www.sciencemag.org/content/359/6381/1233/suppl/DC1

Supplementary Materials for

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, Joshua C. Denny*

*Corresponding author. Email: josh.denny@vanderbilt.edu

Published 16 March 2018, *Science* **359**, 1233 (2018)
DOI: 10.1126/science.aal4043

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/content/359/6381/1233/suppl/DC1)

Materials and Methods

Mapping the clinical features of Mendelian diseases to phecodes

To characterize Mendelian diseases using EHR data, we needed to express the clinical descriptions of Mendelian diseases available in OMIM, a catalog of human genetic disorders and diseases, in terms of EHR-derived phenotypes. We used phecodes to define phenotypic features from the EHR. Phecodes are defined by hierarchical groupings of ICD-9 codes and were developed for phenome-wide associations studies. They have been validated in a broad range of genetic association studies, and have been shown to replicate 66% of known genetic associations for a diverse set of phenotypes with an area under the received operator characteristic curve of 0.83.(*21*) The current version of phecodes (V1.2), used in this paper, can be downloaded at http://phewascatalog.org.

We used an existing map between clinical descriptions in OMIM and the Human Phenotype Ontology (HPO), a controlled vocabulary created to describe phenotypic abnormalities as well as modes of inheritance and clinical course of disease.(*3*) The clinical descriptions in OMIM have been annotated with HPO terms.(*20*) We manually mapped 3,459 of HPO terms to 890 unique phecodes. Many of the mapped pairs were not exact matches. Most often, the HPO term was more specific than the mapped phecode (e.g. HPO term "neurogenic bladder" is mapped to the phecode for "functional disorders of bladder"). In some instances, the best available phecode was narrower than its corresponding HPO term (e.g., the HPO term "retinopathy" is mapped to the phecode for "nondiabetic retinopathy"). 2,358 HPO terms did not map to phecodes, either because the HPO term was too specific to be ascertained by claims data (e.g., "depletion of mitochondrial DNA in liver"), too general (e.g., "abnormality of the kidney"), or pertained to non-pathogenic variation not typically captured in a medical setting (e.g., "thin clavicles").

When creating the map, we annotated several classes of terms that were likely to refer to special populations: those related to pregnancy (e.g. preeclampsia), infancy (e.g. neonatal hypotonia), childhood (e.g. delayed puberty), and abnormalities present at birth (e.g. syndactyly). Because the analyses in this paper were based on adult cohorts, we excluded phenotypes that pertain to congenital conditions, infancy, childhood, or pregnancy. Application of these filters eliminated 1,440 mapped and unmapped HPO terms. The designations of "congenital," "infantile," "pediatric," and "pregnancy-related" are available as part of the phecode-HPO map so that different types of features can be included or excluded in subsequent analyses. In all, 2,382 uniquely mapped HPO terms were used in the analysis to calculate PheRS. It is important to note that while the process of mapping and excluding was subjective (and thus subject to disagreement), this work was done before our association analyses and without referring to the underlying disease definitions so as not to bias our results (See Table S12 for the mapping from HPO term identifier to phecode).

Calculating a phenotype risk score

The PheRS is a means of capturing a phenotypic pattern in a single number. It is analogous to a genetic risk score (GRS), which is commonly used to estimate an individual's risk of a disease or trait by summing up multiple risk SNPs. Using the HPO

terms linked to a Mendelian disease and our HPO-phecode map, we described each disease as a set of phecodes. We then weighted each phecode based on the inverse prevalence of the code in a given population. For our research, we used our entire test population to calculate the frequencies, but one could use an external reference population as well. Given a population of *N* individuals, the weight for phenotype *p* is calculated as:

$$w_p = log \frac{N}{n_p}$$

where n_p is the number of individuals with phenotype *p*. The weighting is analogous to the inverse document frequency statistic commonly used in natural language processing to express the relative importance of a word in a document. We applied this weighting so that an individual with a single rare feature of a disease will have a higher score than an individual with a single common feature.

For an individual *i*, the PheRS for a single disease defined by *m* phecodes is calculated as:

$$PheRS_i = \sum_{p=1}^{m} w_p \, x_{i,p}$$

$$where \; x_{i,p} = \begin{cases} 1 \; if \; individual_i \; has \; phenotype_p \\ 0 \; otherwise \end{cases}$$

Description of cohorts

We validated PheRS using diagnosed cases and matched controls with individuals from the Vanderbilt University Medical Center (VUMC) Synthetic Derivative (SD). The SD comprises the de-identified medical records of ~2.5 million individuals and includes essentially all elements of the EHR (e.g. clinical documents, lab results, billing codes).

Our genetic analyses used BioVU, a de-identified DNA biobank linked to the SD.(*33*) No links to identifiers are preserved in BioVU, and thus data cannot be returned to participants. A total of 30,216 adults (age >= 18) linked to genotype data from the Illumina HumanExome BeadChip were used in the genetic association analyses. The cohort was ascertained previously for five different criteria: (1) eligibility as case or control in one of 31 pharmacogenetics studies (2) availability of longitudinal data with primary care visits (3) presence in the cancer registry (4) old age with longitudinal data (5) presence of rare diseases or conditions ascertained via billing codes. Because our work involved rare variants, signals may easily be drowned out by samples that have non-germline variation or are otherwise compromised. Thus, we filtered out 4,695 individuals with evidence of hematological malignancies, blood transfusions, or stem cell transplants prior to the date the blood sample used for genotyping was drawn, as well as patients who visited VUMC exclusively for cancer care. All individuals included in the study had at least one outpatient encounter, and 63% had at least one inpatient stay at the hospital recorded in their EHR. Using STRUCTURE(*34*) to determine ancestry, we divided the uncompromised population into 21,701 individuals of European ancestry and 3,820 individuals of non-European ancestry (primarily African ancestry). The European

cohort sample was used as the discovery set, while the non-European cohort was used for replication.

We also leveraged a second genotyped set of individuals for replication from the Marshfield Clinic Personalized Medicine Research Project.(*35*) Marshfield Clinic contributed a cohort of 10,124 European adults linked with ICD-9 codes and Illumina HumanExome BeadChip data. From this population, we filtered out 683 compromised individuals (as above), leaving a cohort of 9,441 adults of European ancestry for replication.

Applying the PheRS to data from the EHR

The weights for each phecode phenotype were calculated independently for each cohort. For example, the phecode for proteinuria has a prevalence of 3.9% (853/21,701) in the discovery cohort and 8.6% (814/9,441) in the Marshfield cohort, so the weight for this phenotype is 1.41 and 1.06 respectively (see Table S13 for weights of each phecode for each cohort).

The HPO ontology provides phenotype annotations for 3,927 diseases described in OMIM and linked to one or more genes. Not all of these diseases, however, may be defined by a PheRS. We used only diseases that could be described by at least three unique phecodes and for which at least half of the HPO terms were mapped to unique phecodes. The latter restriction was designed to filter out diseases likely to be insufficiently described by the available phecodes. For example, 3-M Syndrome is annotated by 34 unique HPO terms, only 5 of which are mapped to unfiltered phecodes; the remaining terms describe characteristic facial and skeletal anomalies that are either unmapped or excluded because they pertain to birth defects and neonatal phenotypes. After applying these filters, 1,896 OMIM diseases can be profiled using PheRS linked to 1,667 (known or suspected) causal genes through OMIM (as of 09/01/2015). We further filtered this list to those for which we had relevant SNPs in our discovery cohort, as described below, resulting in a total of 1,204 Mendelian diseases profiled (see Table S14 for details on the Mendelian diseases tested).

Validating the PheRS for six Mendelian diseases

The phenotype risk score is intended to assign a value to an individual based on their similarity to a particular disease profile. To be effective, it needs to be both sensitive and specific. Applied to a population, a PheRS should rank individuals with disease A higher than those who do not have disease A (sensitivity). Furthermore, the PheRS based on disease A should not systematically rank individuals with disease B higher than those who do not have disease B (specificity). We tested this premise on six diseases: cystic fibrosis (CF), Marfan syndrome, phenylketonuria, achondroplasia, Li–Fraumeni syndrome, and hereditary hemochromatosis (HH). Phenylketonuria is unique among these six diseases because of a combination of prenatal screening and a highly effective treatment that essentially eliminates the effects of the disease. Following our analysis, we recognized that this disease effectively served as a negative control. These diseases were chosen in consultation with physician researchers based on the following criteria: (1) there were at least 20 clinically diagnosed adults in the Vanderbilt SD cohort of ~2.5 million (2) their feature set was rich enough to be profiled using PheRS and (3) their

clinical impact was high. These represent the only diseases evaluated for this phase of the project and were selected prior to any analysis.

For each of the six diseases, potential cases were identified using billing code and text phrase searches in clinical notes. All potential cases were then manually reviewed, and only true cases were used in the analysis. All individuals with a text mention of the disease or the causal gene in their notes or with a billing code for the disease were excluded from the possible list of controls. Cases were matched to controls by age, record length (both +/- one year), sex, and race; the ratio of controls to cases was 10:1 or more, depending on availability. Cases and matched controls were assigned a PheRS for the target disease (e.g. CF cases/controls were scored using the CF PheRS definition). We used Wilcox rank-sum to test the null hypothesis that the scores for cases and controls were the same. We then permuted the PheRS definitions for each case/control group, and applied the Wilcox rank-sum test to each combination of PheRS definitions and case-control groups (e.g. comparing the CF-based PheRS score in cases/controls for Marfan syndrome).

Following analysis, we manually reviewed the EHR data for the highest-scoring controls to see if they potentially had the disease in question or had a different genetic disease. High-scoring controls were defined as those controls with a PheRS greater than the third quartile PheRS for cases; for phenylketonuria, we reviewed the top ten scoring controls, since the PheRS was not elevated relative to cases. The review took place six months after the initial case/control sets were defined; information that had accrued in the Synthetic Derivative during that time was included in review, allowing us to find controls who were diagnosed subsequent to the original data pull. See Table S1 for the PheRS quartiles for cases and controls and the number of controls reviewed.

Testing for associations between the PheRS and rare variants in Mendelian genes

Our discovery cohort of 21,701 adults of European ancestry, as well as the two replication cohorts, were genotyped on the Illumina HumanExome BeadChip. The platform is designed to sample variants from the coding region of genes and captures 210,333 missense, 5,158 stop gain, and 9,263 synonymous variants in over 18,000 genes. To focus on rare variants, we excluded all variants with a minor allele frequency >1% in the discovery cohort. We required that each variant have at least 10 heterozygotes or homozygotes for the rare allele (roughly 0.02% minor allele frequency). We further filtered variants to those in genes associated with Mendelian conditions that could be used to create a PheRS. Variants with a missingness rate of >1% were excluded, and individuals with a missing genotype rate >1% were excluded. We further excluded 118 SNPs with Hardy-Weinberg equilibrium (HWE) $p<10^{-4}$ from all analyses. After filtering, we had 6,188 rare variants in 1,096 unique Mendelian genes. Most of the variants were non-synonymous.

We annotated the ExomeBead Chip variants using Variant Effect Predictor (VEP) on GRCh37.(*36*) Consequences of variants as described by the Sequence Ontology were ascertained by using the canonical transcript from Ensembl. In the case where multiple consequences were listed for a single variant/gene pair, the most severe consequence was chosen. The canonical transcript ID was used to annotate each variant according to HGVS nomenclature (See Table S15 for variants tested).

All Mendelian diseases in this study have previously been associated with one or more genetic variants. We matched the variants to Mendelian diseases suitable for PheRS profiling using the OMIM GeneMap. We only tested variants in genes that were known to cause the Mendelian disease that defined the PheRS profile (for example, we only tested the scores for cystic fibrosis against variants in *CFTR*). Although this strategy prevented us from finding associations between disease profiles and genes not already associated with a Mendelian disease, it preserved the statistical power necessary to ascertain the phenotype effects of rare variants. After filtering for genes with at least one testable SNP in our cohort, we conducted 7,520 association tests on 1,204 Mendelian diseases and 6,188 SNPs in 1,096 unique genes. Though many known pathogenic variants are autosomal recessive, all tests for association assumed a dominant mode of inheritance (see Table S16 for all association results in discovery cohort).

We manually reviewed the charts of all individuals who carried associated variants ($q<0.1$). To determine if they were diagnosed with the target Mendelian disease, we conducted text searches and reviewed problem lists and notes from clinic visits. Urinalyses for calcium oxalate were pulled for all heterozygotes and homozygotes for the *AGXT* variant.

PheWAS analysis

We conducted a PheWAS analysis on the 6,188 variants included in the discovery analysis using the Fisher's exact test. 1,734 phecodes were included in the PheWAS, all with at least ten cases. We excluded pregnancy-related phecodes given the lack of appropriate controls for these phenotypes, and the phecode for "pain in joint" due to phenotypic inflation. We used a false discovery rate on all 10,729,992 phenotype-variant association tests to detect significant associations; all p-values were greater than the false discovery rate $q<0.1$. We also calculated a Bonferroni correction for an individual PheWAS ($0.05/1{,}734=2.8\times10^{-5}$) to analyze the PheWAS results for significant variants found in the discovery analysis. Using custom R scripts, we produced grid plots and PheWAS Manhattan plots for each significant variant found the discovery analysis.

Severe endpoints of disease analysis

We reviewed the Mendelian diseases for significant variants in the discovery analysis and determined that there were three severe outcomes associated with these diseases that could be ascertained using claims data. These outcomes included liver transplant (V42.7), kidney transplant (V42.0), and thyroidectomy (06.4). We defined cases for liver transplant and kidney transplant as having at least four relevant billing codes on unique dates, given that transplants require extensive preparation and follow-up. Controls were defined as individuals with zero relevant codes; all other individuals were excluded. For the thyroidectomy cases, we required cases have only one because 06.4 is a procedure code that is frequently only billed for on the date of the procedure. We ascertained case/control status for the entire discovery cohort. We matched severe outcomes to genes on the basis of the Mendelian disease associated with the gene: liver transplant with *HFE*; kidney transplant with *AGXT*, *FAN1*, and *DGKE*; and thyroidectomy with *TG*. We hypothesized that individuals were more likely to have one of these outcomes when they had a variant in a relevant gene compared with wildtype individuals. We only analyzed variants found in our discovery analysis. P-values

comparing individuals with the rare allele versus background for outcomes were generated using a Fisher's exact one-sided text. (Table S4)

Replicating associations found in discovery analysis

We attempted to replicate our significant associations in two cohorts: a cohort of 9,441 adults of European ancestry from Marshfield Clinic, and a cohort of 3,820 adults of non-European ancestry from Vanderbilt. We only attempted replication when there were at least ten individuals who were heterozygous or homozygous for the target variant, as we had in the discovery cohort. This criterion was met for two variants in the Marshfield cohort (rs142698837 and rs150393409) and three variants in the non-European Vanderbilt cohort (rs116297894, rs13408961, and rs150393409). (Table S5 provides counts and summary statistic for variants for all significant associations found in the discovery cohort.)

Comparison of results to existing methods of determining pathogenicity

We tested whether the SNP type was correlated with functional categories (stop-gain, splice donor/acceptor, missense, splice region synonymous, intron/UTR) as determined by VEP using the Ensembl canonical transcript. For this test, we added results for variants that were non-exonic. These SNPs were not included in our reported results set but were useful as a comparison to the coding variants that were included. We filtered this expanded set of results by a nominal $p<0.05$ and created a boxplot of the betas categorized by functional type and grouped by Ensembl's impact rating. We conducted a pairwise comparison of the distribution of betas using the Wilcox rank-sum test.

We used ANNOVAR(*37*) to annotate the variants tested in the discovery analysis (version downloaded 2015-06-17). To test if our results aligned with functional predictions from various sources, we categorized the variants tested in the discovery cohort as significant ($q<0.1$), marginally significant (uncorrected $p<0.05$), and non-significant ($p>0.05$). We considered the variants from our significant results to be predicted as "deleterious" and the variants that were non-significant to be "tolerated." We used Fisher's exact to compare the annotations derived from our result set to 14 annotations. Categorical predictions (e.g. "Deleterious" versus "Tolerated) were available in ANNOVAR for 10 of these annotation databases. For the remaining four, we labeled the top quartile scores as deleterious (See Table S17 for details on annotations from ANNOVAR).

Statistical methods

Statistical analyses were conducted using Plink v1.90b3y(*38*) and R version 3.2.1 (http://www.r-project.org/). A linear regression with age and sex as covariates was performed for both the discovery and replication cohorts. All tests of significance used a dominant genetic model and assumed a two-tailed distribution. For the discovery cohort as well as the PheWAS analysis, the false discovery rate (FDR) was calculated using the Benjamini and Hochberg method. Significance was determined by an FDR of $q<0.1$. We found 18 associations with $q<0.1$ for 17 unique variants in our discovery set. Adding the first three principle components to the linear model did not significantly impact the results overall, and the same 18 associations remained $q<0.1$. Three variants were

available in the Marshfield cohort for replication and three in the non-European cohort. We used $p<0.05$ to define replication in these cohorts.

Calculating residual PheRS

We developed a means to assign each individual a PheRS score that could be compared between individuals and across diseases. In the discovery cohort, we characterized the burden of various phenotypes in groups of individuals by calculating the deviance residuals from the same regression models used in the primary analysis, but without the genotype in the model. Calculated this way, the "residual PheRS" (rPheRS) corresponds to how much an individual's PheRS deviates from what is predicted, given their sex and age. The rPheRS for person *i* is calculated as:

$$rPheRS_i = PheRS_i - E(PheRS)$$

Where E(PheRS | X) is estimated by the regression model PheRS ~ AGE + SEX (i.e., without genetics). We then represented each rPheRS by its deviation from the mean by standard deviation (z-score). While the magnitude of the PheRS and rPheRS is dependent on the number and rarity of symptoms for a particular disease, the z-score allows for comparison of PheRS across different diseases.

Whole exome sequencing sample selection

We undertook whole exome sequencing (WES) on a subset of individuals with significant variant from the discovery analysis. Our goals were to (1) confirm the variant call made with the Exome BeadChip, (2) identify variants in linkage disequilibrium with the target variant, and (3) identify other exonic or intronic variants which may be related to the disease phenotype. We restricted our analysis to the targeted genes of interest from the discovery analysis.

In the discovery phase of our analysis, we found 18 statistically significant associations with variants from 16 genes. In total, 1,401 individuals were either homozygous or heterozygous for the rare allele for a variant in Table 1 (807 from the discovery cohort and 594 from the VUMC replication cohort). 80% of these samples had DNA available for sequencing. We selected half of the genes with significant variants for WES: *AGXT*, *CHRNA4*, *DGKE*, *PLCG2*, *SH2B3*, *SPTBN2*, *SUOX*, and *TG*. These genes were selected by a group of researchers and clinicians on the basis of sample availability, lack of prior literature evidence, and clinical importance. Three of the genes had dominant inheritance patterns (*CHRNA4*, *SH2B3*, and *PLCG2*), and five had recessive inheritance patterns.

For six of the eight genes, the number of heterozygotes and homozygotes for the rare variant was small enough that we sequenced all available samples, including individuals with elevated and non-elevated PheRS. We also included the individuals from non-European cohort when available. Because there were many individuals with variants in *AGXT* and *TG*, we selected a sample of symptomatic (PheRS greater than the expected value) and asymptomatic individuals (PheRS less than or equal to expected value) for both European and non-European cohorts. For *AGXT*, we also included the single homozygous individual in the European cohort and three homozygous individuals from the non-European cohort. In all, we selected 132 samples for sequencing.

WES variant calling and quality control (QC)

Of the 132 samples we attempted to WES, 13 samples failed at various points along the pipeline. Nine had insufficient samples for sequencing, including the individuals with the highest PheRS for spinocerebellar ataxia and sulfocysteinuria. Thus we cannot exclude the possibility that the most symptomatic individuals with these variants also harbor additional second variants (as we found for *DGKE*, *AGXT*, *PLCG2*, and *CFTR*). An additional four samples failed to sequence. The remaining 119 samples were processed using the GATK pipeline.(*39*)

We carried out variant calling following best-practice procedures implemented in an in-house pipeline. We mapped raw paired-end reads to the reference human genome GRCh37 using BWA-0.7.4(*40*) with default settings, masked duplicates using picard-tools-1.92 (http://broadinstitute.github.io/picard/), and re-calibrated base quality scores using GATK-3.7-0(*41*). We then used the GATK HaplotypeCaller for joint variant calling across all 119 samples. We calibrated variant quality scores using VQSR and filtered out low-quality variants with VQSR<99.0.

Concordance between the variants called on the Exome BeadChip and WES was calculated using Plink. Before checking concordance, we filtered out variants that were monomorphic according to WES as well as variants with a missingness > 5%. The overall genotypic concordance rate for 42,077 variants was 99.9%. We also calculated the non-reference concordance with Plink and an in-house Perl script. For the 91,728 non-reference calls made by WES, the concordance rate with the Exome BeadChip was 99.1%. We calculated a non-reference concordance rate per individual and excluded three individuals with concordance < 95%.

We further checked the concordance rate for between the Exome BeadChip and WES for the variants in the target genes (Table S8). All variants except *AGXT* and *CHRNA4* had a concordance of 100%. Two individuals called as heterozygous for the rare variant in *AGXT* on the Exome BeadChip had discordant calls by WES (homozygous and no-call), and were excluded from subsequent analyses. The 17 individuals sequenced for the *CHRNA4* variant were excluded from subsequent analysis due to the variant's high missingness rate in WES and low concordance with the Exome BeadChip (this region may be difficult to cover by short read sequencing as it is also had low coverage in ExAC). We also excluded one *AGXT* heterozygote with a no call at the target variant. After exclusions based on the concordance QC step, we had 97 WES samples remaining in our analysis, 84 of which were from the discovery cohort.

WES variant curation

After VCF files were produced, all variants found in the eight genes of interest were annotated using ANNOVAR and VEP. Variants were curated in terms of their functional impact and ExAC allele frequencies. Clinical significance was retrieved from ClinVar. Variants in the 5' or 3' UTR regions which were covered by the sequencing were curated as well. Variants were labeled "of interest" if they were of moderate or high impact (missense, stop gain, or splicing donor/acceptor) and rare (Filtering MAF of < 1% in ExAC).

We found 45 variants in the target genes, not including the target variants. Five of these were rare variants. 36 variants were listed in ClinVar, with 34 of these variants

specified as benign or likely benign and two specified as "uncertain significance." The benign/likely benign variants all had filtering MAF > 1% in ExAC.

Three variants were assessed as "of interest" given our criteria of being high impact (nonsynonymous) and having a filtering MAF of < 1% in ExAC. One of these variants, rs151185188 (p.R381K), was a missense variant in *AGXT* with a filtering MAF of 0.12% in ExAC. This variant is in ClinVar as "uncertain significance." SIFT and Polyphen assess this variant as benign and tolerated, respectively. Two rare, nonsynonymous variants were found in *PLCG2*: rs72824905 is a missense variant (p.P522R) with a filtering MAF of 0.76% which is predicted to be benign/tolerated by SIFT and Polyphen. The other is 16:81953095 (p.R687S), a missense variant that is not found in ExAC and does not have an rsID number. This variant was predicted to be deleterious by SIFT and tolerated by Polyphen.

The more common of these *PLCG2* variants, rs72824905 (p.P522R), was found in every individual with the discovery variant (p.I251V) as well as others without elevated PheRS. p.P522R was genotyped on the Exome BeadChip and was not associated with the PheRS (linear regression adjusted for age and sex; p=0.23; p=0.54 excluding p.I251V heterozygotes). Based on this evidence, p.P522R was excluded as a variant of interest.

The remaining two variants of interest both occur in the most highly symptomatic individuals for their respective diseases (among those who carry at least one copy of the variant). The heterozygote for the *PLCG2* variant that was discovered through WES (p.R687S) had a z-score of 5.0, which is 2.5 S.D. higher than the next highest scorer among the ten heterozygotes for the variant in the discovery analysis (p.I251V). The individual with rs151185188 (p.R381K), the rare nonsense variant in *AGXT* found through WES, had a z-score of 6.0 and was again the highest scorer among the 36 European ancestry individuals with p.A295T, the next highest scorer in that set having a z-score 2.2 S.D. lower than this. Due to the rarity of many alleles observed in sequencing, correct assignment of alleles to haplotypes by statistical phasing algorithms is difficult. Consequently, we cannot definitively establish with those methods that the individuals with second variants in *AGXT* and *PLCG2* are compound heterozygotes. However, since the alleles are rare, the likelihood is that the two alleles are not in *cis* on the same chromosome because randomly occurring mutations are much more likely to impact different haplotypes than the same haplotype.

When available, we sequenced individuals from our non-European cohort to detect variant linkages not present in the European ancestry individuals. We did not find any additional rare, non-synonymous variants in the 13 individuals from the non-European cohort that we selected for WES (nine for *AGXT*, two for *TG* and two for *SH2B3*).

Selection of *SH2B3*, *TG*, and *SUOX* associations for biologic validation

After reviewing the novel associations for potential biologic validation, we selected three candidates that had local scientific interest and for which we could identify validation methods. These were the only variants we tested for in vitro validation. These methods are described below.

SH2B3 Materials and DNA plasmids

Antibodies were obtained from Santa Cruz Biotechnology (EPOR polyclonal M-20), Cell Signaling Technology (JAK2, GAPDH, and phospho-ERK 1/2), and Invitrogen

(anti-V5). V5 epitope-tagged constructs pcDNA3.1-SH2B3 (WT) and pcDNA3.1-SH2B3-R392E were kindly provided by Dr. Linyi Chen (National Tsing Hua University, Taiwan). pcDNA3.1-SH2B3-E395K was generated by site-directed mutagenesis, and the sequence of the entire insert was verified by Sanger sequencing. pRc/CMV-EPOR and pcDNA3 DNA constructs were generated as previously described.(*42*)

SH2B3 cell transfection, EPO stimulation and Western blot analysis

Subconfluent HEK293T cells were co-transfected with pRc/CMV-EPOR (0.29 ug) and pcDNA3-JAK2 (0.57 ug), and either pcDNA3.1-SH2B3-WT, pcDNA3.1-SH2B3-R392E, pcDNA3.1-SH2B3-E395K, or empty pcDNA3.1 vector (1.14 ug) using FuGENE 6 transfection reagent (Promega). After 48 h, cells were serum-starved for 4 h and then treated with 20 units/mL EPO (Amgen) for the indicated times. A time-course indicated that maximum levels of pERK were observed at ten minutes (Fig. S20). Stimulation was stopped by a cold phosphate-buffered saline wash, followed by protein extraction in SDS sample buffer. Equal amounts of protein were separated by 10% SDS-PAGE and transferred to PVDF membranes, which were probed with the indicated primary antibodies followed by HRP-conjugated secondary antibodies. ECL signals were detected and quantified using the FluorChem® SP imaging system from Alpha Innotech. Mean band intensities were compared using an unpaired two-tailed t-test, and represent the results of 4 independent experiments. Results were consistent whether pERK levels were normalized to EPOR, SH2B3, or GAPDH levels.

Splicing predictions for *SUOX* and *TG* variants

In silico splicing analysis was performed on all PheRS-identified variants except for *CFTR* using Human Splicing Finder (HSF) 3.0 and MaxEntScan prediction algorithms accessed on the following website (http://www.umd.be/HSF3/), by entering gene name and variant cDNA change and position.(*43*) Interpreted results and raw data tables were examined. Both HSF and MaxEnt algorithms predicted that the SUOX variant would break the native 5' donor site, with a 17-19% decrease in splice donor strength. The *TG* variant showed a 58% or 198% increase in strength as an acceptor site relative to the wildtype sequence with HSF and MaxEnt algorithms, respectively. Although the strength of this cryptic acceptor is relatively weak compared to the native site (scores of 3.87 and 9.4, MaxEnt), the MaxEnt algorithm did not predict any other 3' splice acceptor sites in this exon. *SUOX* and *TG* variants were selected for further investigation because of the concordance of HSF and MaxEnt predictions, and their potential to alter donor and acceptor splice sites.

Evaluating the effect on *SUOX* and *TG* variants on splicing in vitro

Wildtype and variant versions of *SUOX* exon 5 and *TG* exon 3 flanked by 100 bp of intron sequence were synthesized GenScript (Piscataway, NJ) and subcloned into minigene assay vector pET01 (MoBiTec GmbH, Göttingen, Germany) using SalI and XbaI restriction sites. Gene block sequences were derived from the following reference sequences: NM_000456.2, NM_003235.4, NG_008136.1, and NG_015832.1. Resulting constructs were sequence-verified and transiently transfected into HEK293T cells using the calcium-phosphate method. RNA was extracted 50 hours after transfection using the RNeasy mini kit (Qiagen), and first-strand cDNA synthesis was performed using 3.0 ug

of RNA in the SuperScript® III System (ThermoFisher). cDNA was diluted 10-fold and used in a touch-down PCR reaction containing primers specific to small exons within the pET01 vector, ETPR04, and ETPR05 (MoBiTec). Expected product sizes of fragments including the exon of interest (exon-included) were 250 bp for *SUOX* and 170 bp for *TG*; the fragment containing exons derived from the parent construct alone (exon-skipped) was 72 bp. Cells transfected with an unrelated expression plasmid, pIRES2-EGFP, were used as a negative control. PCR products were gel extracted and Sanger sequenced to identify which splice i soform they represented. Images were acquired using Bio-Rad's Quantity One® software and analyzed using ImageJ1 software.(*44*) An empirically-derived correction factor for normalizing band intensity to molecular mass was obtained by running experiment samples alongside a standard (Low DNA Mass Ladder, ThermoFisher). Corrected intensity values for each band were expressed as percent of the total intensity within each gel lane; results represent mean percent ± SEM from 5 independent experiments. P values from an unpaired two-tailed *t* test are reported.

Interpretation of variants using ACMG guidelines

We interpreted the variants in Table 1 according to ACMG guidelines(*32*). We used ExAC for minor allele frequencies and counts for heterozygotes and homozygotes.(*13*) For PP3 and BP4 evidence, we used REVEL scores exceeding 0.52 (specificity > 0.90) or below 0.26 (sensitivity > 0.90), respectively.(*11*) For PP2 evidence, we used ExAC constraint scores, considering z-scores >= 3.09 as evidence of a low rate of missense variation (applied only to variants in genes for which missense variation is a known mechanism of disease). We used the rules for combining criteria to classify variants based on existing evidence. We then added the evidence from this paper, assigning PS4 criteria to the statistically significant variants from Table 1, and PS3 evidence from the in vitro experiments (Table S9). In total, we changed the status of seven variants: three were changed from uncertain to pathogenic, two from likely benign to uncertain, one from uncertain to likely pathogenic, and one from likely pathogenic to pathogenic.

Fig. S1: Disease burden of novel associations identified in this study.

Bar chart of the number of individuals who are both carriers of a variant and have the phenotype of the diseases reported in the significant results table. Phenotypes pertaining to symptoms (e.g. Fatigue) were pruned from the chart, as well as those phenotype/variant pairs that did not show enrichment. The *CFTR* variants were also excluded because the individuals had all been diagnosed with cystic fibrosis. None of the individuals represented in this chart have been diagnosed with the corresponding Mendelian disease. *Left panel*: Each circle represents an individual who both carries the novel variant and has that phenotype, each part of an associated Mendelian disease. Solid-filled circles represent individuals with disease beyond expected by random chance. *Right panel*: The bars on the right panel represent the population attributable fraction for each variant/phenotype pair.

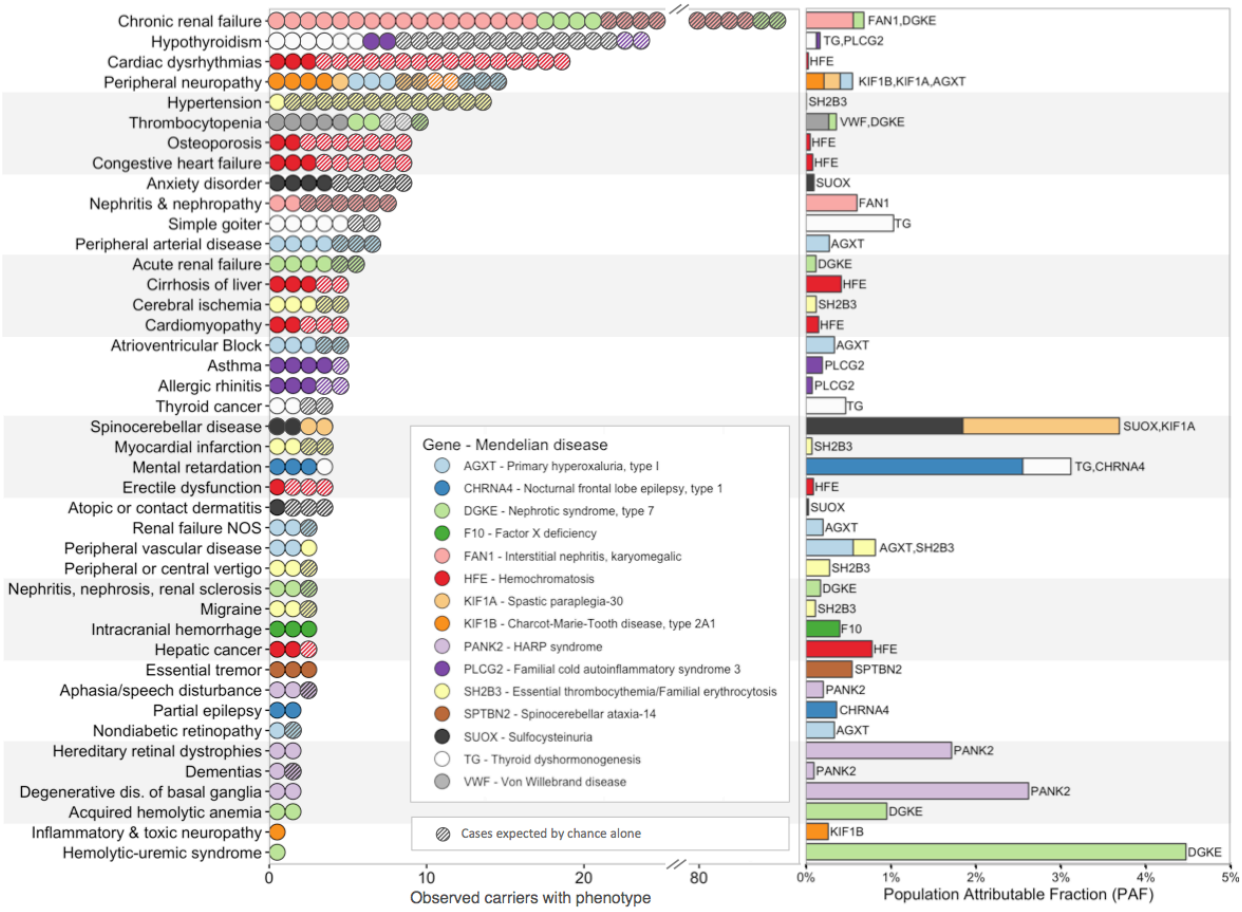


Fig. S2-17: Phenotype grids and PheWAS plots for significant variants identified in the discovery analysis.

For phenotype grids S2A-S17A, each column represents an individual in the discovery cohort who is heterozygous or homozygous for the specified variant; each row represents a feature of the Mendelian disease. Black squares indicate an individual has the phecode specified in the row label. The bar to the left of the grid indicates the relative risk of the phenotype among those displayed in the grid compared with wildtype individuals. The grid for S5A cannot be displayed due to the large number of heterozygotes. Figures S2B-S17B show PheWAS plots for each variant. A point represents a single association test for a phenotype. The y-axis indicates the –log(P) from the association using Fisher's exact test. The constituent phenotypes that define the PheRS are starred. Constituent phenotypes as well as those 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. S2: Phenotype grid and PheWAS plot for Primary Type 1 Hyperoxaluria - rs13408961 (*AGXT*, p.A295T)

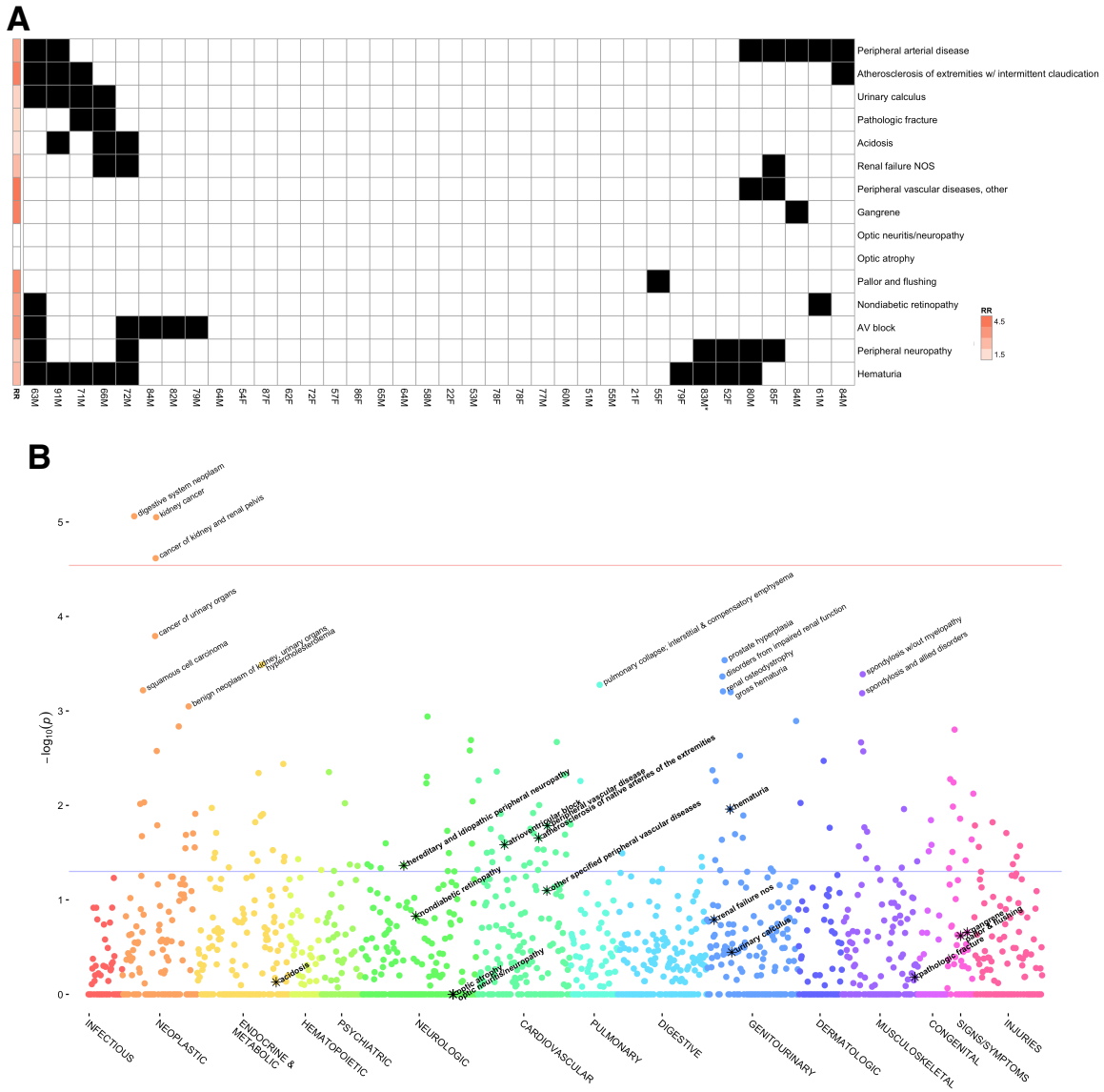


Fig. S3: Phenotype grid and PheWAS plot for Cystic Fibrosis - rs74597325 (*CFTR*, p.R553*)

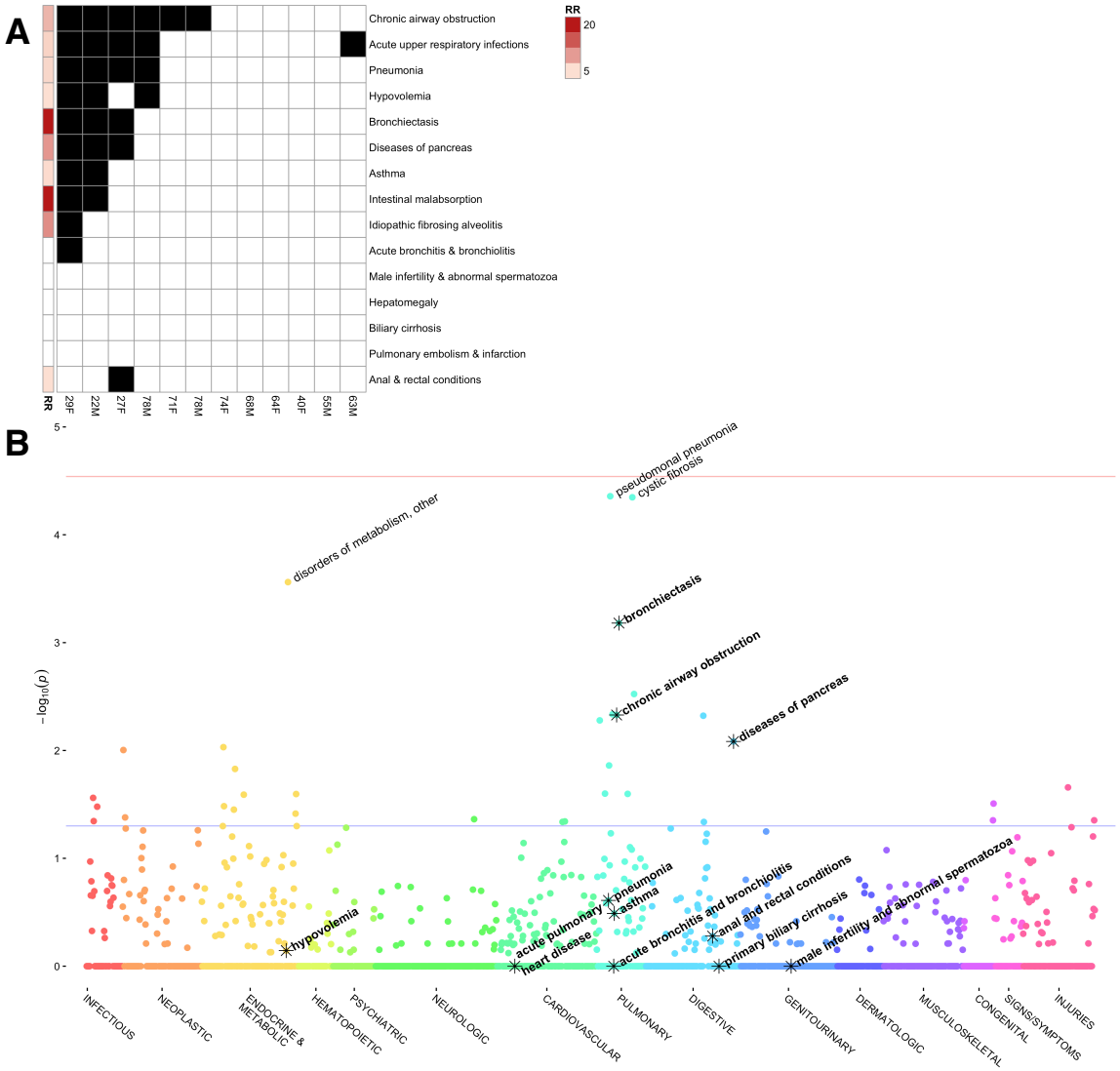


Fig. S4: Phenotype grid and PheWAS plot for Nocturnal frontal lobe epilepsy - rs55855125 (*CHRNA4*, p.R483Q

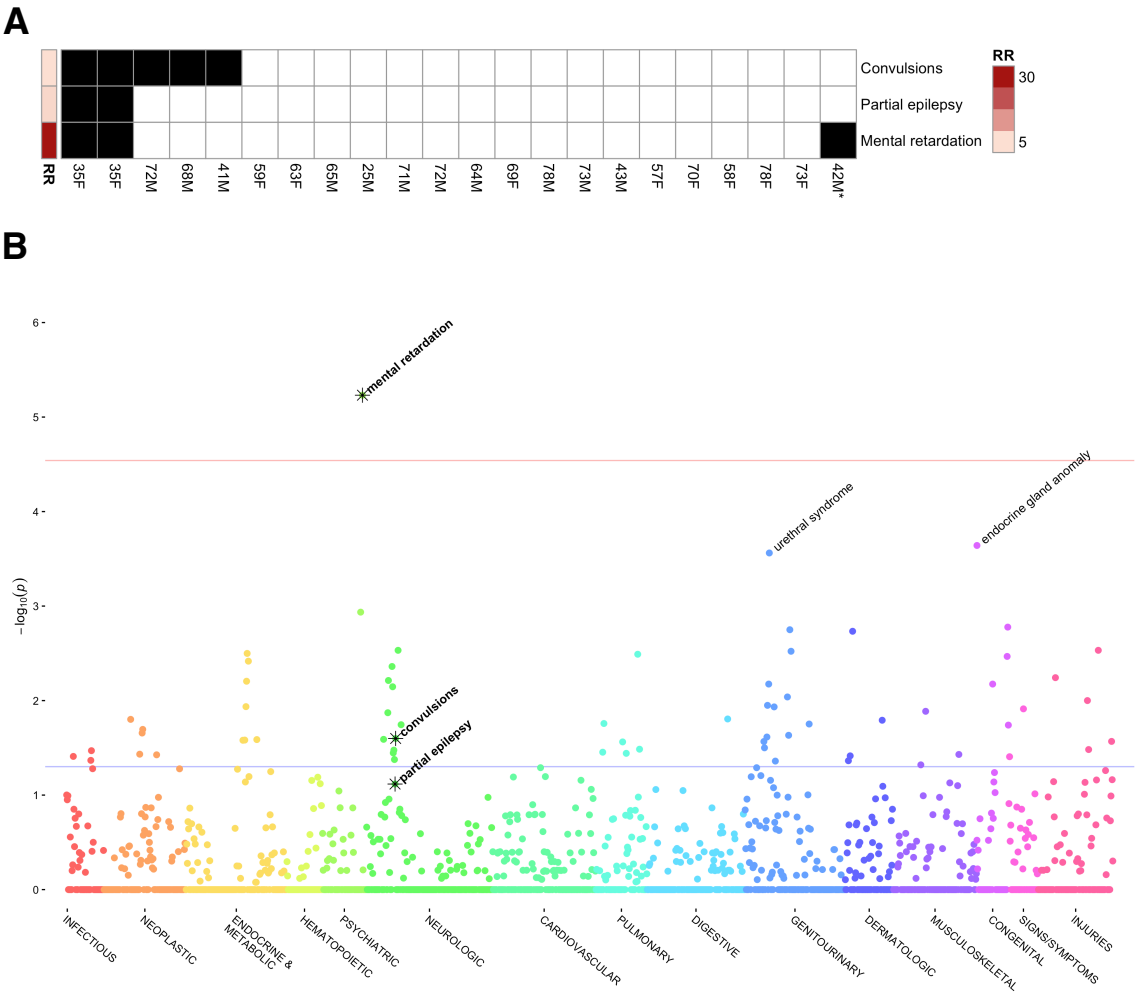


Fig. S5: Phenotype grid and PheWAS plot for Factor X deficiency – rs149212700 (*F10*, p.R291Q)

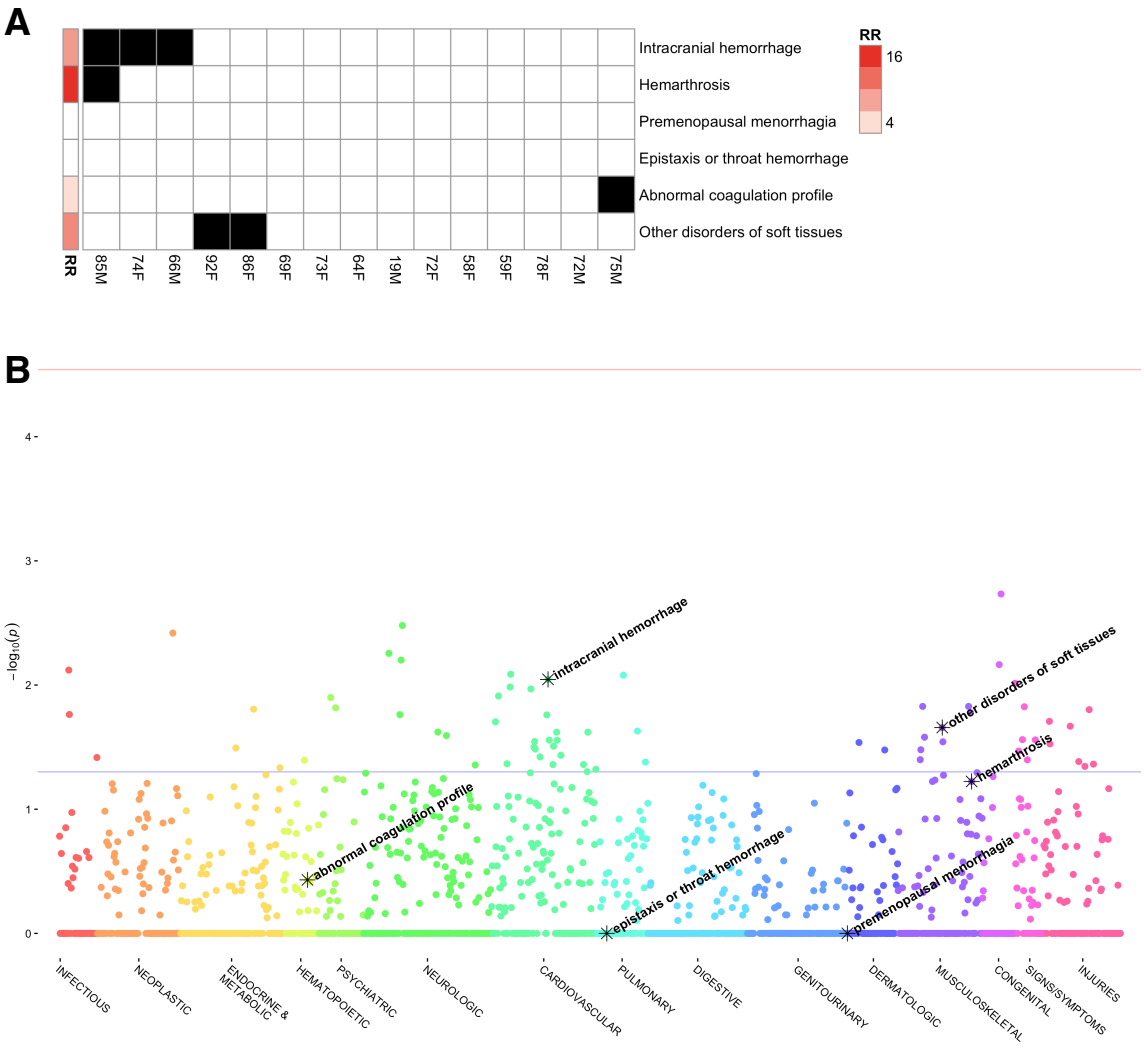


Fig. S6: Phenotype grid and PheWAS plot for Interstitial nephritis, karyomegalic – rs150393409 (*FAN1*, p.R507H)

A

Grid not available due to large number of carriers

B

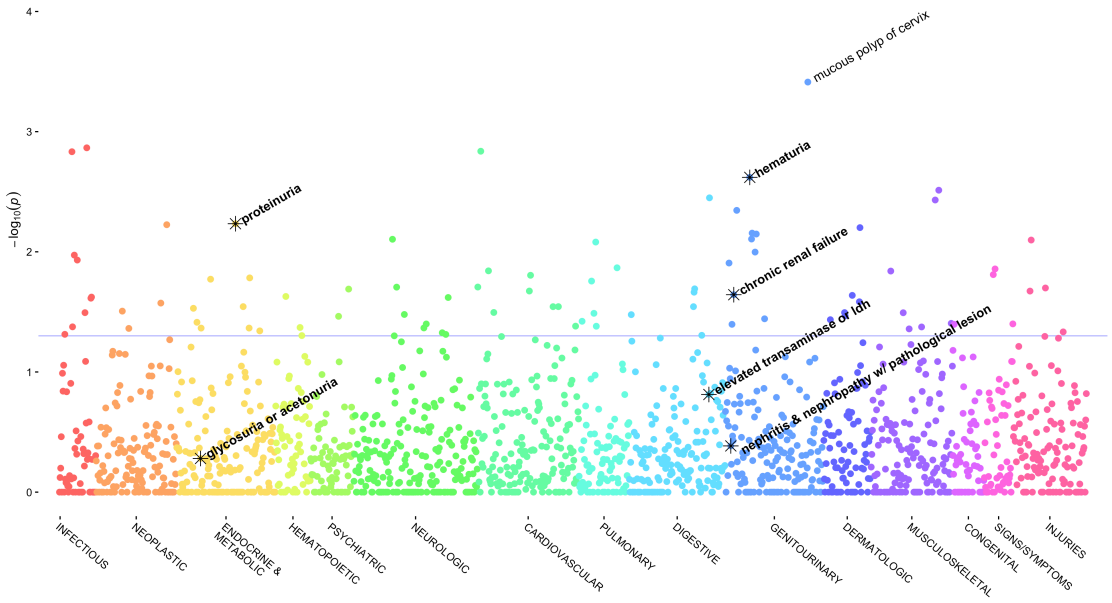


Fig. S7: Phenotype grid and PheWAS plot for Hemochromatosis - rs146519482 (*HFE*, p.E168Q)

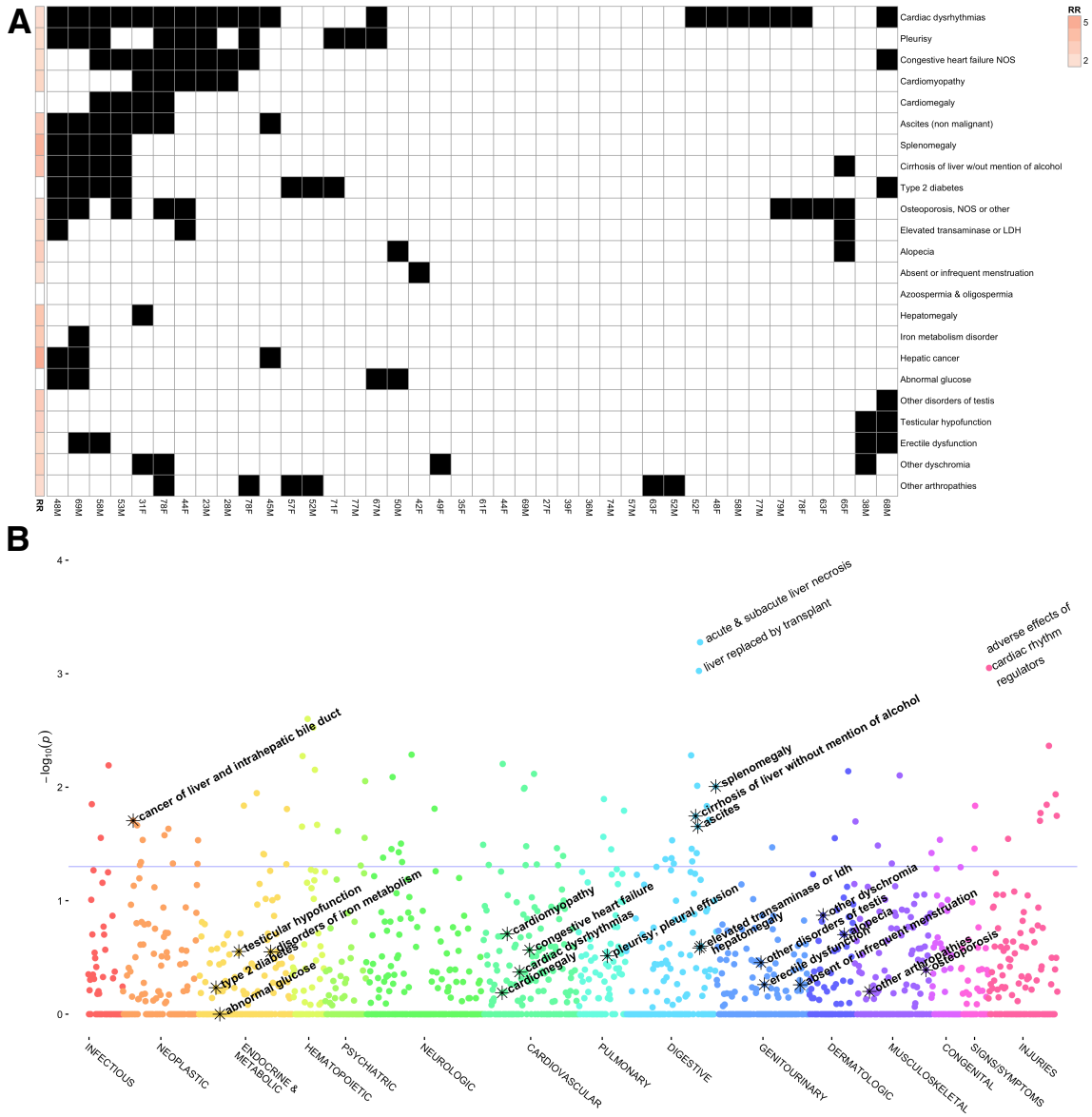


Fig. S8: Phenotype grid and PheWAS plot for Spastic paraplegia 30 – rs116297894 (*KIF1A*, p.A993A)

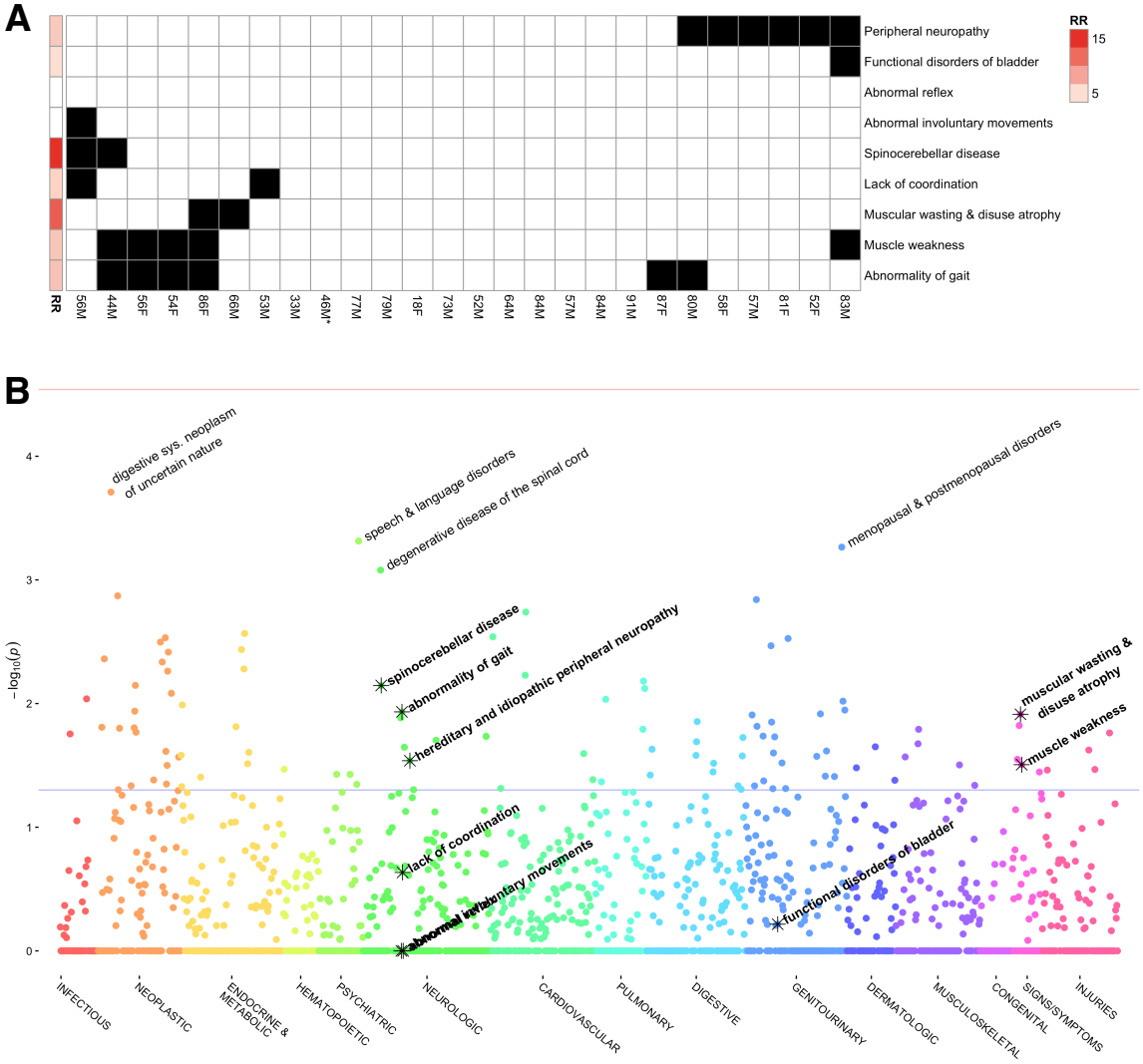


Fig. S9: Phenotype grid and PheWAS plot for Charcot-Marie-Tooth disease, type 2A1 – rs41274468 (*KIF1B*, p.T674I)

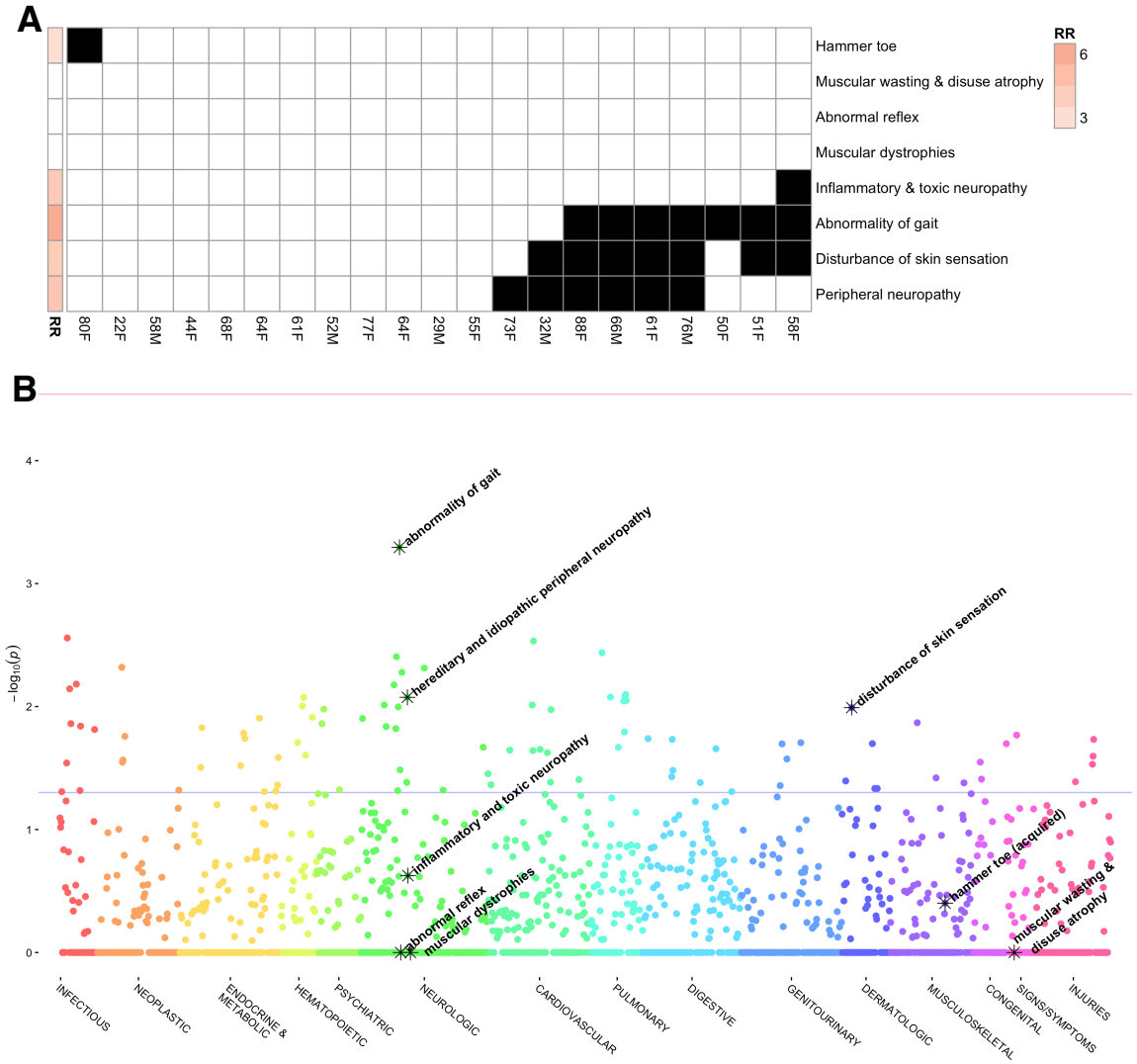


Fig. S10: Phenotype grid and PheWAS plot for Hypoprebetalipoproteinemia, acanthocytosis, retinitis pigmentosa, & pallidal degeneration – rs137852959 (*PANK2*, p.G521R)

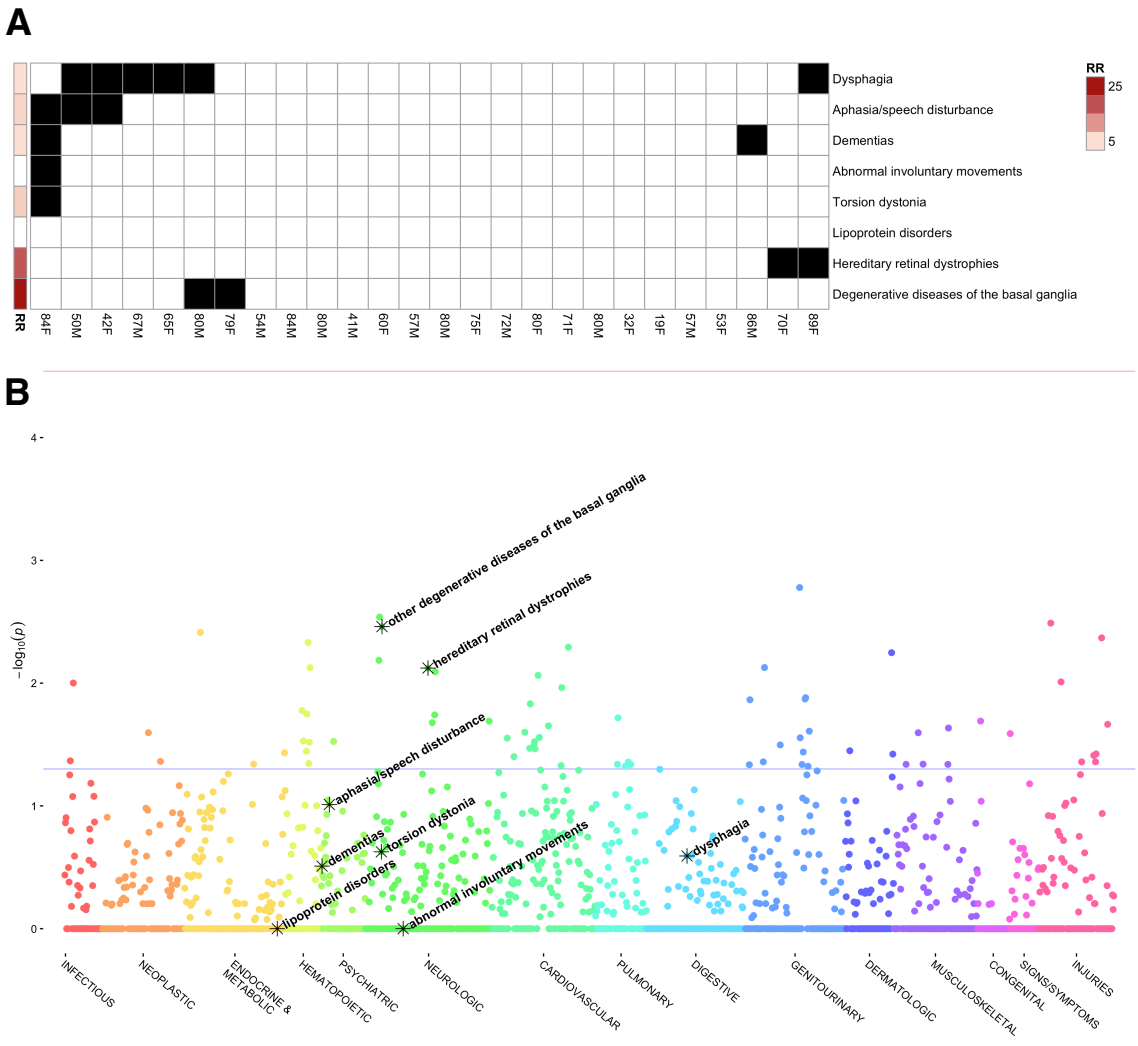


Fig. S11: Phenotype grid and PheWAS plot for Familial cold autoinflammatory syndrome 3 – rs190840748 (*PLCG2*, p.I251V)

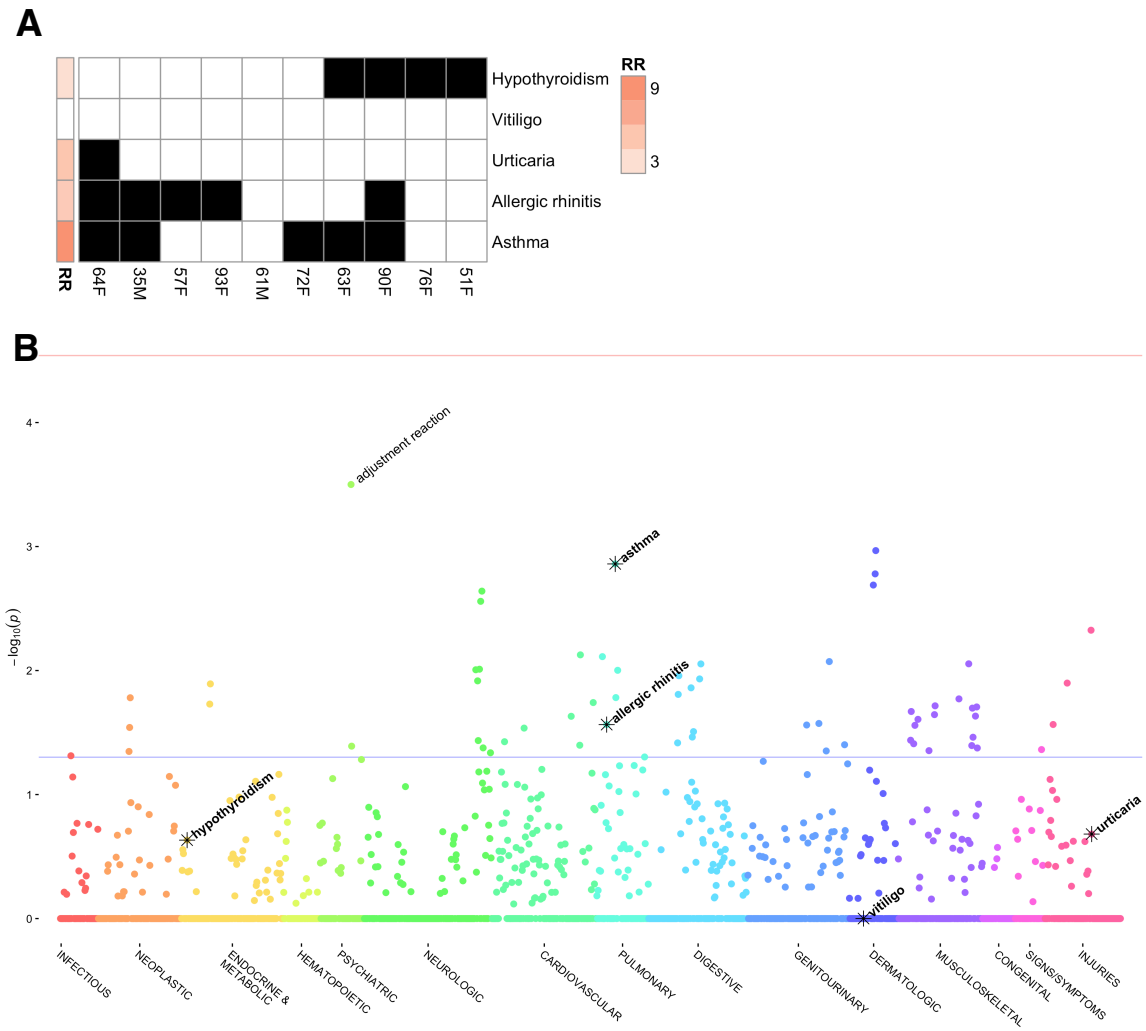


Fig. S12: Phenotype grid and PheWAS plot for Familial erythrocytosis – rs148636776 (*SH2B3*, p.E395K)

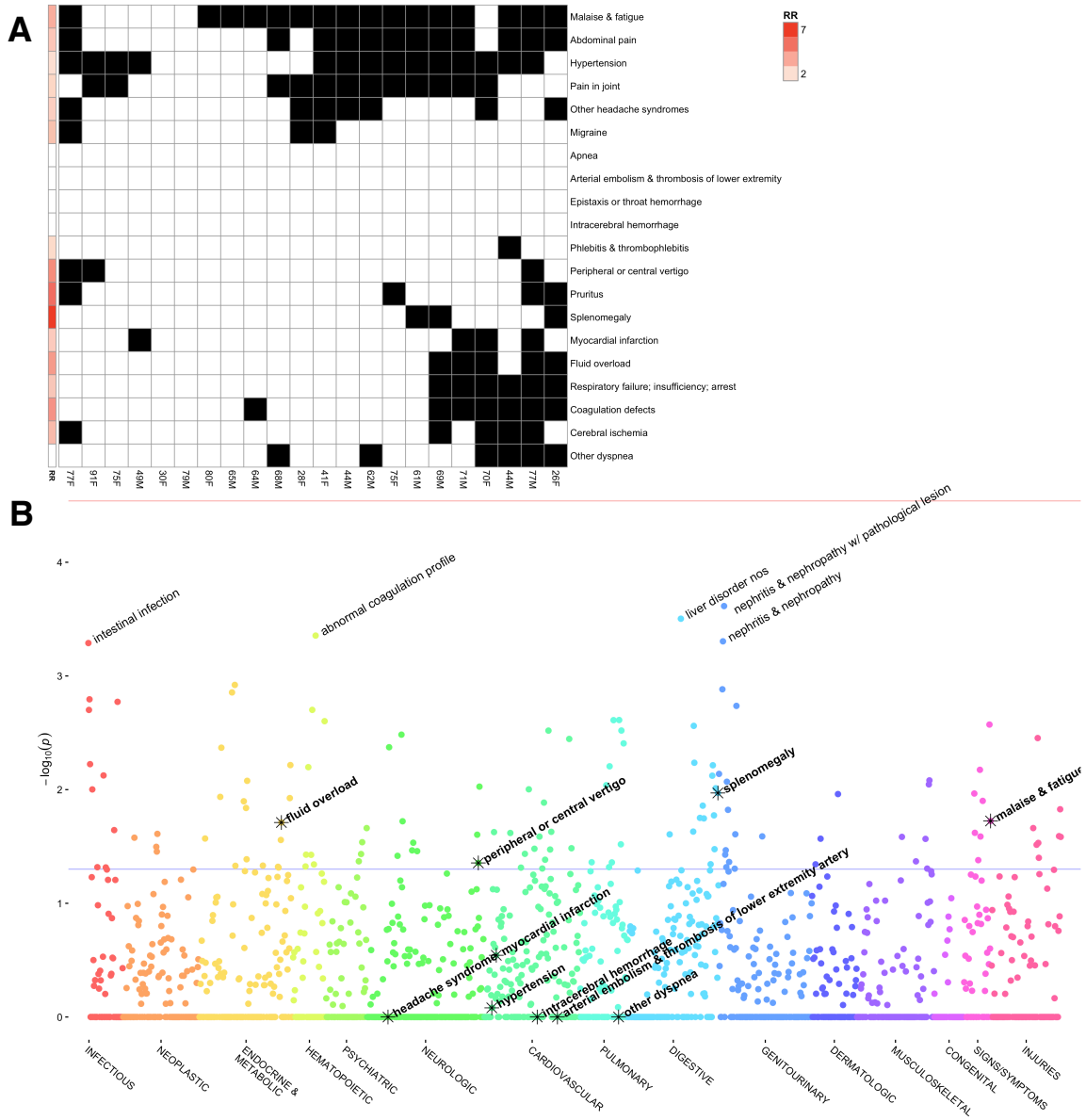


Fig. S13: Phenotype grid and PheWAS plot for Essential thrombocythemia – rs148636776 (*SH2B3*, p.395K)

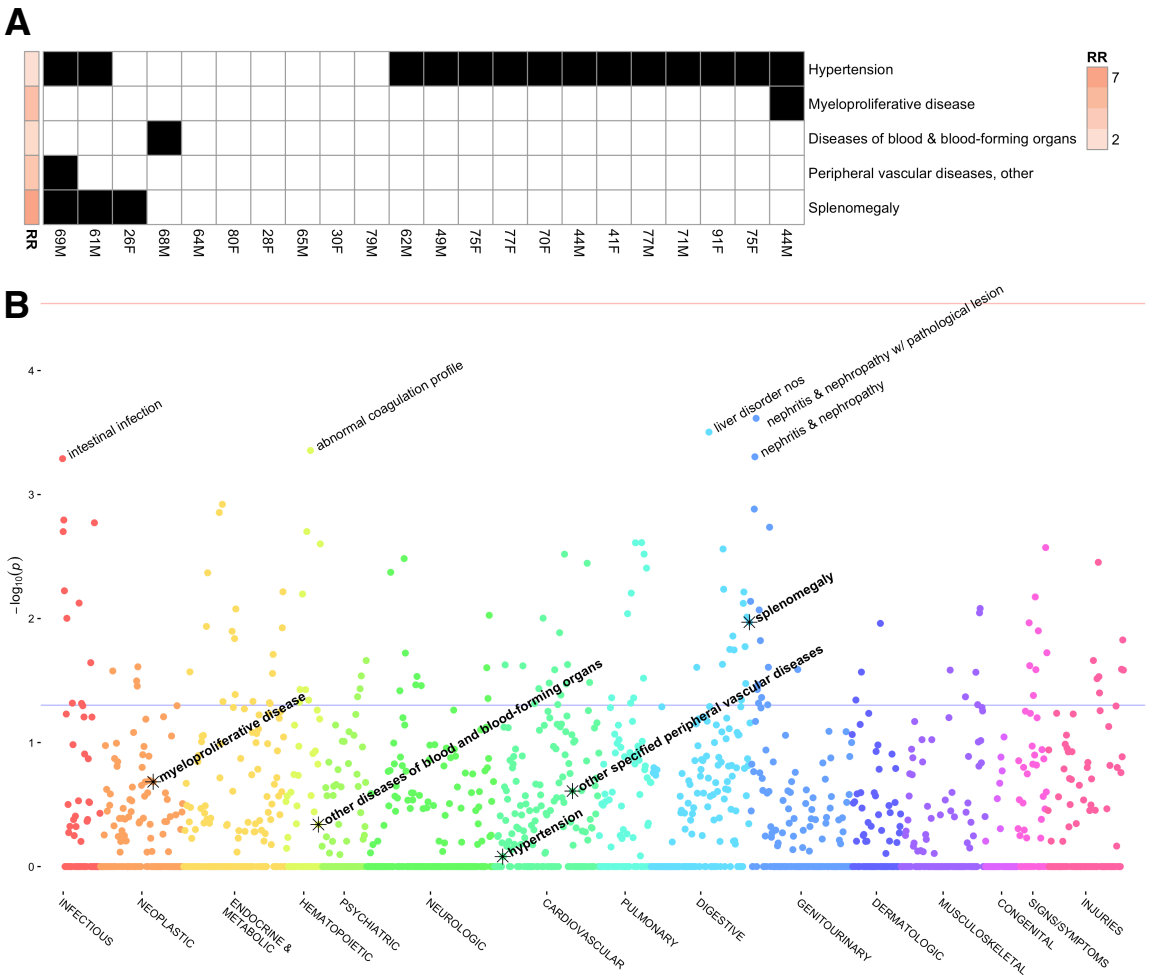


Fig. S14: Phenotype grid and PheWAS plot for Sulfocysteinuria – rs202085145 (*SUOX*, p.R76S)

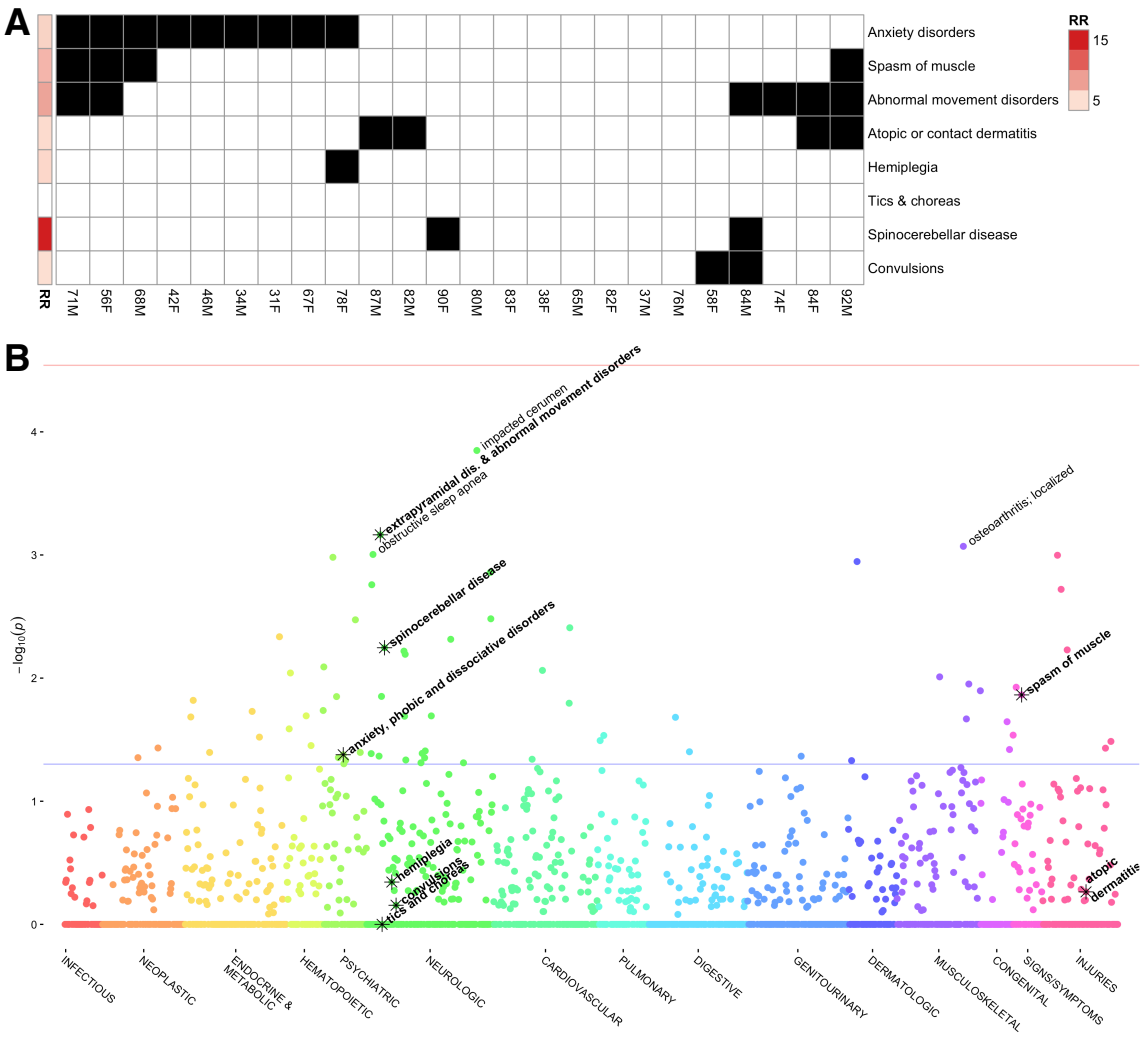


Fig. S15: Phenotype grid and PheWAS plot for Spinocerebellar ataxia, autosomal recessive 14 – rs145522851 (*SPTBN2*, p.R2370H)

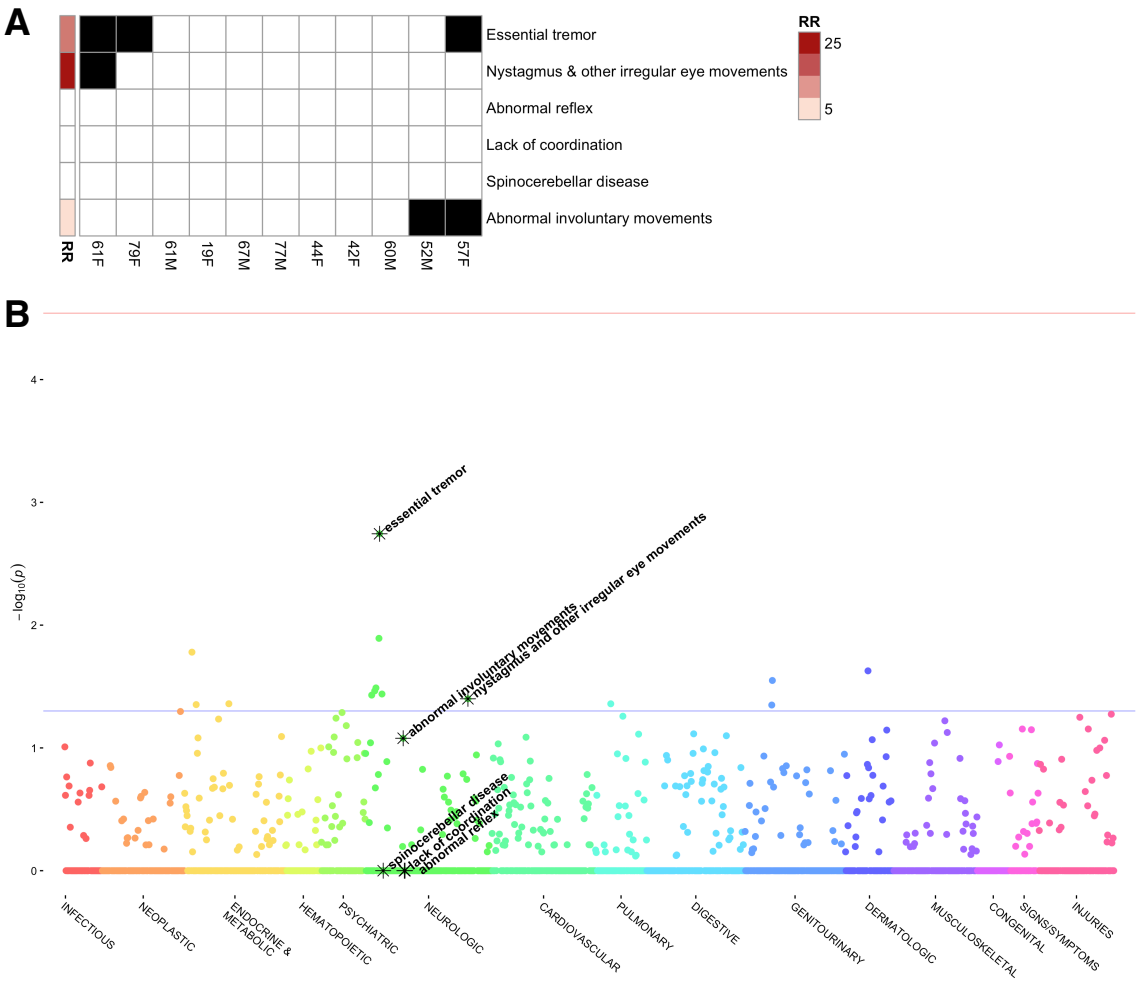


Fig. S16: Phenotype grid and PheWAS plot for Thyroid dyshormonogenesis – rs142698837 (*TG*, p.G77S)

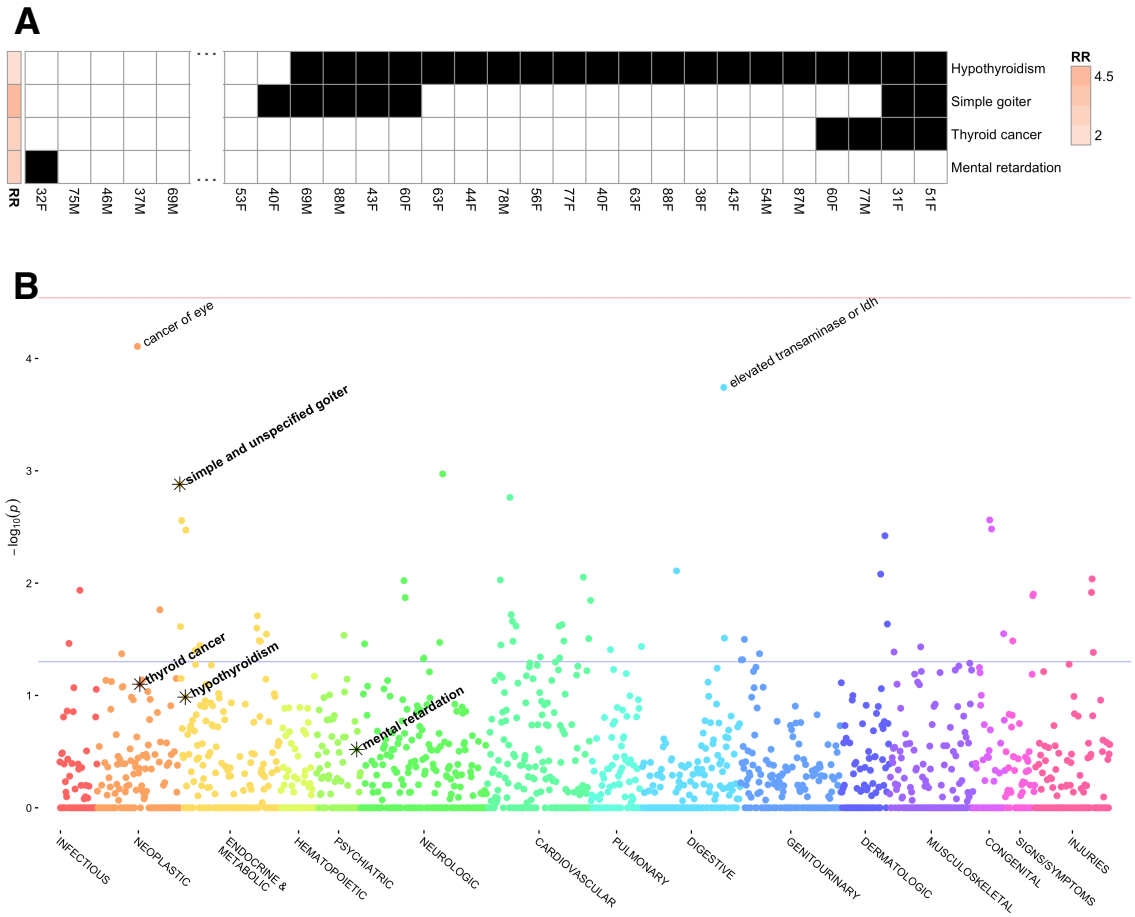


Fig. S17: Phenotype grid and PheWAS plot for Von Willebrand disease – rs144072210 (*VWF*, p.T1951A)

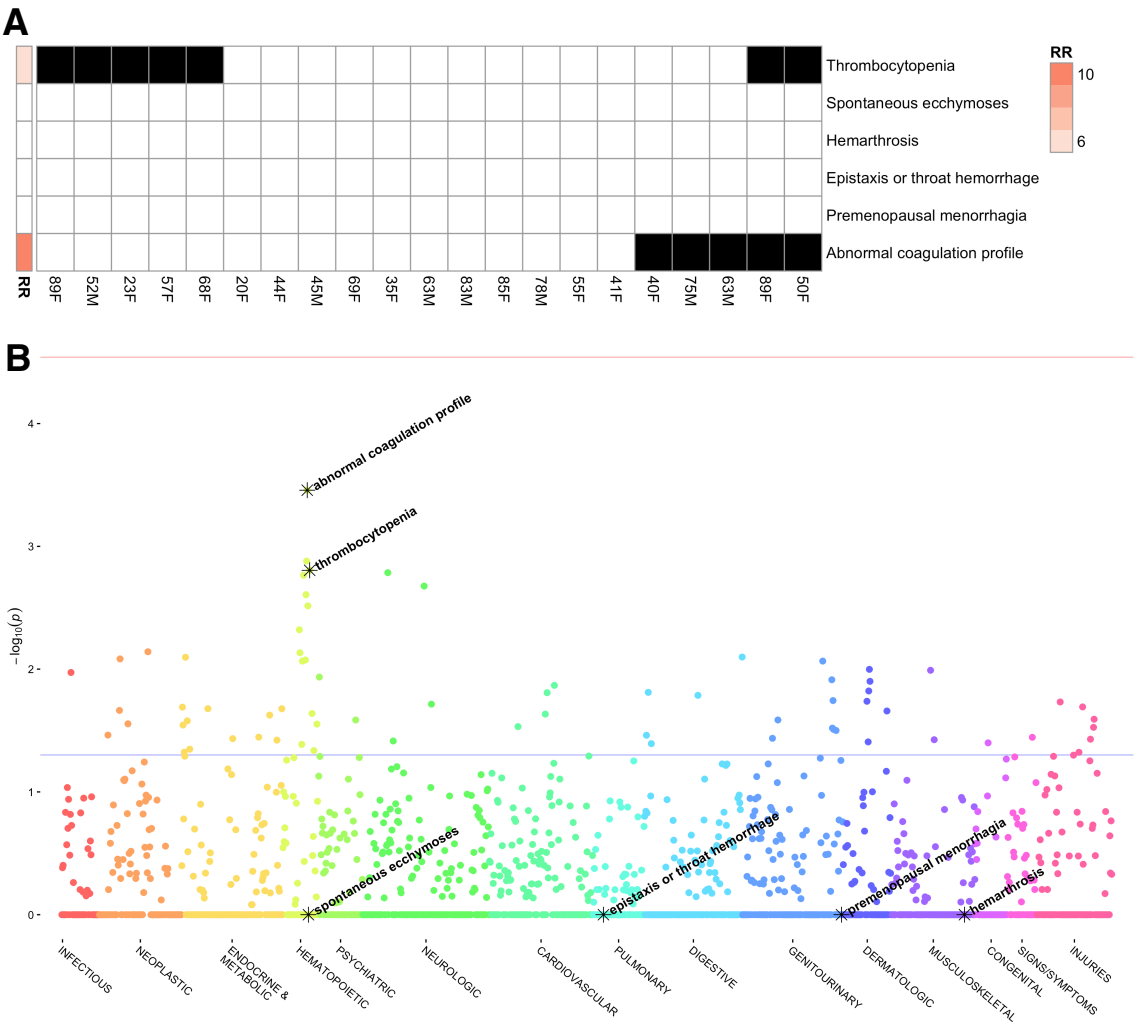


Fig. S18: Disease burden of novel associations identified in this study.

Each point represents and individual who is heterozygous or homozygous for a significant variant from the discovery analysis. The x-axis shows the z-score for residual from the PheRS for the disease paired with the variant label to the left; the established inheritance mode is in parentheses and mean z-scores for variant carriers are under gene label. PheRS is tested by linear regression assuming a dominant model adjusted for age and sex. Individuals in the WES analysis are triangles; all others are circles. Findings from chart review and WES are labeled. Homozygotes confirmed with WES are labeled HOM.

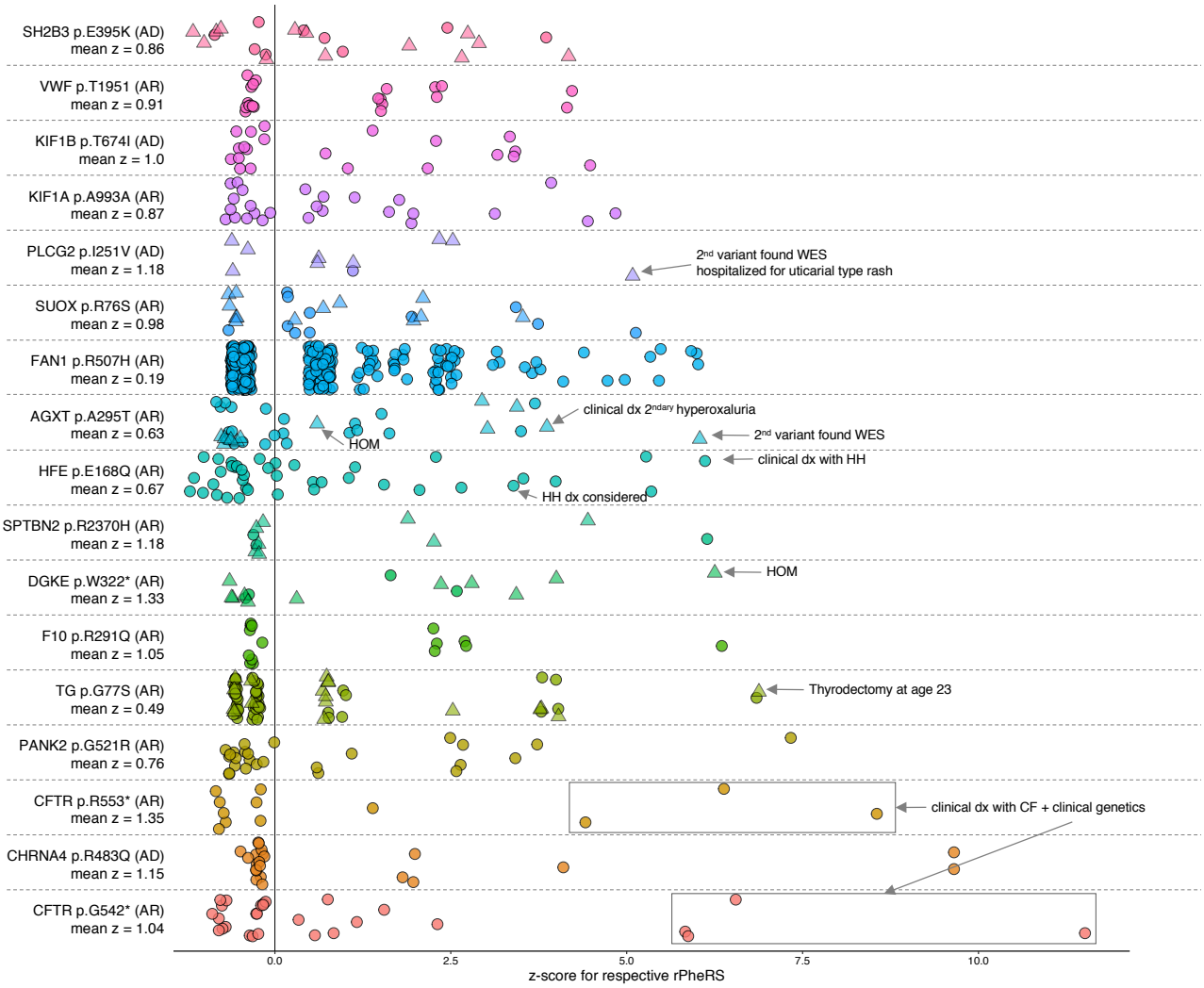
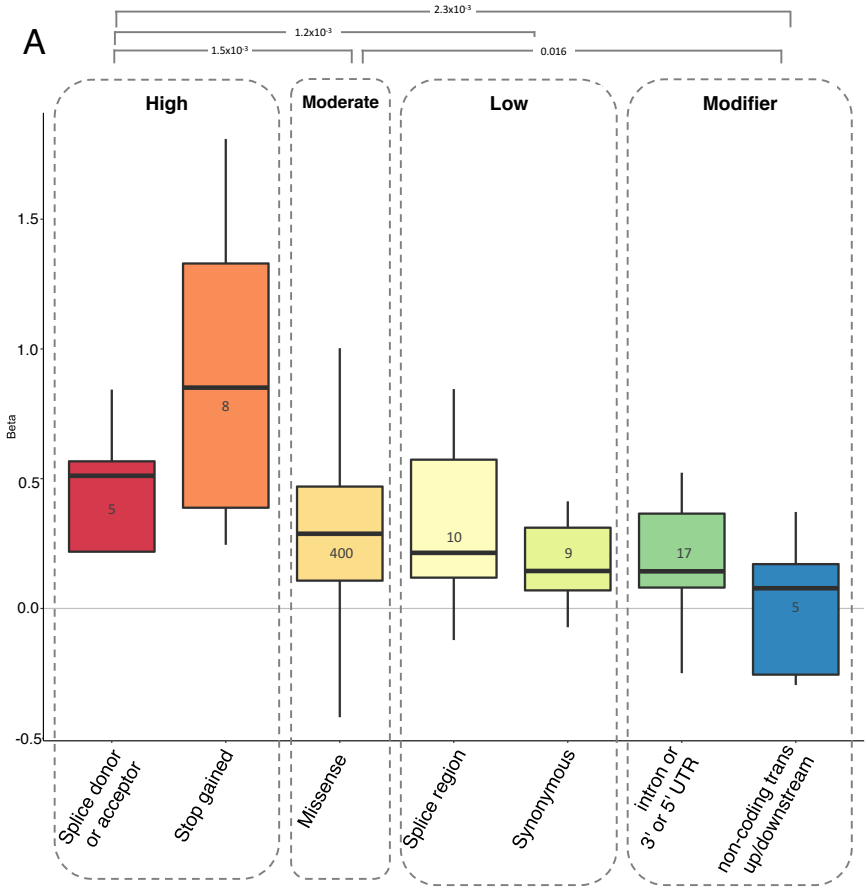


Fig. S19: PheRS associations based on variant function and predicted pathogenicity. (**A**) Boxplots of effect size (beta) for various types of variants. Only variants of nominal significance in the discovery analysis ($p<0.05$) are included. Consequences were derived from the Ensembl canonical transcript using VEP, and are ordered by the impact rating provided by Ensembl. Pairwise statistical significance between variants of different impact was evaluated by Wilcoxon rank-sum test. (**B**) Relationship between statistical significance of association with the relevant PheRS and predicted effect of variant according to various computational methods. We compared our classification of "significant" "non-significant" with predictions from a variety of source, using Fisher's exact. The plot displays the variants that were "weakly significant," though the p-values were generated by comparing "significant" with "non-significant" associations.



Impact 1	Impact 2	P
High	Moderate	$1.5x10^{-3}$
High	Low	$1.2x10^{-3}$
High	Modifier	$2.3x10^{-5}$
Moderate	Low	0.325
Moderate	Modifier	0.016
Low	Modifer	0.400

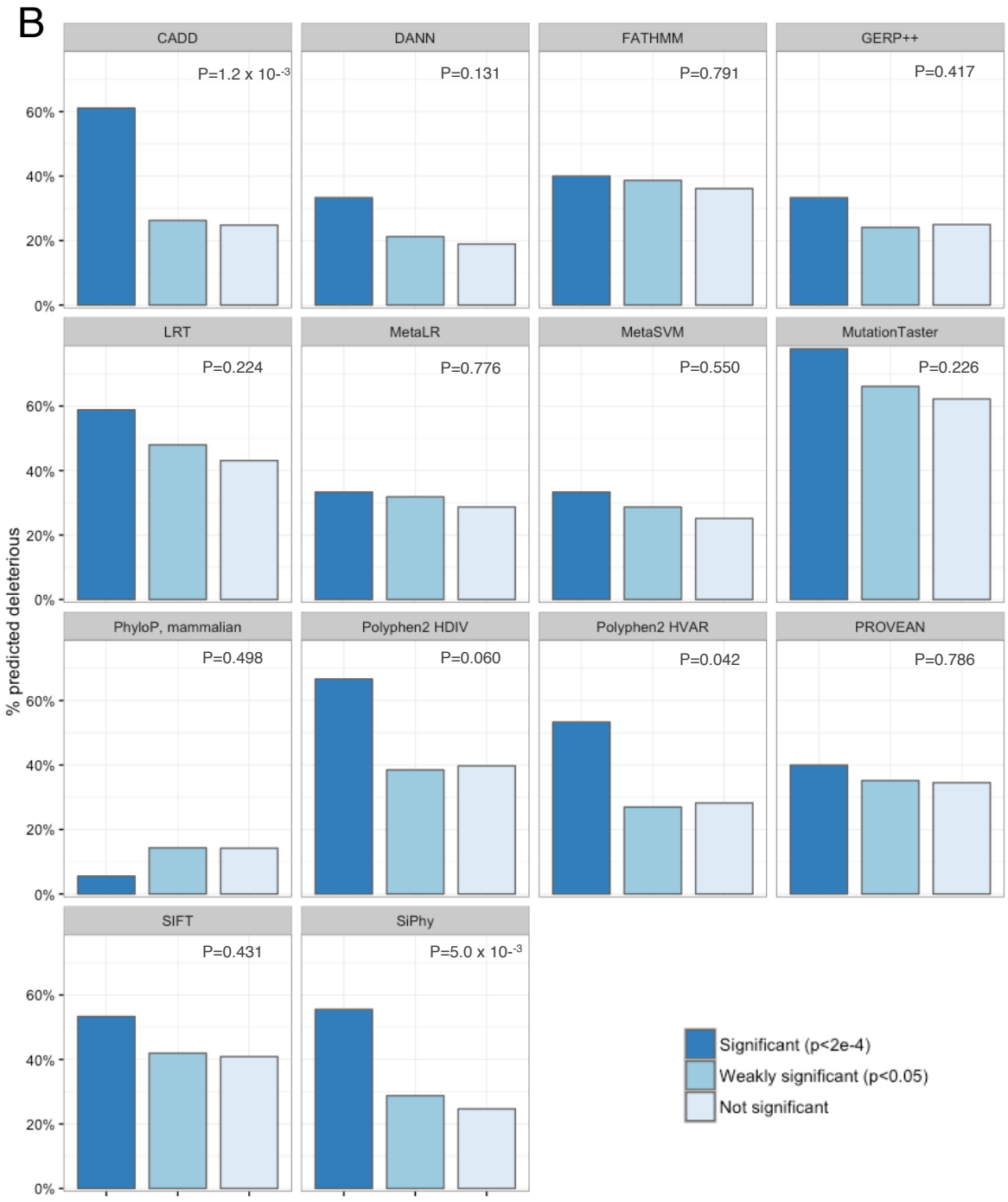

B
CADD
P=1.2 x 10-3
DANN
P=0.131
FATHMM
P=0.791
GERP++
P=0.417
LRT
P=0.224
MetaLR
P=0.776
MetaSVM
P=0.550
MutationTaster
P=0.226
PhyloP, mammalian
P=0.498
Polyphen2 HDIV
P=0.060
Polyphen2 HVAR
P=0.042
PROVEAN
P=0.786
SIFT
P=0.431
SiPhy
P=5.0 x 10-3
% predicted deleterious
60%
40%
20%
0%
Significant (p<2e-4)
Weakly significant (p<0.05)
Not significant

Fig. S20: Timecourse of ERK activation.

Western blot of HEK293T cells transiently transfected with wildtype (WT) versus variant SH2B3 constructs, EPOR, and JAK2 or empty vector. Cells were stimulated with erythropoietin (EPO) for the times indicated at top, and western blotting was performed with the antibodies indicated at left.

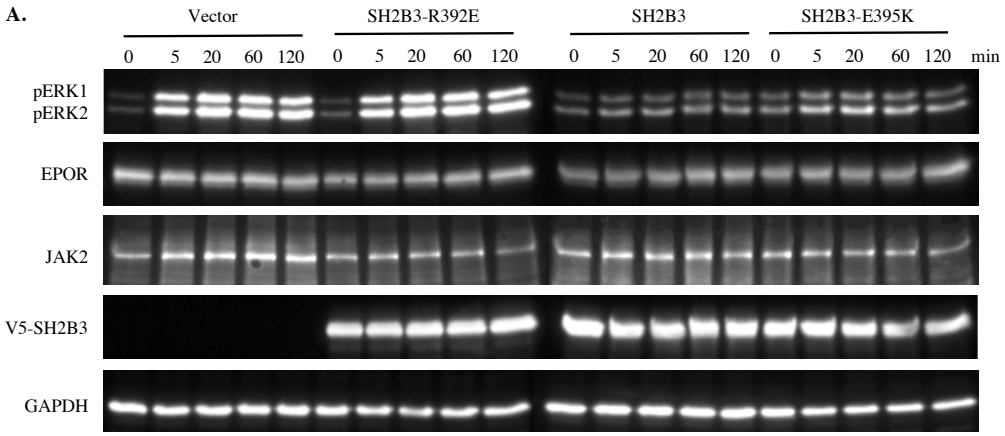


Table S1: Demographic summary for three cohorts.

Ancestry was determined by genetics. Age is the age at last visit available in the EHR. Years recorded is calculated as the age at last visit minus the age at first, plus one.

Site	*Ancestry*	*cohort size*	*%male*	*median age*	*median yrs. recorded*
Vanderbilt	European	21,701	44%	64	11
Marshfield Clinic	European	9,441	41%	61	36
Vanderbilt	Non-European	3,820	36%	56	12

Ascertainment variables for BioVU cohorts. The BioVU cohorts were ascertained for five different purposes. The largest portion of samples were genotyped for a variety of drug side-effect and/or efficacy studies. Two cohorts were ascertained based on rich EHR data: longitudinal and elderly. Another group was ascertained based on the presence of their samples from the tumor registry. Finally, a set of individuals were ascertained because they had a rare disease or condition as determined by billing codes.

Set	*BioVU discovery cohort*	*BioVU replication cohort*
pharmacogenomic studies	5,559 (26%)	1,175 (31%)
longitudinal	4,444 (20%)	9,17 (24%)
cancer registry	4,436 (20%)	6,38 (17%)
elderly	4,030 (19%)	4,81 (13%)
rare	3,232 (15%)	6,09 (16%)
ALL	*21,701*	*3,820*

EHR Race for BioVU non-European cohort.

EHR Race	*Count*
Black	2,684
White	490
Asian	339
Unknown	272
Indian American	30
Pacific Island	5

Table S2: Manual review of *CFTR* carriers.

Manual chart review revealed that 7 of 40 individuals with one or two copies of the rare variant G542* or R553* were clinically diagnosed with cystic fibrosis. All of the diagnosed individuals had genetic testing as part of their clinical care. Five were found to carry the ΔF508 mutation in addition to G542* or R533*, making them compound heterozygotes. Another was homozygous for G542*. A final individual was confirmed as a heterozygote for G542*, but a second variant was not identified via targeted genotyping and he passed away before higher resolution testing was conducted. The remaining 32 variant carriers had no clinical diagnosis of CF or relevant clinical genetic testing. Genetic clinical testing for *CFTR* was fully concordant with genotyping from the Exome BeadChip.

rs74597325 (R553*)

Subj. sex-age	CF dx?	PheRS	z-score	Genetics from Exome BeadChip	Genetics (from clinical testing)
F-29	Y	12.24	8.61	p.R553* heterozygote	ΔF508/R553*
M-19	Y	9.28	6.44	p.R553* heterozygote	ΔF508/R553*
F-27	Y	6.73	4.45	p.R553* heterozygote	ΔF508/R553*

rs113993959 (G542*)

Subj. sex-age	CF dx?	CF PheRS	z-score	Genetics from Exome BeadChip	Genetics (from clinical testing)
M-18	Y	16.03	11.6	p.G542* heterozygote	Unknown/G542*
F-21	Y	9.52	6.58	p.G542* homozygote	G542*/G542*
F-20	Y	8.61	5.90	p.G542* heterozygote	ΔF508/G542*
F-24	Y	8.58	8.58	p.G542* heterozygote	ΔF508/G542*

Table S3: Severe outcomes associated with variants.

Associations between variants from the discovery analysis and three outcomes were tested using a one-tailed Fisher's exact test. Genes that were matched with outcomes are highlighted in yellow. The three outcome/gene pairs that yielded a $p<0.05$ are starred.

		Kidney transplant		Liver transplant		Thyroidectomy	
Gene	**Variant**	**Odds ratio**	**P**	**Odds ratio**	**P**	**Odds ratio**	**P**
AGXT	p.A295T	4.58	6.9×10^{-3}*	2.25	0.37	0.00	1.00
CFTR	p.G542*	1.65	0.35	0.00	1.00	0.00	1.00
CHRNA4	p.R483Q	0.00	1.00	0.00	1.00	2.57	0.33
DGKE	p.W322*	4.40	0.09	0.00	1.00	0.00	1.00
F10	p.R291Q	0.00	1.00	0.00	1.00	3.77	0.24
FAN1	p.R507H	0.99	0.56	1.31	0.30	0.78	0.77
HFE	p.E168Q	0.00	1.00	8.10	2.1×10^{-3}*	1.42	0.51
KIF1A	p.A993A	2.54	0.20	3.11	0.28	2.18	0.38
KIF1B	p.T674I	0.00	1.00	3.86	0.24	0.00	1.00
PANK2	p.G521R	3.81	0.05	0.00	1.00	0.00	1.00
PLCG2	p.I251V	0.00	1.00	8.10	0.13	5.66	0.18
SH2B3	p.E395K	3.00	0.16	3.68	0.25	2.57	0.33
SPTBN2	p.R2370H	0.00	1.00	0.00	1.00	0.00	1.00
SUOX	p.R76S	0.00	1.00	0.00	1.00	0.00	1.00
TG	p.G77S	1.44	0.36	2.35	0.22	3.28	0.04*
VWF	p.T1951A	3.14	0.15	0.00	1.00	5.39	0.06

Additional Data table S4-S17 are provided in a separate excel file:

Table S4: Demographics of variant carriers in discovery and replication cohorts.

Table S5: Mean PheRS for carriers, heterozygotes, and homozygotes in discovery and replication cohorts.

Table S6: Results from replication analysis of novel genetic associations in two independent cohorts.

Table S7: Whole exome sequencing sample selection and QC information.

Table S8: Summary of variants identified in WES analysis.

Table S9: Complete report of variants found in WES.

Table S10: Variants from discovery analysis reinterpreted using ACMG guidelines with evidence from this study.

Table S11: Treatments available for PheRS-associated Mendelian diseases.

Table S12: HPO to phecode map.

Table S13: Phecode frequencies and weights for discovery and replication cohorts.

Table S14: Description of Mendelian diseases tested in discovery analysis.

Table S15: Description variants tested in discovery analysis.

Table S16: Statistical results from PheRS analysis in discovery cohort.

Table S17: ANNOVAR annotations and categorical predictions.

References and Notes

1. J. R. Lupski, J. W. Belmont, E. Boerwinkle, R. A. Gibbs, Clan genomics and the complex architecture of human disease. *Cell* **147**, 32–43 (2011). doi:10.1016/j.cell.2011.09.008 Medline

2. D. R. Blair, C. S. Lyttle, J. M. Mortensen, C. F. Bearden, A. B. Jensen, H. Khiabanian, R. Melamed, R. Rabadan, E. V. Bernstam, S. Brunak, L. J. Jensen, D. Nicolae, N. H. Shah, R. L. Grossman, N. J. Cox, K. P. White, A. Rzhetsky, A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. *Cell* **155**, 70–80 (2013). doi:10.1016/j.cell.2013.08.030 Medline

3. T. Groza, S. Köhler, D. Moldenhauer, N. Vasilevsky, G. Baynam, T. Zemojtel, L. M. Schriml, W. A. Kibbe, P. N. Schofield, T. Beck, D. Vasant, A. J. Brookes, A. Zankl, N. L. Washington, C. J. Mungall, S. E. Lewis, M. A. Haendel, H. Parkinson, P. N. Robinson, The human phenotype ontology: Semantic unification of common and rare disease. *Am. J. Hum. Genet.* **97**, 111–124 (2015). doi:10.1016/j.ajhg.2015.05.020 Medline

4. D. Langlais, N. Fodil, P. Gros, Genetics of infectious and inflammatory diseases: Overlapping discoveries from association and exome-sequencing studies. *Annu. Rev. Immunol.* **35**, 1–30 (2017). doi:10.1146/annurev-immunol-051116-052442 Medline

5. J. C. Denny, L. Bastarache, D. M. Roden, Phenome-wide association studies as a tool to advance precision medicine. *Annu. Rev. Genomics Hum. Genet.* **17**, 353–373 (2016).

6. W. S. Bush, M. T. Oetjens, D. C. Crawford, Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat. Rev. Genet.* **17**, 129–145 (2016). doi:10.1038/nrg.2015.36 Medline

7. E. Marouli, M. Graff, C. Medina-Gomez, K. S. Lo, A. R. Wood, T. R. Kjaer, R. S. Fine, Y. Lu, C. Schurmann, H. M. Highland, S. Rüeger, G. Thorleifsson, A. E. Justice, D. Lamparter, K. E. Stirrups, V. Turcot, K. L. Young, T. W. Winkler, T. Esko, T. Karaderi, A. E. Locke, N. G. D. Masca, M. C. Y. Ng, P. Mudgal, M. A. Rivas, S. Vedantam, A. Mahajan, X. Guo, G. Abecasis, K. K. Aben, L. S. Adair, D. S. Alam, E. Albrecht, K. H. Allin, M. Allison, P. Amouyel, E. V. Appel, D. Arveiler, F. W. Asselbergs, P. L. Auer, B. Balkau, B. Banas, L. E. Bang, M. Benn, S. Bergmann, L. F. Bielak, M. Blüher, H. Boeing, E. Boerwinkle, C. A. Böger, L. L. Bonnycastle, J. Bork-Jensen, M. L. Bots, E. P. Bottinger, D. W. Bowden, I. Brandslund, G. Breen, M. H. Brilliant, L. Broer, A. A. Burt, A. S. Butterworth, D. J. Carey, M. J. Caulfield, J. C. Chambers, D. I. Chasman, Y. I. Chen, R. Chowdhury, C. Christensen, A. Y. Chu, M. Cocca, F. S. Collins, J. P. Cook, J. Corley, J. C. Galbany, A. J. Cox, G. Cuellar-Partida, J. Danesh, G. Davies, P. I. W. de Bakker, G. J. de Borst, S. de Denus, M. C. H. de Groot, R. de Mutsert, I. J. Deary, G. Dedoussis, E. W. Demerath, A. I. den Hollander, J. G. Dennis, E. Di Angelantonio, F. Drenos, M. Du, A. M. Dunning, D. F. Easton, T. Ebeling, T. L. Edwards, P. T. Ellinor, P. Elliott, E. Evangelou, A.-E. Farmaki, J. D. Faul, M. F. Feitosa, S. Feng, E. Ferrannini, M. M. Ferrario, J. Ferrieres, J. C. Florez, I. Ford, M. Fornage, P. W. Franks, R. Frikke-Schmidt, T. E. Galesloot, W. Gan, I. Gandin, P. Gasparini, V. Giedraitis, A. Giri, G. Girotto, S. D. Gordon, P. Gordon-Larsen, M. Gorski, N. Grarup, M. L. Grove, V. Gudnason, S. Gustafsson, T. Hansen, K. M. Harris, T. B. Harris, A. T. Hattersley, C. Hayward, L. He, I. M. Heid, K. Heikkilä, Ø. Helgeland, J. Hernesniemi, A. W. Hewitt, L.

J. Hocking, M. Hollensted, O. L. Holmen, G. K. Hovingh, J. M. M. Howson, C. B. Hoyng, P. L. Huang, K. Hveem, M. A. Ikram, E. Ingelsson, A. U. Jackson, J.-H. Jansson, G. P. Jarvik, G. B. Jensen, M. A. Jhun, Y. Jia, X. Jiang, S. Johansson, M. E. Jørgensen, T. Jørgensen, P. Jousilahti, J. W. Jukema, B. Kahali, R. S. Kahn, M. Kähönen, P. R. Kamstrup, S. Kanoni, J. Kaprio, M. Karaleftheri, S. L. R. Kardia, F. Karpe, F. Kee, R. Keeman, L. A. Kiemeney, H. Kitajima, K. B. Kluivers, T. Kocher, P. Komulainen, J. Kontto, J. S. Kooner, C. Kooperberg, P. Kovacs, J. Kriebel, H. Kuivaniemi, S. Küry, J. Kuusisto, M. La Bianca, M. Laakso, T. A. Lakka, E. M. Lange, L. A. Lange, C. D. Langefeld, C. Langenberg, E. B. Larson, I.-T. Lee, T. Lehtimäki, C. E. Lewis, H. Li, J. Li, R. Li-Gao, H. Lin, L.-A. Lin, X. Lin, L. Lind, J. Lindström, A. Linneberg, Y. Liu, Y. Liu, A. Lophatananon, J. Luan, S. A. Lubitz, L.-P. Lyytikäinen, D. A. Mackey, P. A. F. Madden, A. K. Manning, S. Männistö, G. Marenne, J. Marten, N. G. Martin, A. L. Mazul, K. Meidtner, A. Metspalu, P. Mitchell, K. L. Mohlke, D. O. Mook-Kanamori, A. Morgan, A. D. Morris, A. P. Morris, M. Müller-Nurasyid, P. B. Munroe, M. A. Nalls, M. Nauck, C. P. Nelson, M. Neville, S. F. Nielsen, K. Nikus, P. R. Njølstad, B. G. Nordestgaard, I. Ntalla, J. R. O'Connel, H. Oksa, L. M. O. Loohuis, R. A. Ophoff, K. R. Owen, C. J. Packard, S. Padmanabhan, C. N. A. Palmer, G. Pasterkamp, A. P. Patel, A. Pattie, O. Pedersen, P. L. Peissig, G. M. Peloso, C. E. Pennell, M. Perola, J. A. Perry, J. R. B. Perry, T. N. Person, A. Pirie, O. Polasek, D. Posthuma, O. T. Raitakari, A. Rasheed, R. Rauramaa, D. F. Reilly, A. P. Reiner, F. Renström, P. M. Ridker, J. D. Rioux, N. Robertson, A. Robino, O. Rolandsson, I. Rudan, K. S. Ruth, D. Saleheen, V. Salomaa, N. J. Samani, K. Sandow, Y. Sapkota, N. Sattar, M. K. Schmidt, P. J. Schreiner, M. B. Schulze, R. A. Scott, M. P. Segura-Lepe, S. Shah, X. Sim, S. Sivapalaratnam, K. S. Small, A. V. Smith, J. A. Smith, L. Southam, T. D. Spector, E. K. Speliotes, J. M. Starr, V. Steinthorsdottir, H. M. Stringham, M. Stumvoll, P. Surendran, L. M. 't Hart, K. E. Tansey, J.-C. Tardif, K. D. Taylor, A. Teumer, D. J. Thompson, U. Thorsteinsdottir, B. H. Thuesen, A. Tönjes, G. Tromp, S. Trompet, E. Tsafantakis, J. Tuomilehto, A. Tybjaerg-Hansen, J. P. Tyrer, R. Uher, A. G. Uitterlinden, S. Ulivi, S. W. van der Laan, A. R. Van Der Leij, C. M. van Duijn, N. M. van Schoor, J. van Setten, A. Varbo, T. V. Varga, R. Varma, D. R. V. Edwards, S. H. Vermeulen, H. Vestergaard, V. Vitart, T. F. Vogt, D. Vozzi, M. Walker, F. Wang, C. A. Wang, S. Wang, Y. Wang, N. J. Wareham, H. R. Warren, J. Wessel, S. M. Willems, J. G. Wilson, D. R. Witte, M. O. Woods, Y. Wu, H. Yaghootkar, J. Yao, P. Yao, L. M. Yerges-Armstrong, R. Young, E. Zeggini, X. Zhan, W. Zhang, J. H. Zhao, W. Zhao, W. Zhao, H. Zheng, W. Zhou, EPIC-InterAct Consortium, CHD Exome+ Consortium, ExomeBP Consortium, T2D-Genes Consortium, GoT2D Genes Consortium, Global Lipids Genetics Consortium, ReproGen Consortium, MAGIC Investigators, J. I. Rotter, M. Boehnke, S. Kathiresan, M. I. McCarthy, C. J. Willer, K. Stefansson, I. B. Borecki, D. J. Liu, K. E. North, N. L. Heard-Costa, T. H. Pers, C. M. Lindgren, C. Oxvig, Z. Kutalik, F. Rivadeneira, R. J. F. Loos, T. M. Frayling, J. N. Hirschhorn, P. Deloukas, G. Lettre, Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017). doi:10.1038/nature21039 Medline

8. D. G. MacArthur, T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, R. B. Altman, S. E. Antonarakis, E. A. Ashley, J. C. Barrett, L. G. Biesecker, D. F. Conrad, G. M. Cooper, N. J. Cox, M. J. Daly, M. B. Gerstein, D. B. Goldstein, J. N. Hirschhorn, S. M. Leal, L. A. Pennacchio, J. A. Stamatoyannopoulos, S. R. Sunyaev, D. Valle, B. F. Voight, W. Winckler, C. Gunter, Guidelines for investigating

causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014). doi:10.1038/nature13127 Medline

9. M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, J. Shendure, A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014). doi:10.1038/ng.2892

10. I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, S. R. Sunyaev, A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010). doi:10.1038/nmeth0410-248 Medline

11. N. M. Ioannidis, J. H. Rothstein, V. Pejaver, S. Middha, S. K. McDonnell, S. Baheti, A. Musolf, Q. Li, E. Holzinger, D. Karyadi, L. A. Cannon-Albright, C. C. Teerlink, J. L. Stanford, W. B. Isaacs, J. Xu, K. A. Cooney, E. M. Lange, J. Schleutker, J. D. Carpten, I. J. Powell, O. Cussenot, G. Cancel-Tassin, G. G. Giles, R. J. MacInnis, C. Maier, C.-L. Hsieh, F. Wiklund, W. J. Catalona, W. D. Foulkes, D. Mandal, R. A. Eeles, Z. Kote-Jarai, C. D. Bustamante, D. J. Schaid, T. Hastie, E. A. Ostrander, J. E. Bailey-Wilson, P. Radivojac, S. N. Thibodeau, A. S. Whittemore, W. Sieh, REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016). doi:10.1016/j.ajhg.2016.08.016 Medline

12. H. L. Rehm, J. S. Berg, L. D. Brooks, C. D. Bustamante, J. P. Evans, M. J. Landrum, D. H. Ledbetter, D. R. Maglott, C. L. Martin, R. L. Nussbaum, S. E. Plon, E. M. Ramos, S. T. Sherry, M. S. Watson; ClinGen, ClinGen—the Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015). doi:10.1056/NEJMsr1406261 Medline

13. M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, Exome Aggregation Consortium, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016). doi:10.1038/nature19057 Medline

14. A. K. Manrai, B. H. Funke, H. L. Rehm, M. S. Olesen, B. A. Maron, P. Szolovits, D. M. Margulies, J. Loscalzo, I. S. Kohane, Genetic misdiagnoses and the potential for health disparities. *N. Engl. J. Med.* **375**, 655–665 (2016). doi:10.1056/NEJMsa1507092 Medline

15. R. Chen, L. Shi, J. Hakenberg, B. Naughton, P. Sklar, J. Zhang, H. Zhou, L. Tian, O. Prakash, M. Lemire, P. Sleiman, W. Y. Cheng, W. Chen, H. Shah, Y. Shen, M. Fromer, L. Omberg, M. A. Deardorff, E. Zackai, J. R. Bobe, E. Levin, T. J. Hudson, L. Groop, J. Wang, H. Hakonarson, A. Wojcicki, G. A. Diaz, L. Edelmann, E. E. Schadt, S. H. Friend,

Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat. Biotechnol.* **34**, 531–538 (2016). doi:10.1038/nbt.3514 Medline

16. S. L. Van Driest, Q. S. Wells, S. Stallings, W. S. Bush, A. Gordon, D. A. Nickerson, J. H. Kim, D. R. Crosslin, G. P. Jarvik, D. S. Carrell, J. D. Ralston, E. B. Larson, S. J. Bielinski, J. E. Olson, Z. Ye, I. J. Kullo, N. S. Abul-Husn, S. A. Scott, E. Bottinger, B. Almoguera, J. Connolly, R. Chiavacci, H. Hakonarson, L. J. Rasmussen-Torvik, V. Pan, S. D. Persell, M. Smith, R. L. Chisholm, T. E. Kitchner, M. M. He, M. H. Brilliant, J. R. Wallace, K. F. Doheny, M. B. Shoemaker, R. Li, T. A. Manolio, T. E. Callis, D. Macaya, M. S. Williams, D. Carey, J. D. Kapplinger, M. J. Ackerman, M. D. Ritchie, J. C. Denny, D. M. Roden, Association of arrhythmia-related genetic variants with phenotypes documented in electronic medical records. *JAMA* **315**, 47–57 (2016). doi:10.1001/jama.2015.17701 Medline

17. I. S. Kohane, Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet.* **12**, 417–428 (2011). doi:10.1038/nrg2999 Medline

18. D. C. Crawford, D. R. Crosslin, G. Tromp, I. J. Kullo, H. Kuivaniemi, M. G. Hayes, J. C. Denny, W. S. Bush, J. L. Haines, D. M. Roden, C. A. McCarty, G. P. Jarvik, M. D. Ritchie, eMERGEing progress in genomics-the first seven years. *Front. Genet.* **5**, 184 (2014). doi:10.3389/fgene.2014.00184 Medline

19. OMIM (Online Mendelian Inheritance in Man), http://omim.org/.

20. S. Köhler, S. C. Doelken, C. J. Mungall, S. Bauer, H. V. Firth, I. Bailleul-Forestier, G. C. M. Black, D. L. Brown, M. Brudno, J. Campbell, D. R. FitzPatrick, J. T. Eppig, A. P. Jackson, K. Freson, M. Girdea, I. Helbig, J. A. Hurst, J. Jähn, L. G. Jackson, A. M. Kelly, D. H. Ledbetter, S. Mansour, C. L. Martin, C. Moss, A. Mumford, W. H. Ouwehand, S.-M. Park, E. R. Riggs, R. H. Scott, S. Sisodiya, S. Van Vooren, R. J. Wapner, A. O. M. Wilkie, C. F. Wright, A. T. Vulto-van Silfhout, N. de Leeuw, B. B. A. de Vries, N. L. Washingthon, C. L. Smith, M. Westerfield, P. Schofield, B. J. Ruef, G. V. Gkoutos, M. Haendel, D. Smedley, S. E. Lewis, P. N. Robinson, The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42**, D966–D974 (2014). doi:10.1093/nar/gkt1026 Medline

21. J. C. Denny, L. Bastarache, M. D. Ritchie, R. J. Carroll, R. Zink, J. D. Mosley, J. R. Field, J. M. Pulley, A. H. Ramirez, E. Bowton, M. A. Basford, D. S. Carrell, P. L. Peissig, A. N. Kho, J. A. Pacheco, L. V. Rasmussen, D. R. Crosslin, P. K. Crane, J. Pathak, S. J. Bielinski, S. A. Pendergrass, H. Xu, L. A. Hindorff, R. Li, T. A. Manolio, C. G. Chute, R. L. Chisholm, E. B. Larson, G. P. Jarvik, M. H. Brilliant, C. A. McCarty, I. J. Kullo, J. L. Haines, D. C. Crawford, D. R. Masys, D. M. Roden, Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013). doi:10.1038/nbt.2749 Medline

22. A. Verma, A. O. Basile, Y. Bradford, H. Kuivaniemi, G. Tromp, D. Carey, G. S. Gerhard, J. E. Crowe Jr., M. D. Ritchie, S. A. Pendergrass, Phenome-wide association study to explore relationships between immune system related genetic loci and complex traits and diseases. *PLOS ONE* **11**, e0160573 (2016). doi:10.1371/journal.pone.0160573 Medline

23. W.-Q. Wei, L. A. Bastarache, R. J. Carroll, J. E. Marlo, T. J. Osterman, E. R. Gamazon, N. J. Cox, D. M. Roden, J. C. Denny, Evaluating phecodes, clinical classification software,

and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLOS ONE* **12**, e0175508 (2017). doi:10.1371/journal.pone.0175508 Medline

24. E. L. Macleod, D. M. Ney, Nutritional management of phenylketonuria. *Ann. Nestlé [Engl.]* **68**, 58–69 (2010). doi:10.1159/000312813 Medline

25. P. D. Stenson, M. Mort, E. V. Ball, K. Evans, M. Hayden, S. Heywood, M. Hussain, A. D. Phillips, D. N. Cooper, The Human Gene Mutation Database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* **136**, 665–677 (2017). doi:10.1007/s00439-017-1779-6 Medline

26. Q. Zhou, G.-S. Lee, J. Brady, S. Datta, M. Katan, A. Sheikh, M. S. Martins, T. D. Bunney, B. H. Santich, S. Moir, D. B. Kuhns, D. A. Long Priel, A. Ombrello, D. Stone, M. J. Ombrello, J. Khan, J. D. Milner, D. L. Kastner, I. Aksentijevich, A hypermorphic missense mutation in *PLCG2*, encoding phospholipase Cγ2, causes a dominantly inherited autoinflammatory disease with immunodeficiency. *Am. J. Hum. Genet.* **91**, 713–720 (2012). doi:10.1016/j.ajhg.2012.08.006 Medline

27. N. Maslah, B. Cassinat, E. Verger, J.-J. Kiladjian, L. Velazquez, The role of LNK/SH2B3 genetic alterations in myeloproliferative neoplasms and other hematological disorders. *Leukemia* **31**, 1661–1670 (2017). doi:10.1038/leu.2017.139 Medline

28. W. Tong, J. Zhang, H. F. Lodish, Lnk inhibits erythropoiesis and Epo-dependent JAK2 activation and downstream signaling pathways. *Blood* **105**, 4604–4612 (2005). doi:10.1182/blood-2004-10-4093 Medline

29. C. Camps, N. Petousi, C. Bento, H. Cario, R. R. Copley, M. F. McMullin, R. van Wijk, WGS500 Consortium, P. J. Ratcliffe, P. A. Robbins, J. C. Taylor, Gene panel sequencing improves the diagnostic work-up of patients with idiopathic erythrocytosis and identifies new mutations. *Haematologica* **101**, 1306–1318 (2016). doi:10.3324/haematol.2016.144063 Medline

30. M. Garber, M. Guttman, M. Clamp, M. C. Zody, N. Friedman, X. Xie, Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54–i62 (2009). doi:10.1093/bioinformatics/btp190 Medline

31. S. J. Hebbring, M. Rastegar-Mojarad, Z. Ye, J. Mayer, C. Jacobson, S. Lin, Application of clinical text data for phenome-wide association studies (PheWASs). *Bioinformatics* **31**, 1981–1987 (2015). doi:10.1093/bioinformatics/btv076 Medline

32. S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, H. L. Rehm; ACMG Laboratory Quality Assurance Committee, Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–423 (2015). doi:10.1038/gim.2015.30 Medline

33. D. M. Roden, J. M. Pulley, M. A. Basford, G. R. Bernard, E. W. Clayton, J. R. Balser, D. R. Masys, Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* **84**, 362–369 (2008). doi:10.1038/clpt.2008.89 Medline

34. J. K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000). Medline

35. C. A. McCarty, R. A. Wilke, P. F. Giampietro, S. D. Wesbrook, M. D. Caldwell, Marshfield Clinic Personalized Medicine Research Project (PMRP): Design, methods and recruitment for a large population-based biobank. *Per. Med.* **2**, 49–79 (2005). doi:10.1517/17410541.2.1.49

36. W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016). doi:10.1186/s13059-016-0974-4 Medline

37. K. Wang, M. Li, H. Hakonarson, ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010). doi:10.1093/nar/gkq603 Medline

38. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007). doi:10.1086/519795 Medline

39. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010). doi:10.1101/gr.107524.110 Medline

40. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). doi:10.1093/bioinformatics/btp324 Medline

41. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011). doi:10.1038/ng.806 Medline

42. R. Zhao, S. Xing, Z. Li, X. Fu, Q. Li, S. B. Krantz, Z. J. Zhao, Identification of an acquired JAK2 mutation in polycythemia vera. *J. Biol. Chem.* **280**, 22788–22792 (2005). doi:10.1074/jbc.C500138200 Medline

43. F.-O. Desmet, D. Hamroun, M. Lalande, G. Collod-Béroud, M. Claustres, C. Béroud, Human Splicing Finder: An online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37**, e67 (2009). doi:10.1093/nar/gkp215 Medline

44. C. A. Schneider, W. S. Rasband, K. W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012). doi:10.1038/nmeth.2089 Medline