Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study

Christopher DeBoever1,2, Yosuke Tanigawa1, Malene E. Lindholm3, Greg McInnes1, Adam Lavertu1, Erik Ingelsson4, Chris Chang5, Euan A. Ashley6, Carlos D. Bustamante1,2, Mark J. Daly6,7 & Manuel A. Rivas1

Protein-truncating variants can have profound effects on gene function and are critical for clinical genome interpretation and generating therapeutic hypotheses, but their relevance to medical phenotypes has not been systematically assessed. Here, we characterize the effect of 18,228 protein-truncating variants across 135 phenotypes from the UK Biobank and find 27 associations between medical phenotypes and protein-truncating variants in genes outside the major histocompatibility complex. We perform phenome-wide analyses and directly measure the effect in homozygous carriers, commonly referred to as “human knockouts,” across medical phenotypes for genes implicated as being protective against disease or associated with at least one phenotype in our study. We find several genes with strong pleiotropic or non-additive effects. Our results illustrate the importance of protein-truncating variants in a variety of diseases.

1 Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA. 2 Department of Genetics, Stanford University, Stanford, CA 94305, USA. 3 Grail, Inc., 1525 O’Brien Drive, Menlo Park, CA 94025, USA. 4 Division of Cardiovascular Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA. 5 Department of Medicine, School of Medicine, Stanford University, Stanford, CA 94305, USA. 6 Analytical and Translational Genetics Unit, Boston, MA 02114, USA. 7 Broad Institute of MIT and Harvard, Cambridge 02142 MA, USA. Correspondence and requests for materials should be addressed to M.A.R. (email: mrivas@stanford.edu)
Protein-truncating variants (PTVs), genetic variants predicted to shorten the coding sequence of genes, are a promising set of variants for drug discovery since identification of PTVs that protect against human disease provides in vivo validation of therapeutic targets. Although tens of thousands of germline PTVs have been identified, their medical relevance across a broad range of phenotypes has not been characterized. Because most PTVs are present at low frequency, assessing the effects of PTVs requires genotype data from many individuals with linked phenotype data for a variety of diseases and physiological measurements. The recent release of genotype and linked clinical and questionnaire data for 488,377 individuals in the UK Biobank provides an unprecedented opportunity to assess the clinical impact of truncating protein-coding genes at a resolution not previously possible.

PTVs are genetic variants that disrupt transcription and lead to a shortened or absent protein that often causes loss of protein function; however, it is also possible to observe gain-of-function effects. PTVs include nonsense single-nucleotide variants (SNVs), frameshift insertions or deletions (indels), large structural variants, and splice-disrupting SNVs. Although most common genetic variants associated with disease have relatively small effect on disease risk, PTVs are expected to have much stronger effects on disease risk as they dramatically alter protein sequence.

Population sequencing efforts have estimated that every human genome contains ~ 100 PTVs although this rate can vary. The ~ 100 loss-of-function variants observed in the 1000 Genomes Project5 were predicted to be loss-of-function carriers in previous studies (Supplementary Data1). We identified 291 genes that had at least one observed homozygous PTV carrier in our study but had no observed homozygous loss-of-function carriers in previous studies (Supplementary Data1).

Here, we test for associations between PTVs and 135 different medical phenotypes including cancers and complex diseases among 337,205 participants in the UK Biobank. We identify 27 PTVs outside of the MHC that are associated with at least one medical phenotype, including several protective associations. We perform phenome-wide association analyses across 206 medical phenotypes for these PTVs as well as PTVs with previously identified associations and find PTVs with pleiotropic effects. We also perform a human “knockout” analysis to identify non-additive associations for homozygous or compound heterozygous PTV carriers and find several genes with non-additive effects. The associations reported here indicate new disease-causing genes that may be promising therapeutic targets.

Results

PTV genetic association analysis. To assess the clinical relevance of PTVs, we cataloged predicted PTVs present on the Affymetrix UK Biobank array and their effects on medical phenotypes from 337,205 unrelated individuals in the UK Biobank study. We defined PTVs as SNVs predicted to introduce a premature stop codon or to disrupt a splice site or small indels predicted to disrupt a transcript’s reading frame. Although methods to predict PTVs, also referred to as loss-of-function or knockout variants, are still being improved and validated, previous work has found that 70% of nonsense PTVs predicted to cause nonsense-mediated decay show evidence for decreased expression of the corresponding transcript and 79% of splice-site variants disrupt splicing, indicating that predicted PTVs are likely to affect gene expression or function.

We identified 18,228 predicted PTVs in the UK Biobank array that were polymorphic across 8750 genes after filtering (Methods, Supplementary Fig. 1). Each participant had 95 predicted PTVs with minor allele frequency (MAF) < 1% on average, and 778 genes were predicted to be homozygous or compound heterozygous for PTVs with MAF < 1% in at least one individual. We observed 291 genes that had at least one observed homozygous PTV in our study but had no observed homozygous loss-of-function carriers in previous studies (Supplementary Data1).

The observed number of PTVs per individual is consistent with the ~ 100 loss-of-function variants observed in the 1000 Genomes project. In contrast, the number of PTV singletons (or observed allele counts < 10) in ExAC suggests approximately five singletons per individual and only ~ 0.2 per individual in highly constrained genes. These observations indicate that the majority of PTVs in an individual are common (or common and low frequency) such that they can be assessed via genotyping.

We used computational matching and manual curation based on hospital in-patient data (National Health Service Hospital Episode Statistics), self-reported verbal questionnaire data, and cancer and death registry data to define a broad set of medical phenotypes including various cancers, cardiometabolic diseases, and autoimmune diseases (Supplementary Data 2). We then performed association analyses between the 3724 PTVs with MAF > 0.01% and 135 medical phenotypes with at least 2000 case samples (Fig. 1, Supplementary Fig. 2) and stratified the association results into three bins based on PTV MAF > 1% (463 PTVs), between 0.1% and 1% (700 PTVs), and between 0.01% and 0.1% (2561 PTVs) to account for expected differences in the statistical power to detect associations for PTVs with different MAFs (Supplementary Fig. 3). We adjusted the nominal association p values separately for each MAF bin using the
Benjamini-Yekutieli (BY) procedure to correct for multiple hypothesis testing and identified 74 significant associations between PTVs and medical phenotypes (BY-adjusted $p < 0.05$, Fig. 2a–c, Supplementary Data 3).

Among the 74 PTV-phenotype associations we identified, 27 involved PTVs in genes outside of the MHC (chr6–25, 477, 797–36, 448, 354). As PTVs in or near the MHC likely tag HLA risk alleles, we focused on associations for PTVs outside of the MHC. We identified 5 PTVs with seven associations consistent with protective effects (odds ratio (OR) < 1, BY-adjusted $p < 0.05$, Fig. 2d, Supplementary Data 3). We found that the rare splice-disrupting PTV rs146597587 in IL33 is associated with protection against asthma (MAF = 0.48%, $p = 7.6 \times 10^{-13}$, OR = 0.64, 95% confidence interval (CI): 0.57–0.72). This PTV is negatively associated with eosinophil counts ($\beta = -0.21$, SD, $p = 2.5 \times 10^{-16}$) and has suggestive evidence of an association with asthma ($p = 1.8 \times 10^{-4}$, OR = 0.47, 95% CI: 0.32–0.70). Our results provide strong evidence in an independent sample that this PTV protects against asthma and suggests that knocking down IL33 function may be a useful therapeutic approach for asthma. We also identified protective associations for the PTV rs11078928 (MAF = 47.1%) in GSDMB against asthma ($p = 6.3 \times 10^{-50}$, OR = 0.90, 95% CI: 0.88–0.91) and bronchitis ($p = 2.6 \times 10^{-6}$, OR = 0.91, 95% CI: 0.87–0.95). GSDMB is associated with asthma in humans and induces an asthma phenotype in mouse when overexpressed. We identified additional protective associations between PTVs in IFI1H1 and hypothyroidism (labeled as hypothyroidism/myxedema) (MAF = 1.5%, $p = 1.7 \times 10^{-6}$, OR = 0.80, 95% CI: 0.73–0.88) and VKORC1 and hypertension (MAF = 25.3%, $p = 1.4 \times 10^{-6}$, OR = 0.97, 95% CI: 0.96–0.98).

We also found 20 risk associations for PTVs in 12 genes outside the MHC (Fig. 2d, Supplementary Data 3). We identified clinically relevant PTV-phenotype associations such as FLG, whose protein product contributes to the structure of epidermal cells, and eczema/dermatitis (MAF = 0.48%, $p = 6.7 \times 10^{-15}$, OR = 1.80, 95% CI: 1.55–2.08) and TSHR, thyroid-stimulating hormone receptor, and hypothyroidism/myxedema (MAF = 0.046%, $p = 1.2 \times 10^{-13}$, OR = 3.30, 95% CI: 2.41–4.53). We replicated known risk genome-wide association study (GWAS) associations such as BRCA2 and family history of lung cancer (MAF = 0.93%, $p = 7.3 \times 10^{-11}$, OR = 1.19, 95% CI: 1.13–1.25) and rs33966350 in ENPEP and hypertension (MAF = 1.3%, $p = 4.8 \times 10^{-11}$, OR = 1.17, 95% CI: 1.12–1.23) and identified risk associations between FANCM, a member of the same gene family as BRCA2, and lung cancer (MAF = 0.11%, $p = 9.7 \times 10^{-10}$, OR = 1.58, 95% CI: 1.36–1.83) as well as NOL3, a regulator of apoptosis in muscle cells, and muscle or soft tissue injury (MAF = 0.11%, $p = 6.5 \times 10^{-9}$, OR = 3.43, 95% CI: 2.19–5.36). To investigate the association between NOL3 and tissue injury, we knocked down NOL3 threefold in differentiated human skeletal muscle cells and used electrical pulse stimulation to induce cell damage and simulate injury. Lower expression of NOL3 resulted in increased activation of caspase 8, an early indicator of apoptosis, in the damaged cells, consistent with the observation that NOL3 inhibits caspase 8 (Supplementary Fig. 4a, b). The degree of DNA fragmentation, another indicator of tissue damage, was also higher in NOL3 knockdown cells compared with control (Supplementary Fig. 4c). We observed higher expression of MAFbx/atrogin-1 (mRNA and protein), a muscle-specific E3 ubiquitin ligase that is activated during skeletal ...
muscle atrophy in NOL3 knockdowns without stimulation (Supplementary Fig. 4d, e), consistent with increased expression of MAFbx in NOL3 knockout mice and general protein degradation after stimulation. These results provide additional evidence that NOL3 has an important role in muscle injury.

Even in the context of variants with strong predicted effects such as PTVs, it is critical to evaluate whether the associated variant is causal in the context of neighboring variants. We initially identified an association between the PTV rs34358 in ANKDD1B and high cholesterol, although this association disappeared upon conditional analysis with rs17238484, an intronic variant in HMGCR known to be associated with cholesterol levels. Another association between rs34358 and family history of diabetes remained upon conditional analysis with rs17238484 (p = 9.1 × 10^-3, OR = 1.03, 95% CI: 1.02–1.05). We performed conditional analyses for the remaining 27 associations outside of the MHC by identifying genotyped variants within 10 kb of the associated PTV and using the genotypes of the nearby variant as a covariate for logistic regression. For PTVs with MAF < 1%, we found that only the association between a PTV in HEATR6 and retinal detachment was explained by a nearby variant rs3744375 (Supplementary Data 3). Six of the common (MAF > 1%) PTVs with associations were in high linkage disequilibrium with other nearby common variants that may explain the observed associations (Supplementary Data 3, Supplementary Fig. 5), though the PTVs remain strong functional candidates for these associations. For instance, the gain-of-function PTV rs328 in LPL (MAF = 10.1%) that we find to be associated with decreased risk for high cholesterol (p = 3.9 × 10^-15, OR = 0.90, 95% CI: 0.88–0.93) and angina (p = 1.3 × 10^-7, OR = 0.91, 95% CI: 0.87–0.94) has been associated with coronary artery disease, lipid metabolism, and lower triglyceride levels. Similarly, a recent study found that the PTV rs11078928 in GSDMB that offers protection against asthma removes exon 6 from the transcript and eliminates the ability of GSDMB to induce cell death. The PTV rs2004640 in IRF5 has previously been associated with rheumatoid arthritis and has been connected to pathogenesis in the mouse model. The PTV rs601338 in FUT2 determines secretor status for ABH blood groups that has been associated with susceptibility to infection and several diseases. The PTV rs2884737 in VKORC1 associated with hypertension is in moderate LD (R^2 = 0.56) with several nearby common variants and the PTV rs776746 in CYP3A5 associated with hayfever/allergic rhinitis is in near perfect LD with one other nearby variant. Additional functional work may be needed to establish whether the PTVs are causal for these two associations.

We identified five significant associations between PTVs and family history phenotypes included in our analysis (Supplementary Data 3). For two of these associations, the variant associated with the family history phenotype was also associated directly with the family history phenotype was also associated directly with the family history of breast cancer (MAF = 49.1%, OR = 1.3 × 10^-7, OR = 1.03, 95% CI: 1.02–1.04), hypertension (p = 5.7 × 10^-13, OR = 1.04, 95% CI: 1.03–1.05), and essential hypertension (p = 5.2 × 10^-8, OR = 1.04, 95% CI: 1.02–1.05). We also found that the PTV rs11571833 in BRC2A2 was associated with lung cancer (MAF = 0.934%, OR = 7.3 × 10^-11, OR = 1.19, 95% CI: 1.13–1.25). These results demonstrate previous approaches for identifying genetic associations using family history information (e.g., ref. 48,49) can be applied even to relatively rare PTVs.

To further characterize the PTV-phenotype associations, we asked whether missense variants with MAF > 0.01% in the genes with significant PTV associations were also associated with the same phenotypes. For each of the 27 PTV-phenotype associations in our GWAS, we performed association analyses between the missense variants in that gene and the phenotype that the PTV was associated with and found 23 missense variant-phenotype associations with p < 0.001 (Supplementary Data 3). Thirteen of these 23 associations remain significant after a conditional analysis including the PTV genotype as a covariate indicating that several genes with PTV associations also contain independent missense associations. For instance, we found two different missense variants in TSHR that were both associated with hypothyroidism independent of the PTV association. We also identified independent missense associations for genes and phenotypes such as ENPEP and hypertension; GSDMB and asthma; IFH1 and hypothyroidism; and PALB2 and lung cancer (Supplementary Data 3). In total, we found at least one missense association for seven genes implicated in our PTV GWAS providing more evidence that these genes are likely important to the etiology of these conditions.

Forty-seven of the 74 significant associations involved PTVs in genes in or near the MHC (Supplementary Data 4). To investigate whether these associations are caused by linkage between these PTVs and HLA susceptibility alleles, we performed association analyses for each of these PTVs conditional on the presence of each of 344 HLA alleles that were polymorphic among the 337,205 subjects (Supplementary Data 4). We found that the p values for all five associations with MAF between 0.1 and 1% were > 0.05 for at least one HLA allele (Fig. 2e). Similarly, the p values for 30 of 42 associations with MAF > 1% were > 0.05 for at least one HLA allele and only three were < 0.001 (Supplementary Data 4). For instance, we identified an association between rs72841509 in BTN3A2 and Celiac disease (categorized as malabsorption/celiac disease) in our initial GWAS (MAF = 0.13, p = 1.8 × 10^-119, OR = 2.33, 95% CI: 2.17–2.50). However, conditioning upon the presence of HLA-B8, which is on the same haplotype as the HLA-DQ2 Celiac risk allele, reduced the p value of the association between rs72841509 and Celiac disease to p = 0.92. Both of these results indicate that the majority of the associations identified here for PTVs in MHC genes are likely due to LD with HLA susceptibility alleles and show that it is important to carefully consider the genomic context of associated variants, even for variants with strong predicted effects.

We next investigated whether we could identify PTV-phenotype associations using imputed genotypes. After filtering (Methods), we identified 546 PTVs outside the MHC with MAF greater than 0.01% among the UK Biobank imputed genotypes. We stratified these PTVs into the same MAF bins as above (0.01–0.1%, 0.1–1%, and 1–50%) and applied the BY adjustment to the association p values for each bin. We found nine significant associations for imputed PTVs (BY-adjusted p < 0.05, Supplementary Data 3) including rs74315329 in MYOC and glaucoma (MAF = 0.0012, p = 1.8 × 10^-30, OR = 4.71, 95% CI: 3.61–6.14), a well-known risk variant for glaucoma, and D2HGHD and asthma (MAF = 0.445, p = 1.6 × 10^-12, OR = 0.95, 95% CI: 0.94–0.96) and hayfever (categorized as hayfever/allergic rhinitis) (p = 8.4 × 10^-79, OR = 0.94, 95% CI: 0.92–0.96). The D2HGHD PTV is in partial LD with an intronic variant rs34290285 in D2HGHD (r^2 = 0.366, LDlink) that has been associated with asthma and allergic disease in the initial UK Biobank data release. We also identified an association between the PTV rs754512 in MAPT and Parkinson’s disease (MAF = 0.23, p = 1.1 × 10^-56, OR = 0.94, 95% CI: 0.92–0.97). This variant is predicted to be a PTV but is in the intron of the canonical MAPT transcript and
lies on the same haplotype as three MAPT missense variants (rs17651549, rs62063786, rs10445337) so conditional analysis could not establish the causal allele. We found associations between a PTV in RPL3L and atrial flutter (MAF = 0.0021, OR = 5.0 × 10⁻¹⁰, 95% CI: 0.44–0.66) and atrial fibrillation (p = 2.3 × 10⁻⁹, OR = 0.55, 95% CI: 0.46–0.67). The missense variant rs140185678 in MAPT is also independently associated with atrial fibrillation (MAF = 0.0363, p = 5.4 × 10⁻⁹, OR = 1.21, 95% CI: 1.14–1.30) and atrial flutter (p = 1.1 × 10⁻⁷, OR = 1.20, 95% CI: 1.12–1.28). Overall, we were able to recover a small number of associations using imputed PTVs, indicating that better imputation methods are likely needed in the absence of direct genotyping of PTVs.

Targeted PTV phenome-wide association study. To further assess the role of PTVs across medical phenotypes, we performed a phenome-wide association analysis (pheWAS) to determine whether PTVs that have been implicated in disease predisposition may impact other diseases or commonly measured traits. We focused this analysis on PTVs with minor allele frequency > 0.01%, which had many PTVs on the same haplotype as three MAPT missense variants (rs17651549, rs62063786, rs10445337) so conditional analysis could not establish the causal allele. We found associations between a PTV in RPL3L and atrial flutter (MAF = 0.0021, OR = 5.0 × 10⁻¹⁰, 95% CI: 0.44–0.66) and atrial fibrillation (p = 2.3 × 10⁻⁹, OR = 0.55, 95% CI: 0.46–0.67). The missense variant rs140185678 in MAPT is also independently associated with atrial fibrillation (MAF = 0.0363, p = 5.4 × 10⁻⁹, OR = 1.21, 95% CI: 1.14–1.30) and atrial flutter (p = 1.1 × 10⁻⁷, OR = 1.20, 95% CI: 1.12–1.28). Overall, we were able to recover a small number of associations using imputed PTVs, indicating that better imputation methods are likely needed in the absence of direct genotyping of PTVs.

Human gene knockout analysis. Homozygous carriers of PTVs, referred to as homozygous knockouts (KOs), may have dramatically altered medical outcomes compared with carriers with only one PTV (heterozygous KOs). Genetic association analyses typically assume that genetic effects are additive; that is, the log OR of a homozygote is expected to be twice the log OR of a heterozygote. Given the large difference between having one functional copy and no functional copies of a gene, however, we expect that homozygote KOs may have non-additive effects that are stronger or weaker than would be predicted given the effect size for heterozygote KOs. To assess whether any of the 17 genes with significant associations in our GWAS or the eight genes with published protective effects have non-additive effects on medical phenotypes, we estimated the KO status in each subject for each of these 25 genes. Subjects with one PTV in a gene were considered heterozygote KOs for that gene and subjects with two or more PTVs were considered homozygote KOs. In total, 16 of the 25 genes had at least one predicted homozygous KO carrier. We fit additive and non-additive models to test for associations between KO status

---

**Fig. 3** PheWAS for IFIH1. Phenome-wide associations (logistic regression, p < 0.01) for four PTVs in IFIH1 with minor allele frequency > 0.01%. The left panel shows the number of cases per phenotype in thousands. The middle panel shows the logistic regression −log₁₀ p values. The right panel shows the estimated odds ratios and 95% confidence intervals.
PTVs, a class of variants with functional consequences likely to be critical for prioritizing putative drug targets and clinical interpretation. We systematically characterized the association of non-additive PTV associations across several phenotypes including hypertension and mumps (Fig. 4, Supplementary Data 5). FUT2 regulates the expression of the H antigen on the gastrointestinal mucosa and genetic variation in FUT2 is associated with Crohn’s disease, psoriasis, plasma vitamin B12 levels, levels of two tumor biomarkers, and urine fucose levels. Under a non-additive model, the ORs for heterozygous FUT2 KOs are all nearly one while FUT2 homozygous KOs have ORs ranging from 1.05 (95% CI: 1.03–1.07) to 1.51 (95% CI: 1.29–1.77). Given the frequency of the rs601338 PTV, our results indicate that FUT2 function may have an important role in a wide range of phenotypes.

We also found evidence that the association between GSDMB KO and asthma described in our GWAS analysis above is non-additive (Figure 9, Supplementary Data 5). In total, we identified 168,025 heterozygous KOs and 74,534 homozygous KOs for GSDMB. Under an additive model, GSDMB heterozygote KOs are predicted to have a decreased risk for asthma with OR = 0.90 (p = 5.9 × 10^{-56}, 95% CI: 0.88–0.91). Under a non-additive model, however, GSDMB heterozygote KOs are predicted to have OR = 0.86 (p = 4.3 × 10^{-38}, 95% CI: 0.84–0.88) while GSDMB homozygote KO offers only modestly higher protection (p = 9.7 × 10^{-46}, OR = 0.81, 95% CI: 0.79–0.84). Variants that increase expression of GSDMB in humans are associated with asthma risk, and increased GSDMB causes an asthma phenotype in mice. Our results suggest that knocking out just one copy of GSDMB provides most of the protective effect on asthma risk. Overall, we identified non-additive PTV associations for six of 16 genes tested demonstrating that the effect of PTVs on disease risk can be complex.

Discussion
Assessing the medical relevance of protein-truncating variants is critical for prioritizing putative drug targets and clinical interpretation. We systematically characterized the association of PTVs, a class of variants with functional consequences likely to be consistent with inhibition, with medical phenotypes using data from the UK Biobank study. We estimated the effects of PTVs across 135 phenotypes and identified 27 associations between PTVs in 17 genes and 20 different phenotypes. We found four associations for PTVs with minor allele frequency < 0.1%, indicating that more subjects or case/control studies design may be needed to test for associations between ultra-rare PTVs and relatively low prevalence diseases that are not well-represented in biobank datasets. We performed 25 phenome-wide association analyses for the genes implicated by GWAS in this study plus eight genes curated from the literature (Supplementary Table 1) and identified eight genes that were associated with eight or more phenotypes (p < 0.01). Six of these 25 genes showed evidence for non-additive associations across several phenotypes including non-additive associations between GSDMB and asthma and FUT2 and eight phenotypes including hypertension and cholesterol.

The genetic associations reported here directly link gene function to disease etiology and provide attractive targets for drug discovery. Naturally occurring human knockouts that protect against disease provide in vivo validation of safety and efficacy and may be relatively simple to target with drugs. Protective associations between PTVs in IL33 and asthma; GSDMB and asthma; and IFIH1 and hypothyroidism represent particularly attractive drug targets, whereas risk associations between PTVs in FANCM and lung cancer and NOL3 and muscle injuries implicate these genes as important to the development of these conditions. Our results illustrate the value of deep population-scale health and genomic datasets for prioritizing genetic variants and genes with translational potential.

Methods
Quality control of genotype data. We used genotype data from UK Biobank dataset release version 2 for all aspects of the study except the imputed PTV GWAS. To minimize the impact of cofounders and unreliable observations, we used a subset of individuals that satisfied all of the following criteria: (1) self-reported white British ancestry; (2) used to compute principal components; (3) not marked as outliers for heterozygosity and missing rates; (4) do not show putative sex chromosome aneuploidy; and (5) “excess relatives.” We removed 151,169 individuals that did not meet these criteria. For the remaining 337,205 individuals, we used PLINK v1.90b4.176 to compute the following statistics for each variant: (a)
genotyping misalignment rate, (b) p values of Hardy–Weinberg test, and (c) allele frequencies.

**Protein-truncating variant annotation.** We annotated 784,257 autosomal variants extracted from the mapping bin files provided by the UK Biobank using VEP version 87 and the LRT-EVE plugin (https://gtexportal.org/svtools/lrt) and identified 27,057 putative PTVs
d27. We first removed 8118 PTVs specific to the UK BiLEVE Axiom Array or with misalignment > 1% among the subjects genotyped on the UK Biobank Axiom Array. Despite a misalignment rate of 28% on the Axiom Biobank Array, we kept rs141923399 (CARD9) in the analysis. We removed 11 variants with low cluster plots that indicated unusual genotype patterns. We removed Affx-89018997 because the REF/ALT annotation caused problems with analysis software.

We next matched our PTVs to PTVs annotated in gnomAD (gnomad.exomes.r.2.6.1.sites.vcf.gz) based on genomic position, reference, and alternate alleles and compared the allele frequencies in the UKB and gnomAD by (1) performing a Fisher’s exact test using the minor allele counts from the 337,205 UKB subjects and the minor allele counts from gnomAD and (2) calculating the log odds ratio of observing the minor allele in the UKB vs. gnomAD. We stratified our PTVs by minor allele frequency into the following three bins: (0.01%, 1%), (0.1%, 1%), (1%, 5%). For each bin, we performed a Fisher’s exact test comparing the allele frequencies in the UKB and gnomAD. We removed Affx-89018997 because the REF/ALT annotation caused problems with analysis software.

We then identified 784,257 autosomal variants that were not present in gnomAD using VEP (http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=20001). We kept the PTVs if they were present in the UK BiLEVE Axiom Array or with missingness > 1% among the subjects genotyped on the UK BiLEVE Axiom Array. For variants that were specific to one array, we did not use array as a covariate. We stratified GWAS p values from PLINK into three minor allele frequency bins: 0.01–0.1% (2562 PTVs), 0.1–1% (700 PTVs), and >1% (463 PTVs). We corrected p values separately for each bin using the Benjamini–Yekutieli approach implemented in R’s p.adjust function. We considered associations with BY-corrected p values < 0.05 as significant which controls the false discovery rate at 5%. As we identified 74 significant associations in our main analysis, we would expect ~4 false-positive associations. We also applied the Bonferroni correction for each MAF bin and for all tests for reference (Supplementary Data 3).

The missense variant GWAS was performed using the gnomAD genotype as a covariate to evaluate whether the association signals were independent for significant missense variants.

**Cancer phenotype definitions.** We combined cancer diagnoses from the UK Cancer Register with self-reported diagnoses from the UK Biobank questionnaire to define cases and controls for cancer GWAS. Individual level ICD-10 codes were used to map to the self-reported cancer codes. We manually reviewed these variants on the gnomAD browser to determine whether they were likely to accurately type a PTV in gnomAD. In cases where the PTV was present on the gnomAD browser but was not included in the exome data, we kept the PTV in our analysis. In cases where the UKB array likely typed a non-PTV or there was no variant present on the browser, we removed the PTV from our analysis. In total, 79/134 variants were removed during this step. 18,726 PTVs remained after filtering of which 18,228 were polymorphic. We focused on these 18,228 PTVs for subsequent analyses.

We used the Genotype–Reference Consortium definition (http://www.ncbi.nlm.nih.gov/gene/human/regions/MHC?asm=GRC37) to define each MHC region. We considered any PTV in this region or within 3,000,000 base pairs of this region (to avoid including PTVs in LD with the UK Cancer Register with self-reported diagnoses from the UK Biobank questionnaire Data-Field 40006) and the UK Cancer Register with self-reported non-cancer illness codes Data-Field 41202. We annotated 784,257 autosomal variants with MAF > 0.01% in each of the 17 non-MHC genes that had a significant PTV from the GTEx GWAS. All genes except for IRSF had at least one missense variant. We then performed association analyses as described above for the missense variants from each gene and the phenotypes that PTVs in that gene were associated with. We considered significant any missense-phenotype associations with nominal p < 0.001. We repeated the association analysis using the GTEx phenotype as a covariate to evaluate whether the association signals were independent for significant missense variants.

**HLA conditional analysis.** We performed conditional association analyses for 47 of the 74 significant associations from our GWAS for PTVs in genes in or near the MHC using the HLA alleles provided by the UK Biobank (ukb_hla_v2.txt). For each PTV-phenotype association, we re-ran the association analysis using each of the HLA alleles polymorphic in the 337,205 subjects used here as a covariate in turn. We then identified which HLA allele, when used as a covariate, corresponded to the largest p value for the additive genetic effect. These results are reported in Supplementary Data 4. Note that this HLA allele is not necessarily the associated with the reported trait since LD exists between different HLA alleles.

**ANKD18B conditional analysis.** In our initial GWAS, we found associations between the PTV rs34358 in ANKD18B and family history of diabetes and high cholesterol. Since ANKD18B is near HMGCR, we performed a conditional association analysis between rs34358 and family history of diabetes and high cholesterol using the imputed genotypes for rs17328484, an intrinsic variant in HMGCR associated with cholesterol levels17, as covariates. We found that conditioning on rs17328484 made the association between rs34358 and high cholesterol insignificant (p = 0.052) but that the association between rs34358 and family history was only slightly reduced from p = 1.5 × 10–5 to p = 9.1 × 10–5. We therefore decided to include this association in Supplementary Data 3.

**Conidential analysis.** We performed conditional analyses for each of the 27 PTVs outside of the MHC with significant associations. We identified all variants genotyped on the UK Biobank array within 10 kb of the PTVs that passed filtering and had MAF > 0.01. For each variant within 10 kb of a PTV, we ran a logistic regression as described above using PLINK but added the genotype of the nearby variant as a covariate. For each PTV-phenotype association, we identified which nearby variant resulted in the largest p value for the PTV association. We report this nearby variant (cond_variant), p value for the PTV association (cond_p), and the MAF of the nearby variant (cond_maf) in Supplementary Data 3. For Supplementary Fig. 5, we plotted the linkage disequilibrium (LD) between the PTV and nearby variants (minimum LD 0.9) for PTVs with MAF > 1% and for which conditional analysis identified a nearby variant that reduced the p value by at least 0.05. For each locus, we report the magnitude of LD26 and the p value between the PTV and the variant with the lowest p value for each locus. For rs2040640 in IRSF, we plotted variants with LD > 0.6. LD values were calculated using the same UK Biobank subjects used for the GWAS.
Imputed PTVs GWAS. We identified 962 PTVs among the UK Biobank imputed genotypes that were not multi-allelic, had MAF > 0.01, and were not already included in our study by comparing the chromosomal coordinates and alternative alleles of PTVs annotated in gnomAD to the UK Biobank positions and alleles for the UK Biobank data. We only considered PTVs in the HRC site list version 1.1 (http://www.haplotype-reference-consortium.org/site). We removed 408 imputed PTVs that had an imputation score < 0.8, missingness > 1%, or whose MAF differed substantially from the non-Finnish European MAF in gnomAD. We removed eight more imputed PTVs that were in genes near the MHC. In total we were left with 546 imputed PTVs that we stratified into the following MAF bins: 0.01–0.01% (247 PTVs), 0.1–1% (153 PTVs), and > 1% (146 PTVs). We corrected p values separately for each bin using the Yekutieli approach implemented in R’s p.adjust. We assessed linkage disequilibrium between imputed PTVs and other variants using LDmatrix in LDLink.

For the missense variant rs140185678 (MAF = 0.0363) in RPL3L, we ran GWAS as described above and found that the variant was associated with atrial fibrillation (p = 5.4 × 10⁻⁶, OR = 1.21, 95% CI: 1.14–1.30) and atrial flutter (p = 1.8 × 10⁻⁵, OR = 1.20, 95% CI: 1.12–1.28). We re-reference analysis using the genotype of the RPL3L PTV rs140192228 as a covariate and found that the association between rs140185678 and atrial fibrillation (p = 4.3 × 10⁻⁸, OR = 1.21, 95% CI: 1.14–1.29) and atrial flutter (p = 8.4 × 10⁻⁶, OR = 1.20, 95% CI: 1.12–1.28) were still significant. The PTV was also significant under these models for atrial fibrillation (p = 1.1 × 10⁻⁷, OR = 1.10, 95% CI: 0.79–0.91) and atrial flutter (p = 2.1 × 10⁻⁶, OR = 0.84, 95% CI: 0.78–0.90).

Phenome-wide association analyses. We performed pheWAS on the 17 genes with at least one significant association in our GWAS as well as 8 genes reported to have protective genetic associations: CARD9, RNF186, IL23R, ANGPTL4, PCSK9, LPA, APOC3, and MAFbx. We generated associations between PTVs in these genes with MAF greater than 0.01% and 135 medical phenotypes from UK Biobank, the UK Biogem, and the NHLBI Exome Sequencing Project. We found 250 PTVs per subject on average and 1173 genes with at least one PTV per subject. We included all PTV associations with p < 0.01 in Supplementary Data 3 ("all_phewas" tab).

NOL3 siRNA knockdown in human skeletal muscle cells. Adult human skeletal muscle cells (150–05A, Sigma-Aldrich) were cultured in skeletal muscle cell growth medium (Sigma-Aldrich). For differentiation, cells were dissociated using 0.25% trypsin-EDTA (Gibco) and cultured in skeletal muscle cell differentiation medium (Sigma-Aldrich). After 24 h, differentiation was initiated by changing medium to skeletal muscle cell differentiation medium. After 24 h, differentiation was initiated by changing medium to skeletal muscle cell differentiation medium (Sigma-Aldrich), which was subsequently exchanged every second day. We transfected differentiating skeletal muscle cells in DMEM/F-12 medium with 30 pmol siRNA against NOL3 (s301, Thermofisher Scientific) or a scramble control siRNA (ThermoFisher Scientific) using the Lipofectamine RNAiMAX Transfection reagent (ThermoFisher Scientific) according to the manufacturer’s protocol.

Electrical pulse stimulation. Four days after siRNA-treatment, we electrically stimulated skeletal muscle cells using a C-pose unit and a six-well C-Dish without supply of an electric current. Another 5 s delay for a total of 5 h. Unstimulated cells were exposed to the six-well C-Dish without supply of an electric current.

RNA isolation and qRT-PCR. Total RNA was extracted by the Trizol method. In all, 300 ng of total RNA was used to generate cDNA through the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems) in a total volume of 20 µl. Gene expression was quantified using standard TaqMan gene expression assays (NOL3 Hs01126088_g1; ACTB Hs01060665_g1; MAFbx/FBXO32 Hs01041408_m1; Thermofisher Scientific).

Protein analyses. Fifteen mg of total protein was loaded onto a 4–15% polyacrylamide gel (Bio-Rad), separated and subsequently transferred onto a PVDF membrane (Merck Millipore). The membrane was blocked in Odyssey blocking buffer (LI-COR), incubated overnight with primary antibodies; Fbx20/MAbFbx 1.500 (ab168372 Lot: GR822135_2, Abcam) and GAPDH 1:1000 (sc-48167 Lot: B2112, Santa Cruz Biotechnology) as a loading control. After washing and incubation with the appropriate fluorescent secondary antibody (a-rabbit 1:5000 (925–32211 Lot: C70926_01 and a-goat 1:5000 925–32214 Lot: CS0330–07), the membranes were imaged and protein quantified using the LI-COR Odyssey Fc imaging system. We measured the activity of caspase 8 as an early apoptotic signal inhibited by NOL3. A chromogenic Caspase 8 assay kit (ab39700, Abcam) was used according to the manufacturer’s protocol. To increase the efficiency of the homogenization, the homogenate was snap-frozen in liquid nitrogen prior to protein quantification and the centrifugation step performed to remove solid material was done at a lower speed. The assay was quantified by measuring the absorbance of p-nitroanilide chromophore from the sequence Ile-Glu-Thr-Asp, and the signal was measured at OD = 405 nm. The resulting values were related to the total protein content measured with the Pierce BCA protein assay kit (ThermoFisher Scientific).

DNA fragmentation. We analyzed the level of DNA fragmentation as a measure of the degree of apoptosis induced by the stimulation. The Cell Death Detection ELISA plus kit (Sigma-Aldrich) was used to quantify the level of cytoplasmic histone-associated DNA fragments, according to the manufacturer’s specifications. In brief, the cells were lysed directly in the culture wells, scrapped off the plate and centrifuged at 200 x g. The supernatant (cytoplasmic fraction) was loaded onto a streptavidin-coated microplate and incubated for 2 h with a biotin-labeled histone antibody and a peroxidase-conjugated DNA antibody. An ABTS substrate was subsequently added, and the enzyme-linked immunosorbant assay was read at OD = 405 nm. The level of fragmentation was related to the total amount of protein.

Knockout status. We estimated PTV knockout carrier status for each individual by summing the total number of PTVs present in an individual for each gene that had at least one PTV. If a PTV was predicted to effect more than one gene, we counted that PTV for each gene. If an individual was heterozygote for two different PTVs in the same gene, we considered the individual as a heterozygous KO. If an individual was predicted to carry > 2 PTVs in a given gene, we set his or her count to two. We thus obtained carrier statuses for each gene in each subject that ranged from no KO, heterozygous KO, or homozygous KO. For all 18,228 predicted PTVs, we found 262 PTVs per subject on average and 1173 genes with at least one putative KO. If we restrict to only high confidence PTVs, we observe 174 PTVs per subject on average and 995 genes with at least one putative KO. If we restrict to PTVs with MAF < 1%, we observe 95 PTVs per subject on average and 778 genes with at least one putative KO.

Additivity analyses. To test for departures from additivity, we tested for associations between PTV carrier status and phenotype status for 16 of the 25 genes used in the phEWA analysis that had at least one homozygote knockout and 206 phenotypes with at least 1000 cases. For each gene and phenotype, we fit two models using the glm function in R (family = binomial). For the additive model, we provided PTV carrier status as a numeric variable, and for the non-additive model, we provided a reduced PTV carrier status as a factor. We included age, sex, genotyping array, and the first four principal components as covariates for both models. To identify gene-phenotype associations with suspected departures from additivity, we identified genes and phenotypes where either the additive p value or homozygote KO p value was < 10⁻⁵ and the difference between the non-additive model AIC and additive model AIC was < −1.

URLs. For LDlink, see https://analysistools.ncbi.nih.gov/LDLink/; for gnomAD browser, see http://gnomad.broadinstitute.org/; for UK Biobank, see http://www.ukbiobank.ac.uk/.

Data availability. The UK Biobank data are available through the UK Biobank (http://www.ukbiobank.ac.uk/). Analysis scripts and notebooks are available on Github at https://github.com/rivas-lab/public-resources. GWAS results can be browsed on the Global Biobank Engine (biobankengine.stanford.edu).

Received: 7 September 2017 Accepted: 21 March 2018 Published online: 24 April 2018

References
1. Rivas, M. A. et al. A protein-truncating R179X variant in RNF186 confers protection against ulcerative colitis. Nat. Commun. 7, 12342 (2016).
2. Rivas, M. A. et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. 43, 1066–1073 (2011).
3. Cohen, J. et al. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat. Genet. 37, 161–165 (2005).
4. Cohen, J. H., Boerwinkle, E., Mosley, T. H. Jr. & Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N. Engl. J. Med. 354, 1260–1272 (2006).
5. Saleheen, D. et al. Human knockouts in a cohort with a high rate of consanguinity. Nature 544, 235–239 (2015).
6. Narasimhan, V. M. et al. Health and population effects of rare gene knockouts in adult humans with related parents. Science 352, 474–477 (2016).
7. Sulem, P. et al. Identification of a large set of rare complete human knockouts. Nat. Genet. 47, 448–452 (2015).
8. Rivas, M. A. et al. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science 348, 666–669 (2015).
9. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016).
10. Holbrook, J. A., Neu-Yilik, G., Hentze, M. W. & Kulozik, A. E. Nonsense-mediated decay approaches the clinic. Nat. Genet. 36, 801–808 (2004).
11. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. Cell 169, 1177–1186 (2017).
12. MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science 335, 823–828 (2012).
13. Saleheen, D. et al. Human knockout and phenotypic analysis in a cohort with a high rate of consanguinity. Nature 544, 235–239 (2015).
14. Musumuru, K. & Kathiresan, S. Cardiovascular endocrinology: is ANGPTL3 the next PCSK9? Nat. Rev. Endocrinol. 13, 503–504 (2017).
15. Stützle, N. O. et al. ANGPTL3 deficiency and protection against coronary artery disease. J. Am. Coll. Cardiol. 66, 2396–2403 (2015).
16. Lim, E. T. et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. PLoS Genet. 10, e1004494 (2014).
17. Stützle, N. O. et al. Coding variation in ANGPTL4, LPL, and SVEP1 and the risk of coronary disease. N. Engl. J. Med. 374, 1134–1144 (2016).
18. Bycroft, C. et al. Genome-wide genetic data on ~500,000 UK Biobank participants. Preprint at bioRxiv https://doi.org/10.1101/166298 (2017).
19. Allen, N. E., Sudlow, C., Peakman, T. & Collins, R. UK Biobank data: come and get it. Sci. Transl. Med. 2, 224ed4 (2014).
20. MacArthur, D. G. & Tyler-Smith, C. Loss-of-function variants in the genomes of healthy humans. Hum. Mol. Genet. 19, R125–R130 (2010).
21. Narasimhan, V. M., Xue, Y. & Tyler-Smith, C. Human knockout carriers: dead diseases, healthy, or improved? Trends Mol. Med. 22, 341–351 (2016).
22. MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science 335, 823–828 (2012).
23. Kosmicki, J. A. et al. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. Nat. Genet. 49, 504–510 (2017).
24. Zengini, E. et al. Genome-wide analyses using UK Biobank data provide insights into the genetic architecture of osteoarthritis. Nat. Genet. https://doi.org/10.1038/nature18891-018 (2018).
25. Smith, D. et al. A rare IL13 loss-of-function mutation reduces blood eosinophil counts and protects from asthma. PLoS Genet. 13, e1006659 (2017).
26. Zhao, C.-N. et al. The association of GSDB and ORMDL3 Gene polymorphisms with asthma: a meta-analysis. Allergy Asthma Immunol. Res. 7, 175 (2015).
27. Danesh, J. et al. GSDB inducethat a phenotype characterized by increased airway responsiveness and remodeling without lung inflammation. Proc. Natl. Acad. Sci. 113, 13132–13137 (2016).
28. Irvine, A. D., McLean, W. H. I. & Leung, D. Y. M. Filaggrin mutations associated with skin and allergic diseases. New Engl. J. Med. 365, 1315–1327 (2011).
29. Cassin, A. et al. Current loss-of-function mutations in the thyrotropin receptor gene: when to investigate clinical effects, and treatment. J. Clin. Res. Pediatr. Endocrinol. 4, 29–39 (2012).
30. Wang, Y. et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat. Genet. 46, 736–741 (2014).
31. Surendran, P. et al. Trans-ancestry meta-analyses identify rare and common variants associated with blood pressure and hypertension. Nat. Genet. 48, 1151–1161 (2016).
32. Koski, T., Inohara, N., Chen, S. & Núñez, G. ARC, an inhibitor of apoptosis expressed in skeletal muscle and heart that interacts selectively with caspases. Proc. Natl. Acad. Sci. USA 95, 5156–5160 (1998).
33. Davies, J. L., Kwong, K. R., Kim, J. S. N. & Mallard, J. D. Apoptosis repressor with a CARD domain (ARC) restrains Bax-mediated pathogenesis in dystrophic skeletal muscle. PLoS ONE 8, e82053 (2013).
34. Koski, T., Inohara, N., Chen, S. & Núñez, G. ARC an inhibitor of apoptosis expressed in skeletal muscle and heart that interacts selectively with caspases. Proc. Natl. Acad. Sci. USA 95, 5156–5160 (1998).
35. Bodine, S. C. Identification of ubiquitin ligases required for skeletal muscle atrophy. Science 294, 1704–1708 (2001).
36. Mitchell, A. S. et al. Functional morphological, and apoptotic alterations in skeletal muscle of ARC deficient mice. Apoptosis 20, 310–326 (2015).
Franke, A. et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn’s disease susceptibility loci. Nat. Genet. 42, 1118–1125 (2010).

Tsai, L. C. et al. Large scale meta-analysis characterizes genetic architecture for common psoriasis associated variants. Nat. Commun. 8, 15382 (2017).

Hazra, A. et al. Common variants of FUT2 are associated with plasma vitamin B12 levels. Nat. Genet. 40, 1160–1162 (2008).

Lin, X. et al. Genome-wide association study identifies novel loci associated with serum level of vitamin B12 in Chinese men. Hum. Mol. Genet. 21, 2610–2617 (2012).

He, M. et al. A genome wide association study of genetic loci that influence tumour biomarkers cancer antigen 19-9 carcinoembryonic antigen and α fetoprotein and their associations with cancer risk. Gut 63, 143–151 (2013).

Liang, Y. et al. Genetic variations affecting serum carcinoembryonic antigen levels and status of regional lymph nodes in patients with sporadic colorectal cancer from Southern China. PLoS ONE 9, e97923 (2014).

Ruedi, R. et al. Genome-wide association study of metabolic traits reveals novel gene-metabolite–disease links. PLoS Genet. 10, e1004132 (2014).

Verlaan, D. J. et al. Allele-specific chromatin remodeling in the ZBP1P2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. Am. J. Hum. Genet. 85, 377–393 (2009).

Das, S. et al. GSDMB induces an asthma phenotype characterized by increased airway responsiveness and remodeling without lung inflammation. Proc. Natl. Acad. Sci. USA 113, 13132–13137 (2016).

Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12, e1001779 (2015).

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

McLaren, W. et al. The Ensembl variant effect predictor. Genome Biol. 17, 122 (2016).

Hill, A. et al. Stepwise distributed open innovation contests for software development: acceleration of genome-wide association analysis. Gigascience 6, 1–10 (2017).

Yekutieli, D. & Benjamini, Y. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29, 1165–1188 (2001).

Machiela, M. J. & Chanock, S. J. L-DNAnic: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 31, 3555–3557 (2015).

Orfanos, Z. et al. Breaking sarcomeres by in vitro exercise. Sci. Rep. 6, 19614 (2016).

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 24983 and 16698. We thank all the participants in the UK Biobank study. We would like to thank Stefan Stender for suggesting that the association between the PTVG rs34358 and high cholesterol might be due to the LDL GWAS variant rs17238484 in HMGR (https://gitter.im/UK-Biobank/Lobby). M.A.R. and C.D. are supported by the Stanford University Biomedical Informatics Training Program (T32 LM012409). E.A.A. is supported by 1U24EB023674-01 (MoToPAC) and U01HG007708 (Undiagnosed Diseases Network). The primary and processed data used to generate the analyses presented here are available in the UK Biobank access management system (https://amsportal.ukbiobank.ac.uk/) for application 24983, "Generating effective therapeutic hypotheses from genomic and hospital linkage data" (http://www.ukbiobank.ac.uk/wp-content/uploads/2017/06/24983-Dr-Manuel-Rodriguez.pdf), and the results are deposited in the Global Biobank Engine (https://biobankengine.stanford.edu). We thank the Customer Solutions Team from Paradigm4 who helped us implement efficient databases for queries and application of inference methods to the data.

Author contributions

M.A.R. conceived and designed the study. C.D., Y.T., G.M., A.L., and M.A.R. designed and carried out the statistical and computational analyses. C.C. optimized and implemented computational methods. C.D., Y.T., G.M., A.L., and M.J.D. carried out quality control of the data. Browser features in Global Biobank Engine were led and developed by G.M. and M.A.R., with assistance from A.L., Y.T., and C.D. M.E.L. and E.A.A. designed and carried out NOL3 experiments. E.I., E.A.A., M.J.D. and C.D.B. provided analysis and commented on the manuscript. The manuscript was written by C.D., Y.T., and M.A.R. M.A.R. supervised all aspects of the study.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-03910-9.

Competing interests: C.D.B. is a member of the scientific advisory boards for Liberty Biosecurity, Personalis, 23andMe Roots into the Future, Ancestry.com, IdentityGenomics, and Etalon and is a founder of CDB Consulting. M.J.D. is a member of the scientific advisory board for DisplayGenomics and Prime Genomics. E.I. is a scientific advisor for Precision Wellness and Olink Proteomics for work unrelated to the present project. The remaining authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018