform, the first 10 principal components of the genetic relatedness matrix, and smoking history (plus interaction terms when relevant, see above). Model fitting was performed using the maximum-likelihood method with a Firth penalty term, which was included given the risk for Type I error rate inflation in the setting of unbalanced samples and rare predictors[77]. Significance for a given association was assessed using a likelihood-ratio test[78].

The age-of-onset for clinically important Mendelian disease outcomes was also used to assess the effects of polygenic load on disease severity. The outcomes included in this study

were: End-Stage Renal Disease (ESRD; UKBB Data Field: 42026), Chronic Obstructive Pulmonary Disease (COPD; UKBB Data Field: 42016), Recurrent and Persistent Hematuria (UKBB Data Field: 132002), and Cystic Kidney Disease (UKBB Data Field: 132532). Details concerning the construction of these data fields are available through the UKBB. For each outcome, age-of-onset was modeled using Cox Proportional Hazard (CPH) regression:

$$\lambda_i = \beta_{PGS} \times \xi_i + \beta_{P\backslash \mathrm{LP}} \times G_i + \beta_{PG\mathrm{SxP}\backslash \mathrm{LP}} \times G_i \times \xi_i + \vec{a} \times \vec{X}^T,$$

where λ_i represents the logarithm of the partial hazard function for the *i*th subject. The following covariates were included into the model: sex, UKBB array platform, the first 10 principal components of the genetic relatedness matrix, and smoking history (with interaction terms as indicated). Model fitting was performed by maximizing the partial likelihood (using the `lifelines` software package[79]), and significance was assessed using a likelihood-ratio test.

Dataset Availability

The clinical and genetic datasets used in the analyses presented in this manuscript cannot be shared directly with third parties, as both have specific provisions against open data sharing outside of their usual application processes. Information regarding third party access to the UCSF De-Identified Clinical Data Warehouse can be found through UCSF Data Resources: https://data.ucsf.edu/cdrp/research, and the application process for access to the UK Biobank is outlined on their website: https://www.ukbiobank.ac.uk/register-apply. Datasets that were generated to conduct the analyses described in this manuscript are provided as Supplementary Data Files 1-10.

Code Availability

We have deposited the software developed in this study. Latent phenotype model inference was performed using the vLPI software package, which was specifically constructed for the analyses presented in this manuscript. It is available via Github: https://github.com/daverblair/vlpi. A software package that automatically imputes the cryptic phenotypes analyzed in this study using ICD10-CM/ICD10-UKBB codes is available on Github: https://github.com/daverblair/CrypticPhenoImpute. A singularity container with this software already installed can be constructed using the following container script: https://github.com/daverblair/singularity_vlpi. `Plink2`[75], the `LDAK Toolkit`[35,76], and `lifelines`[79] are all freely available from their respective websites.

Human Research Subject Participation

This study used de-identified human genetic and clinical information (UCSF IRB #: 19-29458). It qualified for Exempt status.

Acknowledgements

This work was supported by the Stimulating Access to Research in Residency (StARR) program at UCSF (NHLBI grant 5R38HL143581-03; PI Alison Huang), which provided funding for the first author and data access. The first author was also supported by the UCSF Pediatrics and Medical Genetics Residency Programs. We thank members of the Shieh lab for their feedback given throughout, and we are extremely grateful to the UK Biobank and its participants (Application 53312). We also want to thank Academic Research Systems at UCSF for providing access to the

UCSF Clinical Data Warehouse, and the Wynton High Performance Computing team for their maintenance of the computational resources used in this study.

Contributions

DRB and JTS conceived of the project. DRB designed the study, implemented the analyses, interpreted the results, made the figures, and wrote the manuscript with input from TJH and JTS.

References

1. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **97**, 199–215 (2015).
2. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
3. Wright, C. F., FitzPatrick, D. R. & Firth, H. V. Paediatric genomics: diagnosing rare disease in children. *Nat. Rev. Genet.* **19**, 253–268 (2018).
4. Posey, J. E. *et al.* Insights into genetics, human biology and disease gleaned from family based genomic studies. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **21**, 798–812 (2019).
5. Wenger, B. M. *et al.* A genotype-first approach to exploring Mendelian cardiovascular traits with clear external manifestations. *Genet. Med.* 1–9 (2020) doi:10.1038/s41436-020-00973-2.
6. Akhurst, R. J. TGFβ signaling in health and disease. *Nat. Genet.* **36**, 790–792 (2004).
7. Chen, R. *et al.* Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat. Biotechnol.* **34**, 531–538 (2016).
8. Tarailo-Graovac, M., Zhu, J. Y. A., Matthews, A., van Karnebeek, C. D. M. & Wasserman, W. W. Assessment of the ExAC data set for the presence of individuals with pathogenic genotypes implicated in severe Mendelian pediatric disorders. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **19**, 1300–1308 (2017).
9. Bastarache, L. *et al.* Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* **359**, 1233–1239 (2018).

10. Rahit, K. M. T. H. & Tarailo-Graovac, M. Genetic Modifiers and Rare Mendelian Disease. *Genes* **11**, (2020).
11. Grange, T. *et al.* Quantifying the Genetic Basis of Marfan Syndrome Clinical Variability. *Genes* **11**, (2020).
12. Corvol, H. *et al.* Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nat. Commun.* **6**, (2015).
13. Pottier, C. *et al.* Potential genetic modifiers of disease risk and age at onset in patients with frontotemporal lobar degeneration and GRN mutations: a genome-wide association study. *Lancet Neurol.* **17**, 548–558 (2018).
14. Bonyadi, M. *et al.* Mapping of a major genetic modifier of embryonic lethality in TGF beta 1 knockout mice. *Nat. Genet.* **15**, 207–211 (1997).
15. Zhang, S., Binari, R., Zhou, R. & Perrimon, N. A Genomewide RNA Interference Screen for Modifiers of Aggregates Formation by Mutant Huntingtin in Drosophila. *Genetics* **184**, 1165–1179 (2010).
16. Aubart, M. *et al.* Association of modifiers and other genetic factors explain Marfan syndrome clinical variability. *Eur. J. Hum. Genet.* **26**, 1759–1772 (2018).
17. Pemov, A. *et al.* Genetic Modifiers of Neurofibromatosis Type 1-Associated Café-au-Lait Macule Count Identified Using Multi-platform Analysis. *PLoS Genet.* **10**, (2014).
18. Sturm, A. C. *et al.* Clinical Genetic Testing for Familial Hypercholesterolemia: JACC Scientific Expert Panel. *J. Am. Coll. Cardiol.* **72**, 662–680 (2018).

19. Hindorff, L. A., Gillanders, E. M. & Manolio, T. A. Genetic architecture of cancer and other complex diseases: lessons learned and future directions. *Carcinogenesis* **32**, 945–954 (2011).

20. Ingles, J. & Semsarian, C. Time to Rethink the Genetic Architecture of Long QT Syndrome. *Circulation* **141**, 440–443 (2020).

21. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).

22. Fahed, A. C. *et al.* Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat. Commun.* **11**, 3635 (2020).

23. Mars, N. *et al.* The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nat. Commun.* **11**, 6383 (2020).

24. Nauffal, V. *et al. Monogenic and Polygenic Contributions to QTc Prolongation in the Population*. http://medrxiv.org/lookup/doi/10.1101/2021.06.18.21258578 (2021) doi:10.1101/2021.06.18.21258578.

25. De-Identified Clinical Data Warehouse | Academic Research Systems. https://myresearch.ucsf.edu/de-identified-clinical-data-warehouse.

26. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

27. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med. Inform.* **7**, e14325 (2019).

28. Bastarache, L. *et al.* Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. *J. Am. Med. Inform. Assoc.* **26**, 1437–1447 (2019).

29. ICD - ICD-10-CM - International Classification of Diseases, Tenth Revision, Clinical Modification. https://www.cdc.gov/nchs/icd/icd10cm.htm (2020).

30. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/47/D1/D1018/5198478.

31. Stein-O'Brien, G. L. *et al.* Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends Genet.* **34**, 790–805 (2018).

32. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *ArXiv13126114 Cs Stat* (2014).

33. Kingma, D. P. & Welling, M. An Introduction to Variational Autoencoders. *Found. Trends® Mach. Learn.* **12**, 307–392 (2019).

34. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

35. Zhang, Q., Privé, F., Vilhjálmsson, B. & Speed, D. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat. Commun.* **12**, 4192 (2021).

36. Stoller, J. K., Hupertz, V. & Aboussouan, L. S. Alpha-1 Antitrypsin Deficiency. in *GeneReviews®* (eds. Adam, M. P. et al.) (University of Washington, Seattle, 1993).

37. Serres, F. de & Blanco, I. Role of alpha-1 antitrypsin in human health and disease. *J. Intern. Med.* **276**, 311–335 (2014).

38. Nakanishi, T. *et al.* The undiagnosed disease burden associated with alpha-1 antitrypsin deficiency genotypes. *Eur. Respir. J.* **56**, 2001441 (2020).

39. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

40. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

41. Warady, B. A. *et al.* Alport Syndrome Classification and Management. *Kidney Med.* **2**, 639–649 (2020).

42. Harris, P. C. & Torres, V. E. Polycystic kidney disease. *Annu. Rev. Med.* **60**, 321–337 (2009).

43. Quinlan, C. & Rheault, M. N. Genetic Basis of Type IV Collagen Disorders of the Kidney. *Clin. J. Am. Soc. Nephrol.* **16**, 1101–1109 (2021).

44. Zehnder, A. F. *et al.* Distribution of Type IV Collagen in the Cochlea in Alport Syndrome. *Arch. Otolaryngol. Neck Surg.* **131**, 1007–1013 (2005).

45. Savige, J. *et al.* Ocular Features in Alport Syndrome: Pathogenesis and Clinical Significance. *Clin. J. Am. Soc. Nephrol.* **10**, 703–709 (2015).

46. Savige, J. *et al.* Expert Guidelines for the Management of Alport Syndrome and Thin Basement Membrane Nephropathy. *J. Am. Soc. Nephrol.* **24**, 364–375 (2013).

47. Cornec-Le Gall, E., Torres, V. E. & Harris, P. C. Genetic Complexity of Autosomal Dominant Polycystic Kidney and Liver Diseases. *J. Am. Soc. Nephrol. JASN* **29**, 13–23 (2018).

48. Benonisdottir, S. *et al.* Sequence variants associating with urinary biomarkers. *Hum. Mol. Genet.* **28**, 1199–1211 (2019).

49. Levey, A. S. *et al.* A new equation to estimate glomerular filtration rate. *Ann. Intern. Med.* **150**, 604–612 (2009).

50. Thangaraj, P. M. & Tatonetti, N. P. Medical data and machine learning improve power of stroke genome-wide association studies. *bioRxiv* 2020.01.22.915397 (2020) doi:10.1101/2020.01.22.915397.

51. McCarty, C. A. *et al.* The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics* **4**, 13 (2011).

52. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).

53. Carey, D. J. *et al.* The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **18**, 906–913 (2016).

54. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).

55. Locke, A. E. *et al.* Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature* **572**, 323–328 (2019).

56. Cortes, A. *et al.* Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. *Nat. Genet.* **49**, 1311–1318 (2017).

57. Zhao, J. *et al.* Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of Lipoprotein(a) (LPA). *PLOS ONE* **14**, e0212112 (2019).

58. UKB : Data-Coding 19. https://biobank.ndph.ox.ac.uk/showcase/coding.cgi?id=19.

59. Hartman, K. A., Rashkin, S. R., Witte, J. S. & Hernandez, R. D. Imputed Genomic Data Reveals a Moderate Effect of Low Frequency Variants to the Heritability of Complex Human Traits. *bioRxiv* 2019.12.18.879916 (2019) doi:10.1101/2019.12.18.879916.

60. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinforma. Oxf. Engl.* **26**, 1205–1210 (2010).

61. Blair, D. R. *et al.* A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. *Cell* **155**, 70–80 (2013).

62. Wei, W.-Q. *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS ONE* **12**, (2017).

63. Schriml, L. M. *et al.* Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.* **47**, D955–D962 (2019).

64. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).

65. Whetzel, P. L. *et al.* BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.* **39**, W541–W545 (2011).

66. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267-270 (2004).

67. Stearns, M. Q., Price, C., Spackman, K. A. & Wang, A. Y. SNOMED clinical terms: overview of the development process and project status. *Proc. AMIA Annu. Symp. AMIA Symp.* 662–666 (2001).

68. Dhombres, F. & Bodenreider, O. Interoperability between phenotypes in research and healthcare terminologies—Investigating partial mappings between HPO and SNOMED CT. *J. Biomed. Semant.* **7**, (2016).

69. Human Phenotype Ontology. https://hpo.jax.org/app/download/annotation.

70. sklearn.metrics.average_precision_score — scikit-learn 0.23.1 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html.

71. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*. (Chapman and Hall/CRC, 1993).

72. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

73. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

74. Goh, L. & Yap, V. B. Effects of normalization on quantitative traits in association test. *BMC Bioinformatics* **10**, 415 (2009).

75. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

76. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* **51**, 277–284 (2019).

77. Wang, X. Firth logistic regression for rare variant association tests. *Front. Genet.* **0**, (2014).

78. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).

79. Davidson-Pilon, C. *et al. CamDavidsonPilon/lifelines: 0.26.0*. (Zenodo, 2021). doi:10.5281/zenodo.4816284.