Published in final edited form as:
Science. 2018 March 16; 359(6381): 1233–1239. doi:10.1126/science.aal4043.

Phenotype risk scores identify patients with unrecognized Mendelian disease patterns

Lisa Bastarache[1], **Jacob J. Hughey**[1], **Scott Hebbring**[2], **Joy Marlo**[1], **Wanke Zhao**[3], **Wanting T. Ho**[3], **Sara L. Van Driest**[4,5], **Tracy L. McGregor**[5], **Jonathan D. Mosley**[4], **Quinn S. Wells**[4,6], **Michael Temple**[1], **Andrea H. Ramirez**[4], **Robert Carroll**[1], **Travis Osterman**[1,4], **Todd Edwards**[4], **Douglas Ruderfer**[4], **Digna R. Velez Edwards**[7], **Rizwan Hamid**[5], **Joy Cogan**[5], **Andrew Glazer**[4], **Wei-Qi Wei**[1], **QiPing Feng**[6], **Murray Brilliant**[2], **Zhizhuang J. Zhao**[3], **Nancy J. Cox**[4], **Dan M. Roden**[1,4,6], and **Joshua C. Denny**[*,1,4]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA

[2]Center for Human Genetics, Marshfield Clinic Research Institute, Marshfield, WI, USA

[3]Department of Pathology, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

[4]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

[5]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA

[6]Department of Pharmacology, Vanderbilt University Medical Center, Nashville, TN, USA

[7]Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, TN, USA

Abstract

Genetic association studies often examine features independently, potentially missing subpopulations with multiple phenotypes that share a single cause. We describe an approach that aggregates phenotypes based on patterns described by Mendelian diseases. We mapped the clinical features of 1,204 Mendelian diseases into phenotypes captured from the electronic health record (EHR) and summarized this evidence as phenotype risk scores (PheRS). In an initial validation, PheRS distinguished cases and controls of five Mendelian diseases. Applying PheRS to 21,701 genotyped individuals uncovered 18 associations with rare variants and phenotypes consistent with Mendelian diseases. In 16 patients, the rare genetic variants were associated with severe outcomes such as organ transplants. PheRS can augment rare variant interpretation and may identify subsets of patients with distinct genetic causes for common diseases.

Classically, Mendelian diseases are thought to be rare, caused by variants with large effect sizes, and associated with significant morbidity and mortality. Many are characterized by a

[*]Correspondence to: Josh Denny, josh.denny@vanderbilt.edu.

Supplementary Materials:
Figs. S1 to S20
Tables S1 to S17
References (33-44)

range of clinical phenotypes, often affecting multiple organ systems. Several lines of evidence suggest that genes known to cause Mendelian disease also harbor variants that contribute to complex disease.(1) Studies have identified clinical overlap in patients co-diagnosed with Mendelian and complex disease(2), and SNPs found in genome-wide association studies (GWAS) are enriched for Mendelian loci.(3) A review of evidence from GWAS and whole exome sequencing studies found a striking overlap between primary immunodeficiency genes and complex inflammatory diseases.(4) Collectively, this evidence suggests that variants in Mendelian disease-causing genes may be an under-recognized contributor to complex disease.

Until very recently, the phenotypic effects of rare genetic variants were ascertained primarily in family-based studies of patients with distinctive and often severe phenotypes. Population-level techniques, such as GWAS and phenome-wide association studies (PheWAS)(5, 6), are not easily applied to rare variation since most studies are underpowered. Cohorts large enough to support GWAS of rare variants have only recently been assembled and have demonstrated the potential impact of rare variants on complex traits such as height, finding rare variants with effect sizes much greater than those of common variants.(7)

Estimating the pathogenicity of rare variants remains a challenge and a barrier to use in the clinical setting.(8) Many algorithms have been developed to predict variant pathogenicity,(9–11) and consortia such as ClinGen(12) are aggregating knowledge to enable expert determinations. Resources such as ExAC(13) have helped refine variant interpretation. Some variants previously interpreted as pathogenic are too common in some populations to cause rare, life-threatening disorders,(14) while others thought to be completely penetrant do not always cause disease.(15) Initial studies suggest that electronic health records (EHRs) linked to genetic data may help drive genomic discovery and define clinical phenotypes associated with rare variants.(16–18)

We have developed an approach that increases the power to detect rare variant associations by leveraging the phenotypic patterns of Mendelian diseases. By mapping the clinical manifestations of a Mendelian disease to phenotypes extracted from the EHR, we can compute a “phenotype risk score” (PheRS) that expresses the degree to which an individual’s symptoms overlap with a Mendelian disease. We defined a PheRS as a weighted aggregation of genetically-related phenotypes, analogous to the genetic risk score approach for analyzing multiple variants against a single phenotype. PheRS was validated against clinically diagnosed cases and controls, and a genetic association study of PheRS profiles for 1,204 Mendelian conditions identified both known and novel associations with variants in target genes. The approach presents a method for measuring the phenotypic impact of rare variants and for identifying the heretofore under-recognized contribution of Mendelian disease genes to common medical conditions.

Constructing a phenotype risk score

The Online Mendelian Inheritance in Man (OMIM) provides clinical synopses for thousands of monogenic diseases(19) which have been annotated using the Human Phenotype Ontology (HPO).(20) We created a map from HPO to consolidated billing codes from the

EHR called phecodes. Phecodes enable high-throughput ascertainment of EHR phenotypes, and have been widely used to replicate known genetic associations and discover new ones.(21–23) By mapping HPO terms to phecodes, we can express “phenotype syndromes” patterned after Mendelian diseases in OMIM in terms of clinical phenotypes that can be rapidly derived from the EHR. The PheRS for a given Mendelian disease is defined as the sum of clinical features observed in a given subject weighted by the log inverse prevalence of the feature.

Validating PheRS

We computed the PheRS of clinically diagnosed cases to matched controls for six Mendelian diseases. PheRS was a very strong predictor of case status for five of the diseases (Wilcoxon rank-sum test; $p=8\times10^{-42}$ to 5×10^{-320}) (Fig. 1A and 1B). The exception was phenylketonuria (p=0.28), which effectively served as a negative control since newborn screening and dietary avoidance of phenylalanine essentially eliminates disease manifestations in affected individuals.(24) The PheRS for each Mendelian disease demonstrated specificity for the target disease, as the cases for different Mendelian diseases had similar PheRS distributions as controls (Fig. 1C). The lone exception was that the PheRS for hemochromatosis (HH) was significantly elevated for cystic fibrosis (CF) cases versus controls. However, even in this instance, CF cases had 3-fold higher PheRS for CF compared to HH. A review of controls with a PheRS greater than the 75th percentile identified one individual (PheRS >99th percentile) who was diagnosed with HH in the six months following the case/control ascertainment. Thus, for this individual, the PheRS suggested the diagnosis before it was made by providers.

PheRS identifies potentially pathogenic variants in Mendelian disease genes

We conducted an association analysis based on a cohort of 21,701 adults of European ancestry genotyped on the Exome BeadChip (Table S1). In this cohort, we computed PheRS for 1,204 Mendelian diseases (1,096 causative genes) for which we had sufficient genotype data. We tested for association between PheRS and 6,188 rare variants (minor allele frequency [MAF] < 1%) using linear regression, assuming a dominant genetic model. We only tested the PheRS for a particular Mendelian disease against variants in the gene or genes known to cause that disease. We found 18 significant associations between rare variants and PheRS (q<0.1; Table 1). All significant results had a positive beta coefficient, indicating the variants were associated with an excess of Mendelian disease phenotypes. Four of the genes had an established dominant mode of inheritance, while the remaining 13 genes were known as exclusively or primarily recessive. Four were annotated in ClinVar as “pathogenic” or “likely pathogenic,” and the Human Gene Mutation Database (HGMD) provided evidence of pathogenicity for an additional three variants.(25) The phenotypic impact of the remaining nine variants have not, to our knowledge, been previously described.

Clinical chart review revealed that eight of the 807 individuals with statistically significant variants were diagnosed with the target Mendelian disease. Seven individuals with one of the two *CFTR* variants (p.G542* and p.R553*) were diagnosed with CF. Clinical genetic testing confirmed the variants called on the Exome BeadChip, including one homozygote for p.G542*, and established compound heterozygosity with ΔF508 for five others (Fig. 2A and Table S2). All individuals diagnosed with CF had a PheRS greater than four standard deviations from expected values. Additionally, the highest scoring heterozygote for p.E168Q in *HFE* was diagnosed with hemochromatosis on the basis of clinical findings. The diagnosis of HH was considered but never confirmed for another p.E168Q heterozygote who died of end-stage liver disease.

While the majority of patients with significant variants were undiagnosed, these individuals had a high burden of severe endpoints related to the Mendelian diseases. Of the 40 heterozygotes for *HFE* p.E168Q, four had liver transplants (10% versus 1.2% in background; $p=2.1\times10^{-3}$; Fisher's exact). Individuals with variants in two genes associated with renal failure had elevated rates of kidney transplant: five of 36 (14%) patients with *AGXT* p.A295T were transplanted (another is awaiting transplant), as well as two of 15 (13%) patients with *DGKE* p.W322* (versus 3% in background; $p=6.9\times10^{-3}$ and p=0.088, respectively; Fisher's exact). Four of 69 *TG* p.G77S heterozygotes underwent thyroidectomies (6% versus 2% of non-carriers; p=0.039; Fisher's exact). These are end-stage phenotypes, potentially resulting from the effect of these variants, and did not have an increased prevalence in other significant variant carriers (Table S3). Additionally, we found the population attributable fraction for the constituent PheRS phenotypes averaged 0.5% with a maximum of 4.5%, suggesting that common diseases in adult populations may, in some cases, be attributed to variants in Mendelian genes (Fig. S1).

An examination of PheWAS results using Fisher's exact for variants identified in the discovery analysis revealed that constituent phenotypes were often marginally significant ($p<0.05$), while not crossing the Bonferroni correction level for a single PheWAS. For *CFTR* p.G542*, three features used in the PheRS for CF achieved marginal significance (bronchiectasis, disease of the pancreas, and chronic airway obstruction) (Fig. 2B). However, the constituent phenotypes for CF were only statistically significant when they were analyzed collectively as a PheRS. The association with p.G542* and the PheRS for CF was similar to the association with the phenotype of CF itself (PheRS by linear regression $p=3\times10^{-8}$ vs. CF diagnosis by Fisher's exact 8×10^{-7}). Similarly, while individuals with variant p.W322* in *DGKE* had an excess of nephrotic syndrome features (Fig. 2C), these phenotypes were not significantly associated on their own in the PheWAS analysis (Fig. 2D). A similar pattern was observed for the remaining variants identified in the discovery analysis (Figs. S2–17). A PheWAS analysis of all 6,188 variants tested in the discovery analysis and 1,734 phecodes did not yield any significant associations $q<0.1$.

Replication of novel associations

We attempted to replicate significant associations from the discovery analysis in two independent cohorts: a European ancestry cohort from Marshfield Clinic (n=9,441) and a non-European ancestry cohort from Vanderbilt (n=3,820; Tables S4–5). Each was tested as in the discovery cohort, using linear regression assuming a dominant model, adjusting for age and sex. Only variants with at least 10 heterozygotes or homozygotes for the rare allele were tested. In the Marshfield cohort, both attempted associations replicated: p.G77S in *TG* with thyroid dyshormonogenesis PheRS ($p=5.0\times10^{-4}$) and p.R507H in *FAN1* with karyomegalic interstitial nephritis PheRS ($p=8.2\times10^{-3}$, Table S6). In the Vanderbilt non-European ancestry cohort, we replicated two of three associations: p.A993A in *KIF1A* with spastic paraplegia PheRS ($p=1.9\times10^{-3}$), and p.A295T in *AGXT* with primary hyperoxaluria type 1 PheRS ($p=3.9\times10^{-3}$). The association between p.R507H in *FAN1* and karyomegalic interstitial nephritis PheRS did not replicate in the non-European ancestry cohort, potentially due to the small number of individuals with the allele in the replication cohort (n=15).

Sequencing individuals with novel variants

To test for additional rare variants segregating with high PheRS individuals, we analyzed the whole exome sequences (WES) of 84 individuals from the discovery analysis for seven of the significant variants (Tables S7), including individuals with elevated (n=36) and non-elevated PheRS (n=48). A total of four individuals were found to carry a second rare, nonsynonymous variant in the target gene (Fig. 3, Tables S8–9). Two were possible compound heterozygotes (phase could not be determined in this analysis) (*PLCG2* and *AGXT*) and two were homozygotes for the variant identified in the discovery analysis (*DGKE* and *AGXT*, confirming the results from genotyping). Three of the four individuals with confirmed second variants had the highest PheRS for their respective diseases among those selected for WES.

The heterozygote for *AGXT* p.A295T who was found to have an additional rare *AGXT* variant through WES (p.R381K) had the highest PheRS for primary hyperoxaluria type 1, a recessive condition characterized by nephrocalcinosis and oxalate nephrolithiasis; an EHR review revealed he had calcium oxalate crystals on urinalysis. The second highest scoring individual, a confirmed heterozygote, was diagnosed with hyperoxaluria which was attributed to his Crohn’s disease. The p.A295T homozygotes in the discovery and replication cohorts were no more symptomatic than their heterozygous counterparts. This evidence, along with the persistence of the signal after removing individuals with second variants, suggests that p.A295T may act as a strong risk factor for hyperoxaluria, with more severe manifestations occurring in individuals with additional genetic or environmental risk factors.

The highest scorer for familial cold autoinflammatory syndrome 3 (FCAS3), a dominant condition caused by variants in *PLCG2*, presented in the emergency room with a systemic “urticarial type rash” for which a cause was never identified, and continued to present with blistering rashes and lip and tongue swelling. WES revealed this patient harbors a second rare variant (p.R687S) in the SH2 domain of *PLCG2*. A nearby variant (p.S707Y), also in the SH2 domain, has been implicated in a related dominant disease that has overlapping features with FCAS3 [OMIM #614878].(26)

The confirmed homozygote for p.W322* in *DGKE*, a recessive gene that causes nephrotic syndrome type 7, was diagnosed with hemolytic uremic syndrome as a child and received a kidney transplant in his teens; a genetic etiology for his symptoms was never explored.

Sequencing confirmed the variants called on the Exome BeadChip for *SUOX*, *SH2B3*, *SPTBN2*, and *TG* and did not reveal any additional rare variants in the target genes. For *SH2B3*, the lack of a second variant is consistent with an established dominant inheritance pattern, as well as the high proportion of heterozygotes in our cohort (20 of 22) with at least one feature of familial erythrocytosis. Individuals without a second variant in *AGXT* and *DGKE*, both associated with recessive conditions, also had elevated PheRS (Fig. 3). This stands in contrast to the heterozygotes for the *CFTR* variants, who did not have a significantly elevated PheRS (p=0.51, linear regression assuming a dominant model adjusted for age and sex), consistent with a recessive inheritance model. These findings suggest a blurring of the distinction between dominant and recessive labels for some genes.

Sequencing did not reveal any additional rare nonsynonymous variants in the 36 individuals with non-elevated PheRS. In general, individuals with the highest PheRS were more likely to be clinically diagnosed or have additional genetic variants related to their symptoms (Fig. S18).

Biologic validation of *SH2B3*, *TG*, and *SUOX* associations

We selected three candidate novel associations for biologic validation: *SH2B3*, *SUOX*, and *TG*. SH2B3 is a negative regulator of cytokine signaling in hematopoietic cells that operates via a direct interaction between its SH2 domain and JAK2 to attenuate JAK2-mediated activation of proliferative pathways.(27) The variant identified in this study, p.E395K, is located in a region of the protein that is critical for its inhibitory function(28) and is near known disruptive variants.(29) HEK293T cells stimulated with Erythropoietin (EPO) showed an increase in pERK levels that was quenched in the presence of wildtype SH2B3 but not quenched with both the known p.R392E variant and our p.E395K variant (Fig. 4A, 4B).

Splicing prediction programs suggested a probable reduction in 5′ donor strength for *SUOX* p.R76S and possible generation of an exonic cryptic splice acceptor site by *TG* p.G77S. *SUOX* p.R76S is located at the conserved −1 position of the 5′ donor of exon 5. We demonstrated the *SUOX* variant caused a decrease in exon inclusion from 96% to 35% (unpaired two-tailed t-test, p<0.001, Fig 4C). No transcripts aside from the exon-included and exon-skipped transcripts were detected. Similarly, *TG* p.G77S resulted in altered splicing. The basal rate of exon inclusion was reduced from 65% for the wildtype *TG* exon to only 26% inclusion in the p.G77S exon (unpaired two-tailed t-test, p<0.001). These ratios were consistent across a range of cDNA concentrations and PCR cycle numbers (Fig. 4C, 4D).

Comparison of PheRS to existing methods to determine variant pathogenicity

Across all PheRS variant associations with nominal p < 0.05 (n=454), functional annotations were significantly correlated with PheRS effect size (Wilcoxon rank-sum test); splice donor/acceptor and stop-gain variants tended to have the largest effect size, followed in decreasing order by missense, splice region, synonymous, and intron/UTR variants (Fig. S19A).

Thirteen of 14 functional prediction methods trended or associated with the probability of finding associations with the PheRS; predictions from CADD, SiPhy,(30) and Polyphen2 HVAR(10) were statistically significant (p<0.05 using Fisher's exact; Fig. S19B).

Discussion

In our validation study, PheRS was very effective in identifying patients with diagnosed Mendelian disease using only the phenotypic signatures. Applying PheRS to a genotyped population, we found an increased burden of phenotypes among individuals with rare variants in Mendelian disease genes. Sequencing identified or confirmed second rare variants in four individuals, three of whom had the highest PheRS among all heterozygotes or homozygotes for that variant. In vitro studies provided supporting evidence of pathogenicity for all three variants tested.

While our approach relies on many decades of accumulated knowledge about the phenotypic imprint of Mendelian disease, the method itself is simple to implement. Our ability to replicate in an external cohort suggests that it is portable and would therefore be applicable to datasets like the Million Veteran Program, UK Biobank, and the *All of Us* Research Program.[1] Applied to such large populations, this method could facilitate the discovery of pathogenic variants, refine estimates of penetrance across diverse populations, and provide a more nuanced understanding of inheritance patterns, which this study suggests may be more complex than merely "recessive" or "dominant" for some genes. Incorporation of richer EHR data, such as laboratory results and clinical notes(31), could increase the resolving power of PheRS. Furthermore, this method may be used with other combinations of phenotypes that do not follow established Mendelian patterns, perhaps based on undiagnosed patients with unusual presentations.

The American College of Medical Genetics and Genomics established guidelines for variant interpretation that reflect the need to combined multiple lines of evidence, including population-based genotype-phenotype correlations.(32) Our method provides a high-throughput means to generate such evidence. Using these guidelines, ten of the variants from the discovery analysis were interpreted as having "uncertain significance." By adding data from the PheRS analysis, in combination with evidence from our in vitro studies, four of these variants could be converted to "likely pathogenic" or "pathogenic" (Tables 1, S10).

Our findings suggest that the phenotypic burden of rare variants in Mendelian genes may be greater than previously thought. A combination of PheRS and sequencing identified symptomatic individuals with genetics consistent with established inheritance patterns – heterozygous individuals for dominant genes (*SH2B3*, *PLCG2*) and individuals with confirmed second variants in recessive genes (*DGKE*, *AGXT*) – none of whom were diagnosed with a genetic condition. A much larger number of individuals were heterozygous for variants in genes with a presumed recessive inheritance, and yet still had symptoms consistent with the Mendelian disease pattern. While we cannot exclude the possible influence of structural or non-coding variants, the evidence suggests that these variants

[1] Precision Medicine Initiative and All of Us are service marks of the U.S. Department of Health and Human Services.

increase risk in heterozygotes. These individuals tend to have disease that is mild compared to the classic presentations, but severe relative to the general population. For example, homozygous pathogenic mutations in *TG* are associated with congenital goiter which often progresses to thyroid carcinoma; our most severely affected heterozygote received a thyroidectomy at age 26 for goiter and thyroid carcinoma.

This work adds to the evidence that Mendelian and complex disease are not dichotomous, but rather exist on a spectrum. As a method that is both high-throughput and sensitive to the vast knowledge already acquired, PheRS is a tool that may help bridge the gap between Mendelian and complex disease. A consequential question is whether the treatments designed for a Mendelian condition could be effective in individuals with non-traditional molecular presentations. Of the 17 diseases represented among those patients with suspected but undiagnosed Mendelian disease, 11 have specific treatments available (Table S11), some of which could alter the long-term course of the disease.

The impact this approach will have on accelerating precision medicine depends on three interrelated challenges. First, we must integrate statistical associations generated with PheRS into guidelines used for variant interpretation. Second, as we collect stronger evidence for the phenotypic effects of rare variants, we must learn to rapidly and effectively integrate that knowledge into clinical care. Third, we must determine if PheRS can be used to prospectively identify patients whose symptoms are caused by variants in Mendelian genes. If these challenges are addressed, approaches like ours may ultimately enable the conversion of big data not just to knowledge but also to improved care and outcomes for patients.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by R01-LM010685, K22-LM011939, and T15-LM007359 from the National Library of Medicine, and U01s supporting Vanderbilt's participation in the Electronic Medical Records and Genomics (eMERGE) network (HG004603, HG006378, HG008672) and T32-HG008341 from National Human Genome Research Institute. BioVU received and continues to receive support through the National Center for Research Resources UL1-RR024975, which is now the National Center for Advancing Translational Sciences, UL1-TR000445. Other support comes from R01-GM114128, P50-GM115305, R01-GM120523, and R01-HL133786. The authors thank Bingshan Li and Qiang Wei for technical assistance on the whole exome sequence variant calling pipeline. Software for PheWAS analysis and visualization is available at https://github.com/PheWAS/PheWAS. Software to generate PheRS scores and produce grid visualizations is available at https://github.com/labastar/PheRS. These data have been submitted to dbGaP; the submission is currently under review.

References and Notes

1. Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA. Cell. 2011; 147:32–43. [PubMed: 21962505]
2. Blair DR, et al. Cell. 2013; 155:70–80. [PubMed: 24074861]
3. Groza T, et al. The American Journal of Human Genetics. 2015; 97:111–124. [PubMed: 26119816]
4. Langlais D, Fodil N, Gros P. Annual Review of Immunology. 2017; 35:1–30.
5. Denny JC, Bastarache L, Roden DM. Annu Rev Genomics Hum Genet. 2016; doi: 10.1146/annurev-genom-090314-024956
6. Bush WS, Oetjens MT, Crawford DC. Nat Rev Genet. 2016; 17:129–145. [PubMed: 26875678]

7. Marouli E, et al. Nature. 2017; 542:186–190. [PubMed: 28146470]

8. MacArthur DG, et al. Nature. 2014; 508:469–476. [PubMed: 24759409]

9. Kircher M, et al. Nat Genet. 2014; doi: 10.1038/ng.2892

10. Adzhubei IA, et al. Nat Methods. 2010; 7:248–249. [PubMed: 20354512]

11. Ioannidis NM, et al. Am J Hum Genet. 2016; 99:877–885. [PubMed: 27666373]

12. Rehm HL, et al. New England Journal of Medicine. 2015; 372:2235–2242. [PubMed: 26014595]

13. Lek M, et al. Nature. 2016; 536:285–291. [PubMed: 27535533]

14. Manrai AK, et al. N Engl J Med. 2016; 375:655–665. [PubMed: 27532831]

15. Chen R, et al. Nat Biotech. 2016; 34:531–538.

16. Van Driest SL, et al. JAMA. 2016; 315:47–57. [PubMed: 26746457]

17. Kohane IS. Nature Reviews Genetics. 2011; 12:nrg2999.

18. Crawford DC, et al. Front Genet. 2014; 5doi: 10.3389/fgene.2014.00184

19. OMIM – Online Mendelian Inheritance in Man, available at http://omim.org/

20. Köhler S, et al. Nucleic Acids Res. 2014; 42:D966–974. [PubMed: 24217912]

21. Denny JC, et al. Nat Biotechnol. 2013; 31:1102–1111. [PubMed: 24270849]

22. Verma A, et al. PLoS ONE. 2016; 11:e0160573. [PubMed: 27508393]

23. Wei WQ, et al. PLOS ONE. 2017; 12:e0175508. [PubMed: 28686612]

24. MacLeod EL, Ney DM. Ann Nestle Eng. 2010; 68:58–69. [PubMed: 22475869]

25. Stenson PD, et al. Hum Genet. 2017; 136:665–677. [PubMed: 28349240]

26. Zhou Q, et al. Am J Hum Genet. 2012; 91:713–720. [PubMed: 23000145]

27. Maslah N, Cassinat B, Verger E, Kiladjian JJ, Velazquez L. Leukemia. 2017; 31:1661–1670. [PubMed: 28484264]

28. Tong W, Zhang J, Lodish HF. Blood. 2005; 105:4604–4612. [PubMed: 15705783]

29. Camps C, et al. Haematologica. 2016; 101:1306–1318. [PubMed: 27651169]

30. Garber M, et al. Bioinformatics. 2009; 25:i54–62. [PubMed: 19478016]

31. Hebbring SJ, et al. Bioinformatics. 2015; 31:1981–1987. [PubMed: 25657332]

32. Richards S, et al. Genet Med. 2015; 17:405–424. [PubMed: 25741868]

33. Roden DM, et al. Clin Pharmacol Ther. 2008; 84:362–369. [PubMed: 18500243]

34. Pritchard JK, Stephens M, Donnelly P. Genetics. 2000; 155:945–959. [PubMed: 10835412]

35. McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD. Personalized Medicine. 2005; 2:49–79.

36. McLaren W, et al. Genome Biol. 2016; 17doi: 10.1186/s13059-016-0974-4

37. Wang K, Li M, Hakonarson H. Nucleic Acids Res. 2010; 38:e164. [PubMed: 20601685]

38. Purcell S, et al. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]

39. McKenna A, et al. Genome Res. 2010; 20:1297–1303. [PubMed: 20644199]

40. Li H, Durbin R. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]

41. DePristo MA, et al. Nat Genet. 2011; 43:491–498. [PubMed: 21478889]

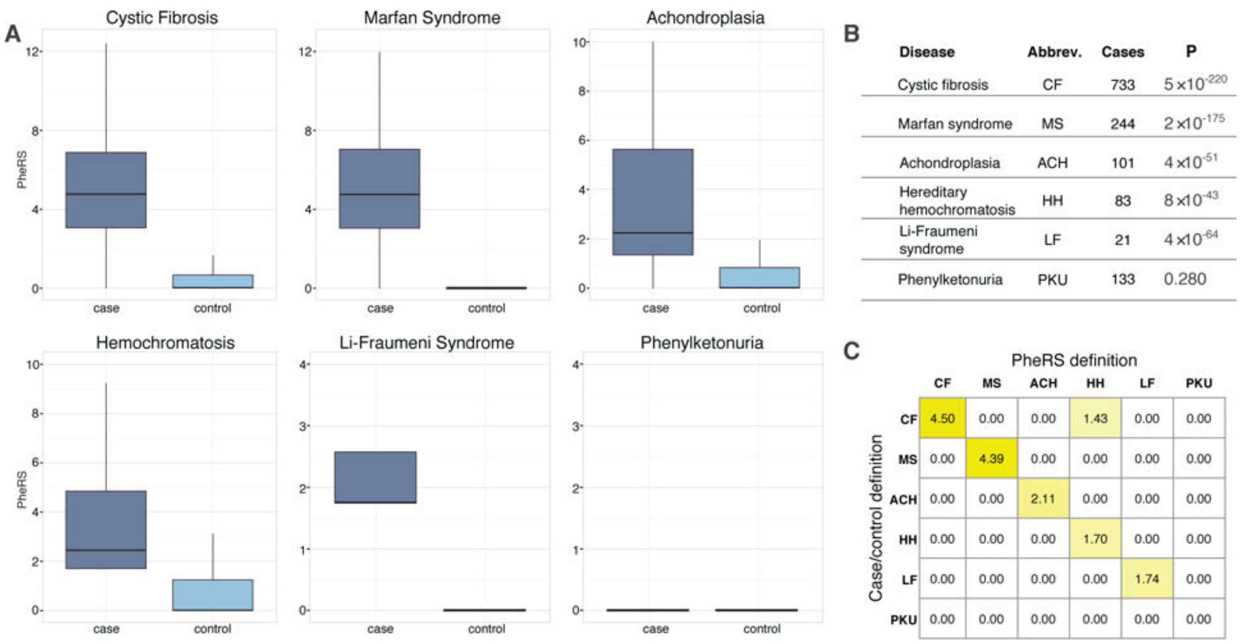
42. Zhao R, et al. J Biol Chem. 2005; 280:22788–22792. [PubMed: 15863514]

43. Desmet FO, et al. Nucleic Acids Res. 2009; 37:e67. [PubMed: 19339519]

44. Schneider CA, Rasband WS, Eliceiri KW. Nat Methods. 2012; 9:671–675. [PubMed: 22930834]

One Sentence Summary

A method to aggregate Mendelian disease phenotypes from the electronic health record finds that complex disease may be explained by a single gene variant in some patients.



Disease	Abbrev.	Cases	P
Cystic fibrosis	CF	733	5×10^{-220}
Marfan syndrome	MS	244	2×10^{-175}
Achondroplasia	ACH	101	4×10^{-51}
Hereditary hemochromatosis	HH	83	8×10^{-43}
Li-Fraumeni syndrome	LF	21	4×10^{-64}
Phenylketonuria	PKU	133	0.280

	CF	MS	ACH	HH	LF	PKU
CF	4.50	0.00	0.00	1.43	0.00	0.00
MS	0.00	4.39	0.00	0.00	0.00	0.00
ACH	0.00	0.00	2.11	0.00	0.00	0.00
HH	0.00	0.00	0.00	1.70	0.00	0.00
LF	0.00	0.00	0.00	0.00	1.74	0.00
PKU	0.00	0.00	0.00	0.00	0.00	0.00

Fig. 1. Phenotype risk scores capture the diagnostic fingerprint of Mendelian disease in EHR data

Scores for six Mendelian diseases were calculated for clinically diagnosed cases and controls matched by age, sex, race, and record length. (**A**) Boxplots of PheRS for cases and controls for each disease. (**B**) Number of cases and statistical significance between cases and controls (Wilcoxon rank-sum test) for each disease. (**C**) Matrix of standardized differences in location (pseudomedian) of the PheRS between cases and controls (by row) and for each Mendelian disease definition (by column).

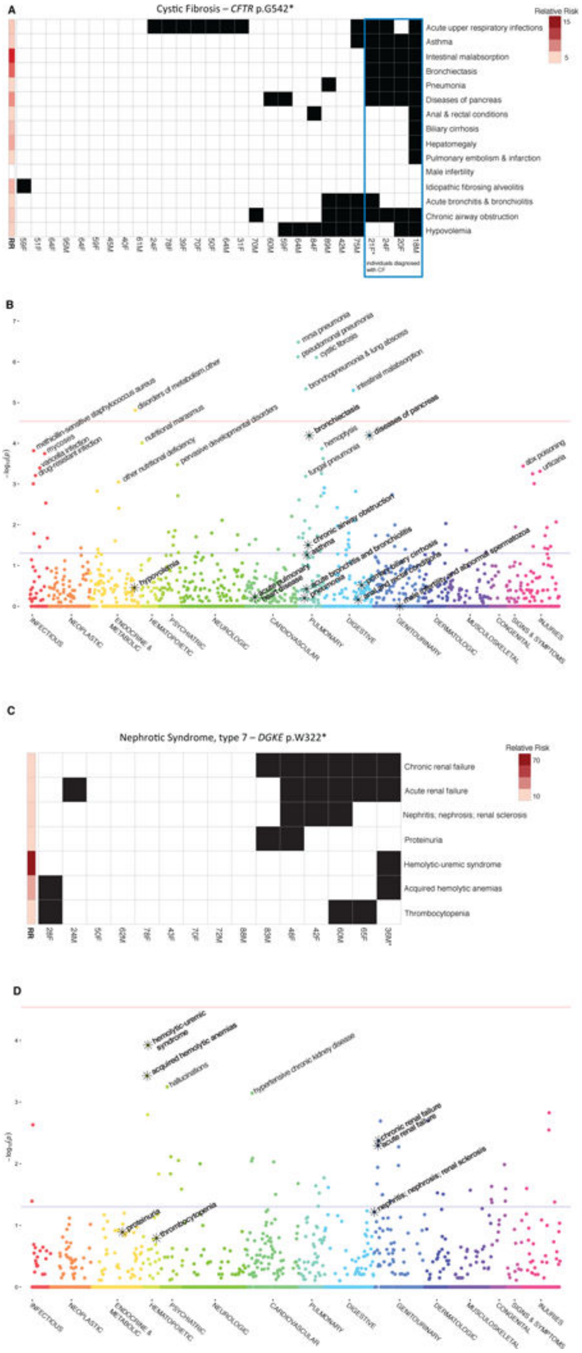


Fig. 2. Phenotypes and PheWAS for two variants associated with PheRS for cystic fibrosis and nephrotic syndrome

For phenotype grids (**A**) and (**C**), each row corresponds to a phenotype used in the PheRS; each column represents an individual who is heterozygous or homozygous (starred) for the variant. The bar on the left of the grid indicates the relative risk for each phenotype compared to wildtype. In grid (A), individuals clinically diagnosed with cystic fibrosis are enclosed by a blue box. PheWAS plots (**B**) and (**D**) show the PheWAS for the variant (Fisher's exact p-value). The constituent phenotypes that define the PheRS are starred. All associations with p<0.001 are labeled. The horizontal red and blue lines show the Bonferroni correction threshold for an individual PheWAS and the nominal (uncorrected) p=0.05, respectively.

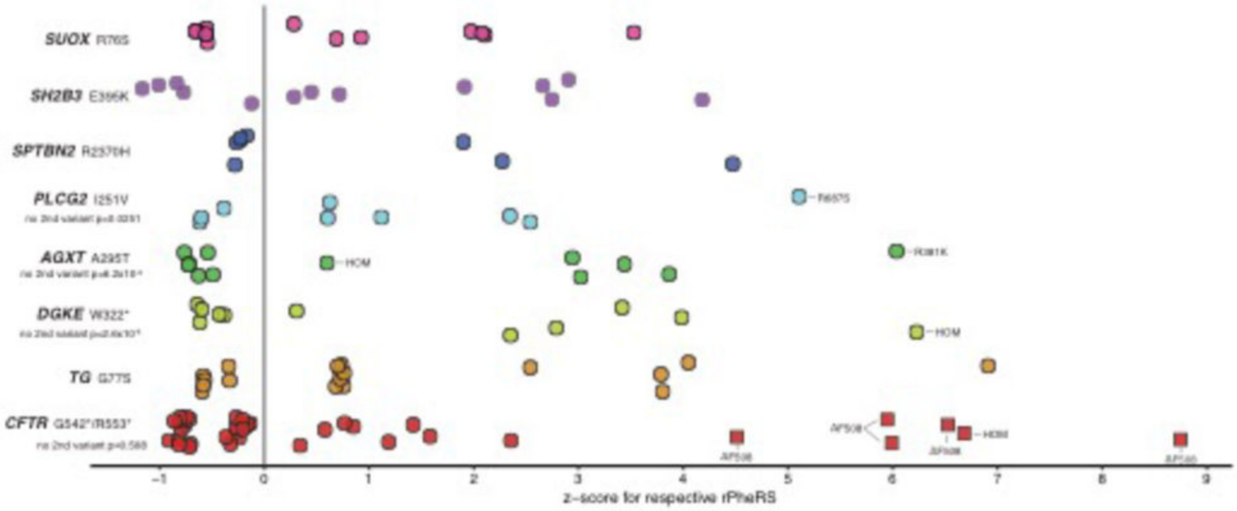


Fig. 3. Whole exome sequencing reveals second variants among individuals with high PheRS and demonstrates disease risk in heterozygotes

Each point represents an individual who is heterozygous or homozygous for the variant labeled on the left. The x-axis represents the z-score for the PheRS relative to what is expected given age and sex (using the residual from the PheRS). All individuals carry at least one copy of the variant indicated on the left; additional variants identified by whole exome sequencing or clinical chart review are labeled for each individual; homozygotes confirmed by sequencing are labeled "HOM." Additional *CFTR* variants were ascertained from clinical testing in the EHR; all other individuals were sequenced for this study. Clinically diagnosed individuals are squares; all others are circles. Where additional variants were found, the association test from the discovery analysis was repeated after dropping individuals with a second variant (p-values generated using linear regression assuming dominant model adjusted for age and sex), and the p-value is recorded under the gene/variant label.

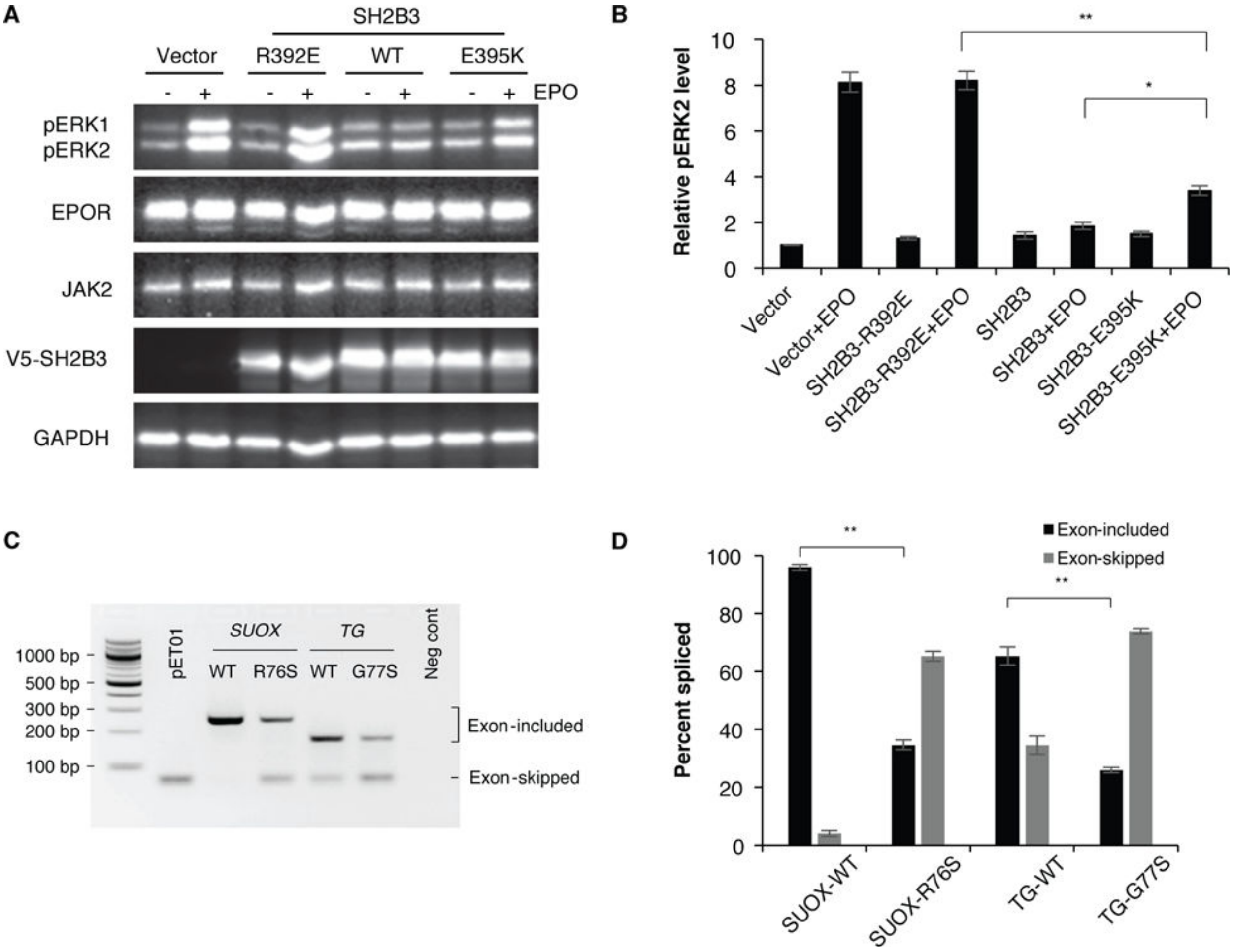


Fig. 4. PheRS enriches for variants with altered function in vitro

Representative Western blots (**A**) and mean phospho-ERK2 levels normalized to EPOR expression (**B**) in EPOstimulated HEK293T cells transiently transfected with wildtype (WT) versus variant SH2B3 constructs, EPOR, and JAK2. As expected, known variant SH2B3-R392E fails to inhibit EPOstimulated ERK phosphorylation. Similarly, SH2B3-E395K shows approximately 1.8-fold elevation of EPO-stimulated ERK activation at 10 min relative to wildtype SH2B3. RT-PCR analysis (**C**) and quantification (**D**) of WT versus variant splicing of *SUOX* and *TG* exons in HEK293T cells transiently transfected with empty minigene vector pET01, pET01 containing exons of interest flanked by 100 bp of intronic sequence, or negative control pIRES2-EGFP. Absolute change in the percent of exon-inclusion was −61% for *SUOX*-VAR and −39% for *TG*VAR. Means ± SEM; *n* = four (A, B) or five (C, D) independent experiments; unpaired twotailed t test; *p=0.003, **p<0.001.

Table 1

Significant associations between phenotype risk scores for Mendelian disease and rare variants

Significant results from the analysis of 7,520 PheRS-variant pairs, generated using linear regression assuming a dominant model, adjusting for age and sex. All associations with a false discovery rate of q<0.1 are included. The established mode of inheritance is listed in the “OMIM Inheritance” column; “Rec*” indicates that disease has also been reported in heterozygotes. ClinVar designations are included, when available: P = pathogenic, LP = likely pathogenic, LB = likely benign, U = uncertain significance. Variants with relevant phenotype associations in HGMD are indicated with “Y.” Results from applying American College of Medical Genetics and Genomics (ACMG) interpretations are found in the last column; an → indicates the interpretation changed in light of evidence presented in this paper.

Gene	Variant	dbSNP	HOM/HET	Associated Mendelian Disease	OMIM Reported inheritance	Phenotype categories in PheRS	Beta	P	ClinVar	HGMD	ACMG
CFTR	c.1624G>T p.Gly542Ter	rs113993959	1/27	Cystic fibrosis	AR	🟪 🟧🟩	1.39	2.9×10^{-8}	P	Y	P
CHRNA4	c.1448G>A p.Arg483Gln	rs55855125	1/21	Nocturnal frontal lobe epilepsy, 1	AD	🟥	0.58	9.0×10^{-8}	U		U
DGKE	c.966G>A p.Trp322Ter	rs138924661	1/14	Nephrotic syndrome, type 7	AR	🟪 🟩	1.31	2.8×10^{-7}	LP	Y	LP→P
SUOX	c.228G>T p.Arg76Ser	rs202085145	0/24	Sulfocysteinuria	AR	🟥 ▨⬛	0.82	1.7×10^{-6}	U		U→P
CFTR	c.1657C>T p.Arg553Ter	rs74597325	0/12	Cystic fibrosis	AR	🟪 🟧🟩	1.81	2.1×10^{-6}	P	Y	P
KIF1B	c.2021C>T p.Thr674 Ile	rs41274468	0/21	Charcot-Marie-Tooth disease, 2A1	AD	🟥 ▨⬛	0.79	5.3×10^{-6}			U
VWF	c.5851A>G p.Thr1951Ala	rs144072210	0/21	Von Willebrand disease	AR*	🟪 🟧 ▨	0.53	8.6×10^{-6}		Y	U
KIF1A	c.2676C>T p.Ala993=	rs116297894	1/25	Spastic paraplegia-30	AR	🟥 🟩 ⬛	0.84	1.3×10^{-5}	LB		LB→U
F10	c.872G>A p.Arg291Gln	rs149212700	0/15	Factor X deficiency	AR *	🟪 🟧 ▨	0.62	1.9×10^{-5}			U
HFE	c.502G>C p.Glu168Gln	rs146519482	0/40	Hemochromatosis	AR	🟦🟪 🟧🟩▨	1.08	4.0×10^{-5}	U	Y	U
TG	c.229G>A p.Gly77Ser	rs142698837	0/69	Thyroid dyshormonogenesis	AR	🟦🟪🟥	0.26	6.0×10^{-5}		Y	U→P
SH2B3	c.1183G>A p.Glu395Lys	rs148636776	0/22	Familial erythrocytosis, 1	AD	🟪🟥🟧🟩▨⬛	1.48	6.1×10^{-5}			U→P

Gene	Variant	dbSNP	HOM/HET	Associated Mendelian Disease	OMIM Reported inheritance	Phenotype categories in PheRS	Beta	P	ClinVar	HGMD	ACMG
SPTBN2	c.7109G>A p.Arg2370His	rs145522851	0/11	Spinocerebellar ataxia	AR*	🟥	0.75	9.0×10^{-5}			U→LP
FAN1	c.1520G>A p.Arg507His	rs150393409	0/434	Interstitial nephritis, karyomegalic	AR	🟪 🟩	0.15	9.9×10^{-5}			LB→U
PANK2	c.1561G>A p 521.Gly Arg	rs137852959	0/26	HARP syndrome	AR	🟪 🟥 🟩 ▨	0.58	1.1×10^{-4}	P	Y	P
SH2B3	c.1183G>A p.Glu395Lys	rs148636776	0/22	Essential thrombocythemia	AD	🟦 🟧 🟩	0.33	1.4×10^{-4}			U→P
AGXT	c.883G>A p.Ala295Thr	rs13408961	1/35	Primary hyperoxaluria, type I	AR	🟪 🟥 🟧 ▨ ⬛	0.82	1.7×10^{-4}	U/LB		LB→U
PLCG2	c.751A>G p.Ile251Val	rs190840748	0/10	Familial cold autoinflammatory syn. 3	AD	🟪 🟧 ▨	0.70	1.9×10^{-4}			U

🟦 Neoplastic

🟪 Endocrine/Metabolic/Blood

🟥 Nervous/Psychiatric/Sensory

🟧 Circulatory/Respiratory

🟩 Digestive/Genitourinary

▨ Musculoskeletal/Dermatologic

⬛ Other symptoms/Injuries