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. Hu,3 Sara L. Van Driest,4,5 Tracy L. McGregor,2 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,3 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 Joshua C. Denny1,4*

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 on the basis of patterns described by Mendelian diseases. We mapped the clinical features of 1204 Mendelian diseases into phenotypes captured from the electronic health record (EHR) and summarized this evidence as phenotype risk scores (PheRSs). In an initial validation, PheRS distinguished cases and controls of five Mendelian diseases. Applying PheRS to 21,701 genotyped individuals uncovered 18 associations between 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 considerable morbidity and mortality. Many are characterized by a 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 diseases (2), and single-nucleotide polymorphisms found in genome-wide association studies (GWASs) are enriched for Mendelian loci (3). A review of evidence from GWASs and whole-exome sequencing studies found an overlap between primary immunodeficiency genes and complex inflammatory diseases (4). Collectively, this evidence suggests that variants in Mendelian disease-causing genes may be an underrecognized contributor to complex disease.

Until very recently, the phenotypic effects of rare genetic variants were ascertained primarily through family-based studies of patients with distinct and often severe phenotypes. Population-level techniques, such as GWASs and phenotype association studies (PheWASs) (5, 6), are not easily applied to rare variation because most studies are underpowered. Cohorts large enough to support GWASs 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 (9–22), and consortia such as ClinGen (22) are aggregating knowledge to enable expert determinations. Resources such as ExAC (23) have helped refine variant interpretation. Some variants previously interpreted as pathogenic are too common in some populations to cause rare, life-threatening disorders (14), whereas 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 1204 Mendelian conditions identified both known and previously unknown associations with variants in target genes. The approach presents a method for measuring the phenotypic impact of rare variants and for identifying the heretofore underrecognized contribution of Mendelian disease genes to common medical conditions.

Constructing a phenotype risk score

The Online Mendelian Inheritance in Man (OMIM) database provides clinical synopses for thousands of monogenic diseases (19) that have been annotated using the Human Phenotype Ontology (HPO) (20). We created a map from HPO terms to consolidated billing codes from the EHR called phenocodes. Phenocodes 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 phenocodes, 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 compared the PheRSs of clinically diagnosed cases with matched controls for six Mendelian diseases. PheRS was a very strong predictor of case status for five of the diseases (Wilcoxon rank sum test; \( P = 8 \times 10^{-30} \) to \( 5 \times 10^{-26} \)) (Fig. 1, A and B). The exception was phenylketonuria (\( P = 0.28 \)), which effectively served as a negative control because 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: The cases for different Mendelian diseases had similar PheRS distributions to those for controls (Fig. 1C). The lone exception was that the PheRS for hereditary hypercholesterolemia (HH) was significantly elevated for cystic fibrosis (CF) cases versus controls. However, even in this instance, CF cases had PheRSs that were three times as high for CF compared with 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 6 months after 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 individuals of European ancestry genotyped on the Exome BeadChip (table S1). In this cohort, we computed PheRSs for 1204 Mendelian diseases (1086 causative genes) for which we had sufficient genotype data. We tested for association between PheRSs and 6188 rare variants (minor allele frequency < 1%) using linear regression, assuming a dominant genetic model. All significant results had a positive beta coefficient, indicating that the variants were associated with an excess of Mendelian disease phenotypes. Four of the genes had an established dominant mode of inheritance, whereas the remaining 13 genes were known as exclusively or primarily Mendelian disease against variants in the gene or gene known to cause that disease. We found 18 significant associations between rare variants and PheRSs (false discovery rate \( q < 0.1 \); Table 1).

All significant results had a positive beta coefficient, indicating that the variants were associated with an excess of Mendelian disease phenotypes. Four of the genes had an established dominant mode of inheritance, whereas 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 AF508 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 HH 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.

Although the majority of patients with significant variants were undiagnosed, these individuals had a high burden of severe end points related to the Mendelian diseases. Of the 40 heterozygotes for \( HFE \) p.E168Q, four had liver transplants (10 versus 1.2% in background; \( P = 2.1 \times 10^{-5}; \) Fisher’s exact test). Individuals with variants in two genes associated with renal failure had elevated rates of kidney transplant: Five of 36 patients (14%) with \( AGXT \) p.A295T received transplants (another is awaiting transplant), as well as two of 15 patients (13%) with \( DGKE \) p.W322* (versus 3% in background; \( P = 6.9 \times 10^{-3}\) and \( P = 0.088\); respectively; Fisher’s exact test). Four of 69 TG p.G778S heterozygotes underwent thyroidec- tomyes (6 versus 2% of noncarriers; \( P = 0.039\); Fisher’s exact test). These are end-stage phenotypes, potentially resulting from the effects of these variants, and did not have an increased prevalence in other significant variant carriers (table S3). Additionally, we found that 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).

Fig. 1. PheRSs 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 PheRSs 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 PheRSs between cases and controls (by row) and for each Mendelian disease definition (by column).
Replication of previously unrecognized associations

We attempted to replicate significant associations from the discovery analysis in two independent cohorts: a European ancestry cohort from Marshfield Clinic (n = 9441) and a non-European ancestry cohort from Vanderbilt (n = 3820) (tables S4 and S5). Each was tested as in the discovery cohort, using linear regression assuming a dominant model, adjusted 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 were replicated: p.G77S in TG with thyroid dyshormonogenesis PheRS (P = 5.0 × 10^{-4}) and p.R507H in FAN1 with karyomegalic interstitial nephritis PheRS (P = 8.2 × 10^{-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 × 10^{-3}) and p.A295T in AGXT with primary hyperoxaluria type 1 PheRS (P = 3.9 × 10^{-3}). The association between p.R507H in FAN1 and karyomegalic interstitial nephritis PheRS was not replicated in the non-European ancestry cohort, potentially owing to the small number of individuals with the allele in the replication cohort (n = 15).

Sequencing individuals with hitherto unrecognized variants

To test for additional rare variants segregating with high-PheRS individuals, we analyzed the whole-exome sequences of 84 individuals from the discovery analysis for seven of the significant variants (table S7), including individuals with elevated (n = 36) and nonelevated (n = 48) PheRSs. A total of four individuals were found to carry a second rare, nonsynonymous variant in the target gene (Fig. 3 and tables S8 and S9). 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 whole-exome sequencing.

The heterozygote for AGXT p.A295T who was found to have an additional rare AGXT variant (p.R381K) through whole-exome sequencing had the highest PheRS for primary hyperoxaluria type 1, a recessive condition characterized by nephrocalcinosis and oxalate nephrolithiasis; an EHR review revealed that he had calcium oxalate crystals evident in 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.

Table 1. Significant associations between phenotype risk scores (PheRSs) for Mendelian disease and rare variants. Shown are significant results from the analysis of 7520 PheRS-variant pairs, generated using linear regression assuming a dominant model, adjusted 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 reported inheritance” column; AD indicates autosomal dominant, AR indicates autosomal recessive; and AR* 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 a “Y.” Results from applying American College of Medical Genetics and Genomics (ACMG) interpretations are given in the last column; an arrow indicates that the interpretation changed in light of evidence presented in this paper. rsIDs are assigned by the dbSNP database.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Variant</th>
<th>rsID</th>
<th>HOM/HET</th>
<th>Associated Mendelian Disease</th>
<th>OMIM Reported inheritance</th>
<th>Phenotype categories in PheRS</th>
<th>Beta</th>
<th>P</th>
<th>ClinVar</th>
<th>HGMD</th>
<th>ACMG</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFTR</td>
<td>c.1654G&gt;A</td>
<td>rs113993959</td>
<td>1/27</td>
<td>Cystic fibrosis</td>
<td>AR</td>
<td></td>
<td>1.39</td>
<td>2.9 × 10^{-8}</td>
<td>P</td>
<td>Y</td>
<td>P</td>
</tr>
<tr>
<td>CHRNA4</td>
<td>c.9170C&gt;G</td>
<td>rs5885125</td>
<td>1/21</td>
<td>Nocturnal frontal lobe epilepsy, 1</td>
<td>AD</td>
<td></td>
<td>0.58</td>
<td>9.0 × 10^{-6}</td>
<td>U</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td>DGKE</td>
<td>c.9170C&gt;G</td>
<td>rs13892461</td>
<td>1/14</td>
<td>Nephrotic syndrome, type 7</td>
<td>AR</td>
<td></td>
<td>1.31</td>
<td>2.8 × 10^{-7}</td>
<td>LP</td>
<td>Y</td>
<td>LP+P</td>
</tr>
<tr>
<td>SUOX</td>
<td>c.228T&gt;G</td>
<td>rs20208514</td>
<td>0/24</td>
<td>Sulfoxyurinuria</td>
<td>AR</td>
<td></td>
<td>0.82</td>
<td>1.7 × 10^{-6}</td>
<td>U</td>
<td>U-P</td>
<td></td>
</tr>
<tr>
<td>CFTR</td>
<td>c.9170C&gt;G</td>
<td>rs7459732</td>
<td>0/12</td>
<td>Cystic fibrosis</td>
<td>AR</td>
<td></td>
<td>1.81</td>
<td>2.1 × 10^{-6}</td>
<td>P</td>
<td>Y</td>
<td>P</td>
</tr>
<tr>
<td>KIF1B</td>
<td>c.2021C&gt;T</td>
<td>rs14274468</td>
<td>0/21</td>
<td>Charcot-Marie-Tooth disease, 2A1</td>
<td>AD</td>
<td></td>
<td>0.79</td>
<td>5.3 × 10^{-4}</td>
<td>U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VWF</td>
<td>c.5851A&gt;G</td>
<td>rs14407221</td>
<td>0/21</td>
<td>Von Willebrand disease</td>
<td>AR*</td>
<td></td>
<td>0.53</td>
<td>8.6 × 10^{-6}</td>
<td>Y</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td>KIF1A</td>
<td>c.7876T&gt;G</td>
<td>rs11629789</td>
<td>1/25</td>
<td>Spastic paraplegia-30</td>
<td>AR*</td>
<td></td>
<td>0.84</td>
<td>1.3 × 10^{-5}</td>
<td>LB</td>
<td>U</td>
<td>LB+U</td>
</tr>
<tr>
<td>F10</td>
<td>c.872C&gt;A</td>
<td>rs14921270</td>
<td>0/15</td>
<td>Factor X deficiency</td>
<td>AR*</td>
<td></td>
<td>0.62</td>
<td>1.9 × 10^{-5}</td>
<td>U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HFE</td>
<td>c.500G&gt;C</td>
<td>rs14651942</td>
<td>0/40</td>
<td>Hemochromatosis</td>
<td>AR</td>
<td></td>
<td>1.08</td>
<td>4.0 × 10^{-4}</td>
<td>U</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td>TG</td>
<td>c.229C&gt;A</td>
<td>rs14269883</td>
<td>0/69</td>
<td>Thyroid dyshormonogenesis</td>
<td>AR</td>
<td></td>
<td>0.26</td>
<td>6.0 × 10^{-5}</td>
<td>Y</td>
<td>U-P</td>
<td></td>
</tr>
<tr>
<td>SH2B3</td>
<td>c.1735G&gt;A</td>
<td>rs14683677</td>
<td>0/22</td>
<td>Familial erythrocytosis, 1</td>
<td>AD</td>
<td></td>
<td>1.48</td>
<td>6.1 × 10^{-4}</td>
<td>U-P</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPTBN2</td>
<td>c.7109G&gt;A</td>
<td>rs14552285</td>
<td>0/11</td>
<td>Spinocerebellar ataxia</td>
<td>AR*</td>
<td></td>
<td>0.75</td>
<td>9.0 × 10^{-5}</td>
<td>U-LP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAN1</td>
<td>c.1505G&gt;A</td>
<td>rs15039349</td>
<td>0/434</td>
<td>Intestinal nephritis, karyomegalic</td>
<td>AR</td>
<td></td>
<td>0.15</td>
<td>9.9 × 10^{-5}</td>
<td>LB-U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PANK2</td>
<td>c.1561G&gt;A</td>
<td>rs13786289</td>
<td>0/26</td>
<td>HARP syndrome</td>
<td>AR</td>
<td></td>
<td>0.58</td>
<td>1.1 × 10^{-4}</td>
<td>P</td>
<td>Y</td>
<td>P</td>
</tr>
<tr>
<td>SH2B3</td>
<td>c.1183G&gt;A</td>
<td>rs14683677</td>
<td>0/22</td>
<td>Essential thrombocythemia</td>
<td>AD</td>
<td></td>
<td>0.33</td>
<td>1.4 × 10^{-4}</td>
<td>U</td>
<td></td>
<td>U-P</td>
</tr>
<tr>
<td>AGXT</td>
<td>c.893G&gt;A</td>
<td>rs13048961</td>
<td>1/35</td>
<td>Primary hyperoxaluria, type I</td>
<td>AD</td>
<td></td>
<td>0.82</td>
<td>1.7 × 10^{-4}</td>
<td>U/LB</td>
<td>LB-U</td>
<td></td>
</tr>
<tr>
<td>PLCG2</td>
<td>c.751A&gt;G</td>
<td>rs19084708</td>
<td>0/10</td>
<td>Familial cold autoinflammatory syn. 3</td>
<td>AD</td>
<td></td>
<td>0.70</td>
<td>1.9 × 10^{-4}</td>
<td>U</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Neoplastic
- Endocrine/Metabolic/Blood
- Nervous/Psychiatric/Sensory
- Circulatory/Respiratory
- Digestive/Genitourinary
- Musculoskeletal/Dermatologic
- Other symptoms/Injuries
The highest scorer for familial cold autoinflammatory syndrome 3 (FCAS3), a dominant condition caused by variants in \( \text{PLCG2} \), 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. Whole-exome sequencing revealed that this patient harbors a second rare variant (p.R687S) in the SH2 domain of \( \text{PLCG2} \). 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 \( \text{DGKE} \), 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 \( \text{SUOX} \), \( \text{SH2B3} \), \( \text{SPTBN2} \), and \( \text{TG} \) and did not reveal any additional rare variants in the target genes. For \( \text{SH2B3} \), 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 \( \text{AGXT} \) and \( \text{DGKE} \), both of which are associated with recessive conditions, also had elevated PheRSs (Fig. 3). This stands in contrast to the heterozygotes for the \( \text{CFTR} \) 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 nonelevated PheRSs. In general, individuals with the highest PheRSs were more likely to be
nized associations for biologic validation: SUOX associations splicing. The basal rate of exon inclusion was re-
tected. Similarly, included and exon-skipped transcripts were de-
Fig. 4C). No transcripts aside from the exon-
inclusion pathogenicity for all three variants tested.

Although our approach relies on many dec-
ades of accumulated knowledge about the pheno-
typic imprint of Mendelian disease, the method it-
tself is simple to implement. Our ability to rep-
licate results in an external cohort suggests that it
is portable and would therefore be applicable to
data sets such as those of the Million Veteran
Program, UK Biobank, and the All of Us Re-
search Program (All of Us is a service mark of
the U.S. Department of Health and Human Services).
Applied to such large populations, this method
could facilitate the discovery of pathogenic var-
iants, 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 labor-
atory results and clinical notes (SI), could in-
crease the resolving power of PheRS. Furthermore,
this method may be used with other combina-
tions of phenotypes that do not follow estab-
lished Mendelian patterns, perhaps based on
undiagnosed patients with unusual presentations.
The American College of Medical Genetics and
Genomics established guidelines for variant inter-
pretation that reflect the need to combine 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, 10 of the

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 previously unrecognized
associations for biologic validation: SH2B3, SUOX, and TG. SH2B3 is a negative regulator of
cytokine signaling in hematopoietic cells that
operates through 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). Human embryonic kidney
(HEK) 293T cells stimulated with erythropoietin
showed an increase in phosphorylated extracellu-
lar signal–regulated kinase (pERK) levels that
was quenched in the presence of wild-type SH2B3
but not quenched with both the known p.R392E
functional mutation and our p.E395K variant (Fig.
4, A and B).

Splicing prediction programs suggested a proba-
ble 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 that 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 de-
tected. Similarly, TG p.G77S resulted in altered
splicing. The basal rate of exon inclusion was re-
duced from 65% for the wild-type 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 concentra-
tions and polymerase chain reaction (PCR) cycle
numbers (Fig. 4, C and D).

Comparison of PheRS with 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/untranslated region variants (fig. S19A).
Thirteen of 14 functional prediction methods
trended or associated with the probability of
finding associations with the PheRS; predictions
from CADD (combined annotation-dependent
depletion) (9), SiPhy (30), and Polyphen2 HVAR
(10) were statistically significant (P < 0.05;
Fishers’s exact test; fig. S19B).

Discussion

In our validation study, PheRS was very effective
in identifying patients with diagnosed Mendelian
disease by 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 dis-
ease genes. Sequencing identified or confirmed
second rare variants in four individuals, three of

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 represented as squares; all others are shown as 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 a dominant model, adjusted for
age and sex), and the P value is recorded under the gene/variant label.

Fig. 3. Whole-exome sequencing reveals second variants among indi-
viduals with high PheRSs and demonstrates disease risk in heterozy-
gotes. Each point represents an individual who is heterozygous or
homzygous 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;

Downloaded from http://science.sciencemag.org on March 15, 2018
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" (Table 1 and table 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 and PLCG2) and individuals with confirmed second variants in recessive genes (DGKE and 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. Although we cannot exclude the possible influence of structural or noncoding variants, the evidence suggests that these variants increase risk in heterozygotes. These individuals tend to have disease that is mild compared with the classic presentations but severe relative to the general population. For example, homozygous pathogenic mutations in TG are associated with congenital goiter, which often progress 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 nontraditional 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 that 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 whether 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.
studies. W.-Q.W., Q.F., D.R., T.E., and D.R.V.E. assisted with designing and analyzing the sequence data. T.E. and J.D.M. provided support for statistical analyses. T.O., A.H.S., J.M., L.B., and J.C.D. reviewed medical records. S.L.V.D., J.D.M., Q.S.W., M.T., R.C., N.I.C., and D.M.R. advised on the interpretation of the results and methods refinement. L.B., J.J.H., D.M.R., and J.C.D. drafted the manuscript. All authors provided critical feedback and helped shape the analysis and interpretation of the results of this study. Competing interests: S.L.V.D. received an honorarium from Merck as an invited speaker. After completing her work on this project, T.L.M. began working for Alnylam Pharmaceuticals and received stock in the company. Neither Merck nor Alnylam supported or were involved in this research. The remaining coauthors have no competing interests to report. Data and materials availability: Software for PheWAS analysis and visualization is available at https://github.com/PheWAS/PheWAS. Software to generate PheRSs and produce grid visualizations is available at https://github.com/labastar/PheRS. Genetic and phenotypic data used in this study are available at dbGap under accession number phs001516.v1.p1.

SUPPLEMENTARY MATERIALS
www.sciencemag.org/content/359/6381/1233/suppl/DC1
Materials and Methods
Figs. S1 to S20
Tables S1 to S17
References (33–44)
14 November 2016; resubmitted 25 August 2017
Accepted 22 January 2018
10.1126/science.aal4043
Phenotype risk scores identify patients with unrecognized Mendelian disease patterns

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

Science 359 (6381), 1233-1239.
DOI: 10.1126/science.aal4043

Hidden effects of Mendelian inheritance
Identifying the determinate factors of genetic disease has been quite successful for Mendelian inheritance of large-effect pathogenic variants. In these cases, two non- or low-functioning genes contribute to disease. However, Mendelian effects of lesser strength have generally been ignored when looking at genomic consequences in human health. Bastarache et al. used electronic records to identify the phenotypic effects of previously unidentified Mendelian variations. Their analysis suggests that individuals with undiagnosed Mendelian diseases may be more prevalent in the general population than assumed. Because of this, genetic analysis may be able to assist clinicians in arriving at a diagnosis.

Science, this issue p. 1233