RESEARCH ARTICLE SUMMARY

HUMAN GENETICS

Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR Study

Frederick E. Dewey *et al.*

INTRODUCTION: Large-scale genetic studies of integrated health care populations, with phenotypic data captured natively in the documentation of clinical care, have the potential to unveil genetic associations that point the way to new biology and therapeutic targets. This setting also represents an ideal test bed for the implementation of genomics in routine clinical care in service of precision medicine.

RATIONALE: The DiscovEHR collaboration between the Regeneron Genetics Center and Geisinger Health System aims to catalyze genomic discovery and precision medicine by coupling high-throughput exome sequencing to longitudinal electronic health records (EHRs) of participants in Geisinger's MyCode Community Health Initiative. Here, we describe initial insights from whole-exome sequencing of 50,726 adult participants of predominantly European ancestry using clinical phenotypes derived from EHRs.

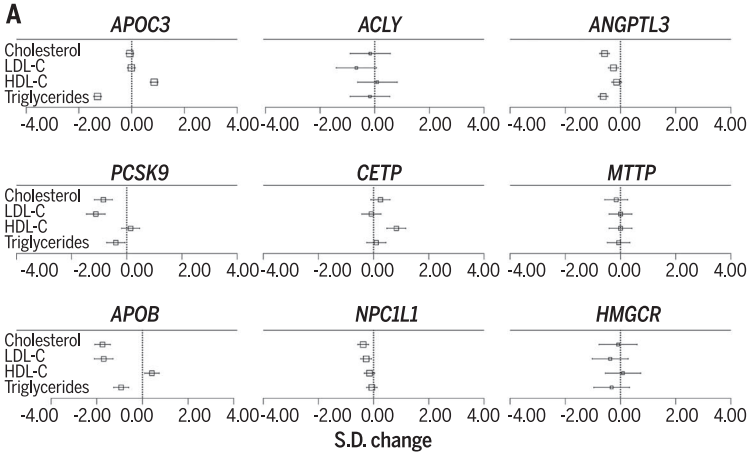
RESULTS: The median duration of EHR data associated with sequenced participants was 14 years, with a median of 87 clinical encounters, 687 laboratory tests, and seven procedures per participant. Forty-eight percent of sequenced individuals had one or more first- or second-degree relatives in the sample, and genome-wide autozygosity was similar to other outbred European populations. We found ~4.2 million single-nucleotide variants and insertion/deletion events, of which ~176,000 are predicted to result in loss of gene function (LoF). The overwhelming majority of these genetic variants occurred at a minor allele frequency of ≤1%, and more than half were singletons. Each participant harbored a median of 21 rare predicted LoFs. At this sample size, ~92% of sequenced genes, including genes that encode existing drug targets or confer risk for highly penetrant genetic diseases, harbor rare heterozygous predicted LoF variants. About 7% of sequenced genes contained rare homozygous predicted LoF variants in at least one individual. Linking these data to EHR-derived laboratory phenotypes revealed consequences of partial or complete LoF in humans. Among these were previously unidentified associations between predicted LoFs in *CSF2RB* and basophil and eosinophil counts, and *EGLN1*-associated erythrocytosis segregating in genetically identified family networks. Using predicted LoFs as a model for drug target antagonism, we found associations supporting the majority of therapeutic targets for lipid lowering. To highlight the opportunity for genotype-phenotype association discovery, we performed exome-wide association analyses of EHR-derived lipid values, newly implicating rare predicted LoFs, and deleterious missense variants in *G6PC* in association with triglyceride levels. In a survey of 76 clinically actionable disease-associated genes, we estimated that 3.5% of individuals harbor pathogenic or likely pathogenic variants that meet criteria for clinical action. Review of the EHR uncovered findings associated with the monogenic condition in ~65% of pathogenic variant carriers' medical records.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aaf6814

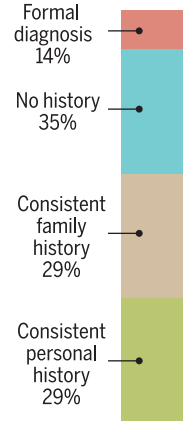
CONCLUSION: The findings reported here demonstrate the value of large-scale sequencing in an integrated health system population, add to the knowledge base regarding the phenotypic consequences of human genetic variation, and illustrate the challenges and promise of genomic medicine implementation. DiscovEHR provides a blueprint for large-scale precision medicine initiatives and genomics-guided therapeutic target discovery. ■

The complete list of authors and affiliations is available in the full article online.
Corresponding author. Email: djcarey@geisinger.edu (D.J.C.); frederick.dewey@regeneron.com (F.E.D.)
Cite this article as F. E. Dewey *et al.*, *Science* 354, aaf6814 (2016). DOI: 10.1126/science.aaf6814



B

	Variant positive/ total	Estimated prevalence
One clinically actionable genetic variant in G76	46/1415	1:31 (3.3%)
Two clinically actionable genetic variants in G76	3/1415	1:472 (0.2%)
Total	**49/1415**	**1:29 (3.5%)**



Therapeutic target validation and genomic medicine in DiscovEHR. (**A**) Associations between predicted LoF variants in lipid drug target genes and lipid levels. Boxes correspond to effect size, given as the absolute value of effect, in SD units; whiskers denote 95% confidence intervals for effect. The size of the box is proportional to the logarithm (base 10) of predicted LoF carriers. (**B** and **C**) Prevalence and expressivity of clinically actionable genetic variants in 76 disease genes, according to EHR data. G76, Geisinger-76.

RESEARCH ARTICLE

HUMAN GENETICS

Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study

Frederick E. Dewey,[1]* Michael F. Murray,[2] John D. Overton,[1] Lukas Habegger,[1] Joseph B. Leader,[2] Samantha N. Fetterolf,[2] Colm O'Dushlaine,[1] Cristopher V. Van Hout,[1] Jeffrey Staples,[1] Claudia Gonzaga-Jauregui,[1] Raghu Metpally,[2] Sarah A. Pendergrass,[2] Monica A. Giovanni,[2] H. Lester Kirchner,[2] Suganthi Balasubramanian,[1] Noura S. Abul-Husn,[1] Dustin N. Hartzel,[2] Daniel R. Lavage,[2] Korey A. Kost,[2] Jonathan S. Packer,[1] Alexander E. Lopez,[1] John Penn,[1] Semanti Mukherjee,[1] Nehal Gosalia,[1] Manoj Kanagaraj,[1] Alexander H. Li,[1] Lyndon J. Mitnaul,[1] Lance J. Adams,[2] Thomas N. Person,[2] Kavita Praveen,[1] Anthony Marcketta,[1] Matthew S. Lebo,[3] Christina A. Austin-Tse,[3] Heather M. Mason-Suares,[3] Shannon Bruse,[1] Scott Mellis,[4] Robert Phillips,[4] Neil Stahl,[4] Andrew Murphy,[4] Aris Economides,[1] Kimberly A. Skelding,[2] Christopher D. Still,[2] James R. Elmore,[2] Ingrid B. Borecki,[1] George D. Yancopoulos,[4] F. Daniel Davis,[2] William A. Faucett,[2] Omri Gottesman,[1] Marylyn D. Ritchie,[2] Alan R. Shuldiner,[1] Jeffrey G. Reid,[1] David H. Ledbetter,[2] Aris Baras,[1] David J. Carey[2]*

The DiscovEHR collaboration between the Regeneron Genetics Center and Geisinger Health System couples high-throughput sequencing to an integrated health care system using longitudinal electronic health records (EHRs). We sequenced the exomes of 50,726 adult participants in the DiscovEHR study to identify ~4.2 million rare single-nucleotide variants and insertion/deletion events, of which ~176,000 are predicted to result in a loss of gene function. Linking these data to EHR-derived clinical phenotypes, we find clinical associations supporting therapeutic targets, including genes encoding drug targets for lipid lowering, and identify previously unidentified rare alleles associated with lipid levels and other blood level traits. About 3.5% of individuals harbor deleterious variants in 76 clinically actionable genes. The DiscovEHR data set provides a blueprint for large-scale precision medicine initiatives and genomics-guided therapeutic discovery.

The application of high-throughput DNA sequencing to human cohorts enables genetic discoveries spanning the development of comprehensive catalogs of rare and common genetic variations (*1*, *2*) and genetic discoveries of previously unidentified Mendelian disease genes (*3*, *4*) and implicates rare variants in common complex diseases (*5*–*7*). Identification of rare loss of gene function variants, or "human knockouts" (*8*–*11*), linked to epidemiological data (*12*) or phenotypes captured in structured research or clinical records (*9*, *10*, *13*, *14*) can facilitate discovery of phenotypic associations that inform human biology and disease pathophysiology. Furthermore, sequencing efforts have identified new therapeutic targets (*15*–*22*), spurring the development of therapeutics for several diseases from lipid disorders to cancer. The implementation of precision medicine requires further investigation of genetic factors that affect health and disease, the development of targeted therapeutics, and further understanding of the utility of genomics for clinical care.

In 2014, the Regeneron Genetics Center, a wholly owned subsidiary of Regeneron Pharmaceuticals, and the Geisinger Health System (GHS), an integrated health system in central and northeastern Pennsylvania, initiated the DiscovEHR collaboration. DiscovEHR strives to elucidate genetic factors that affect a wide range of human diseases and related traits to unveil new biology and drug targets, as well as to scale the implementation of precision medicine by identifying, returning, and acting on clinically actionable genetic variants.

Results

Protein-coding variation in 50,726 exomes

The DiscovEHR cohort is derived from GHS patients who consented to participate in the Geisinger MyCode Community Health Initiative (*23*). MyCode participants consent to provide blood and DNA samples for a system-wide biorepository for broad research purposes, including genomic analyses, and linking to data in the GHS electronic health record (EHR). MyCode participants agree to be recontacted for additional phenotyping and return of clinically actionable results to inform their health care. The DiscovEHR cohort has clinical phenotypes recorded in the GHS EHR over a median of 14 years, with a median of 87 clinical encounters, 658 laboratory tests, and seven procedures captured per participant. Demographics and patient counts for a selection of cardiometabolic, respiratory, neurocognitive, and oncology diseases are described in Table 1.

We sequenced the protein-coding regions of 18,852 genes in 50,726 DiscovEHR participants. Sequence coverage was sufficient to provide at least 20× haploid read depth at >85% of targeted bases in 96% of samples (about 80× mean haploid read depth of targeted bases; fig. S1). We also performed genome-wide array genotyping using the Omni Express exome platform. Per person, we identified a median of 21,409 single-nucleotide variants (SNVs) and 1031 indel variants in protein-coding regions of the genome (fig. S2); a median of 887 variants in each individual was novel. Among all study participants, we identified a total of 4,028,206 unique SNVs and 224,100 unique indels (Table 2), of which 98% occurred at an alternative allele frequency of less than 1%. This abundance of rare variants is consistent with recent accelerated population growth in European-American populations, which comprise an overwhelming majority of the DiscovEHR cohort (fig. S4), and weak purifying selection (*2*, *24*, *25*).

Although our ascertainment protocols did not specifically target families, we expected close familial relationships in this stable regional health care population. We therefore examined the extent of these relationships inferred from whole-exome sequence data using Pedigree Reconstruction and Identification of the Maximally Unrelated Set (PRIMUS) (*26*). Forty-eight percent of sequenced participants had one or more first- or second-degree relatives in the data set (fig. S5), comprising more than 6000 pedigrees with a median pedigree size of two sequenced individuals. The fraction of the autosomal genome existing in runs of homozygosity, which arise from shared parental ancestry, was consistent with previous estimates for European and European-derived populations (*27*) (fig. S6). Collectively, these findings indicate substantial familial substructure in the DiscovEHR cohort, with modest rates of autozygosity similar to other outbred European populations (*27*).

Distribution and clinical impact of predicted loss-of-function variants

Each individual had a median of 21 rare variants predicted to result in a loss of gene function (predicted LoFs or pLoFs) and several hundred more common pLoFs (table S1); an average of 43% of these pLoF variants were frameshift indels, and the remainder were SNVs. We found 176,365 pLoF variants among all study participants, of which 114,340 (65%) are predicted to cause loss of function of all RefSeq (*28*) transcripts. Functionally

[1]Regeneron Genetics Center, Tarrytown, NY 10591, USA. [2]Geisinger Health System, Danville, PA 17822, USA. [3]Laboratory for Molecular Medicine, Cambridge, MA 02139, USA. [4]Regeneron Pharmaceuticals, Tarrytown, NY 10591, USA. *Corresponding author. Email: djcarey@geisinger.edu (D.J.C.); frederick.dewey@regeneron.com (F.E.D.)

Table 1. Demographics and clinical characteristics of adult (≥18 years old) DiscovEHR study population. Unless otherwise noted, values are expressed as median (interquartile range). Diseases are defined by *International Classification of Disease, Ninth Edition* (ICD-9) diagnosis codes.

Basic demographics	GHS active patients	DiscovEHR sequenced patients
N	1,219,522	50,726
Female, *n* (%)	651,248 (53)	30,028 (59)
Age, years	48 (29–65)	61 (48–73)
Body mass index, kg/m^2	27 (23–32)	30 (28–33)
Years of EHR data	5 (1–11)	14 (11–17)
Medication orders per patient	18 (5–65)	129 (37–221)
Laboratory results per patient	116 (38–368)	658 (197–1,119)
Race		
American Indian or Alaska Native, *n* (%)	1,344 (0.1)	51 (0.1)
Asian, *n* (%)	9,625 (0.8)	129 (0.3)
Black or African American, *n* (%)	41,861 (3)	547 (1)
Native Hawaiian or other Pacific Islander, *n* (%)	3,306 (0.3)	41 (0.08)
Other, *n* (%)	6,347 (0.5)	3 (0.01)
Unknown, *n* (%)	23,319 (2)	63 (0.1)
White, *n* (%)	1,133,720 (93)	49,892 (98)
Ethnicity		
Hispanic or Latino, *n* (%)	35,516 (3)	549 (1)
Not Hispanic or Latino, *n* (%)	863,354 (71)	48,477 (96)
Unknown, *n* (%)	320,652 (26)	1,700 (3)
Cardiometabolic phenotypes		
Coronary artery disease, *n* (%)	64,043 (5)	12,298 (24)
Type 2 diabetes, *n* (%)	87,185 (7)	11,474 (23)
Heart failure, *n* (%)	43,807 (4)	5,596 (11)
Bariatric surgery, *n* (%)	6,258 (0.5)	3,112 (6)
Respiratory and immunological phenotypes		
COPD, *n* (%)	55,278 (5)	6,181 (12)
Asthma, *n* (%)	83,901 (7)	7,363 (15)
Rheumatoid arthritis, *n* (%)	10,964 (1)	1,586 (3)
Ulcerative colitis, *n* (%)	4,708 (0.4)	553 (1)
Neurodegenerative phenotypes		
Alzheimer's disease, *n* (%)	6,605 (0.5)	233 (0.5)
Parkinson's disease, *n* (%)	6,513 (0.5)	555 (1)
Multiple sclerosis, *n* (%)	4,349 (0.4)	487 (1)
Myasthenia gravis, *n* (%)	735 (0.06)	90 (0.2)
Oncology phenotypes		
Breast cancer, *n* (%)	15,752 (1)	1,362 (3)
Prostate cancer, *n* (%)	11,268 (1)	1,349 (3)
Lung cancer, *n* (%)	7,398 (0.6)	550 (1)
Colorectal cancer, *n* (%)	7,272 (0.6)	616 (1)

Table 2. Sequence variants identified using whole-exome sequencing of 50,726 DiscovEHR participants.

Variant type	All variants	Allele frequency ≤1%
SNVs	4,028,206	3,947,488
Insertion/deletion variants	224,100	218,785
Predicted loss-of-function variants	176,365	175,393
Nonsynonymous variants	2,025,800	2,002,912
Total	4,252,306	4,166,273

deleterious SNVs and indels exhibited an allele frequency spectrum skewed toward rarity as compared with other SNVs (Fig. 1, A and B): 55.1% of pLoF SNVs and 58.3% of pLoF indels were singletons compared with 46.5% of all SNVs and 49.9% of all indels. The proportion of pLoF variant sites with a derived allele frequency of <0.1% [fraction of rare variants (FRV), 98.5%] was greater than that of missense variants (97.2%) and synonymous variants (95.4%). This is consistent with previous reports of higher FRV for more functionally impactful variants (*13*, *29*). We observed a higher abundance of pLoF variants in the terminal portion of transcripts, suggesting greater tolerance to pLoF variants that result in near–full-length proteins, as previously described (*8*) (fig. S7). Examination of the ratio of observed to possible predicted premature stop mutations (*12*) (Fig. 1, C and D) revealed lower tolerance to pLoF variation in essential genes, cancer-associated genes, and genes associated with autosomal dominant human diseases than in genes associated with autosomal recessive disease genes, drug targets, and olfactory receptors. These results suggest that pLoF variants are under stronger purifying selection compared to variants of other functional classes and that functional context can influence tolerance of genes to pLoF variation.

We estimated the accrual of sequence variants by functional class as sample size grows (Fig. 1, E and F). At the current sample size, rare pLoF variants were observed in 92.4% (17,414) of targeted genes in at least one individual; 15,525 genes (82.4%) harbored rare pLoFs in at least one individual that are predicted to cause loss of function of all protein-coding transcripts of that gene. Homozygous pLoF variants were found in at least one individual in 7.0% of targeted genes (1313 genes), and 868 genes (4.6% of targeted genes) harbored rare pLoFs that affected all transcripts of that gene (Table 3). Of these genes harboring homozygous pLoFs, 654 (49.8%) have not been observed to harbor homozygous pLoFs in other surveys (table S2). This collection of partial and complete human gene knockouts provides opportunities for phenotypic association discovery for highly deleterious gene variants.

To assess the clinical impact of partial or complete loss of gene function in humans, we performed gene-based burden tests of association in mixed linear models with 80 EHR-documented laboratory traits, adjusting for sex, age, and genetic estimates of ancestry and using an experiment-wide significance criterion of 3.3×10^{-8} [0.05/(18,852 genes × 80 traits); tables S3 and S4]. The most statistically significant association for genes harboring at least one homozygous pLoF was between pLoFs in *CSF2RB*, encoding colony-stimulating factor receptor 2β common subunit, and basophils [β = −0.58 standard deviations (SDs) per allele, $P = 8.6 \times 10^{-15}$]. These associations are consistent with loss of function of CSF2RB, the common β chain of the high-affinity receptors for the basophil- and eosinophil-inducing cytokines interleukin-3 (IL-3), IL-5, and the cytokine and myeloid differentiation factor granulocyte-macrophage colony-stimulating

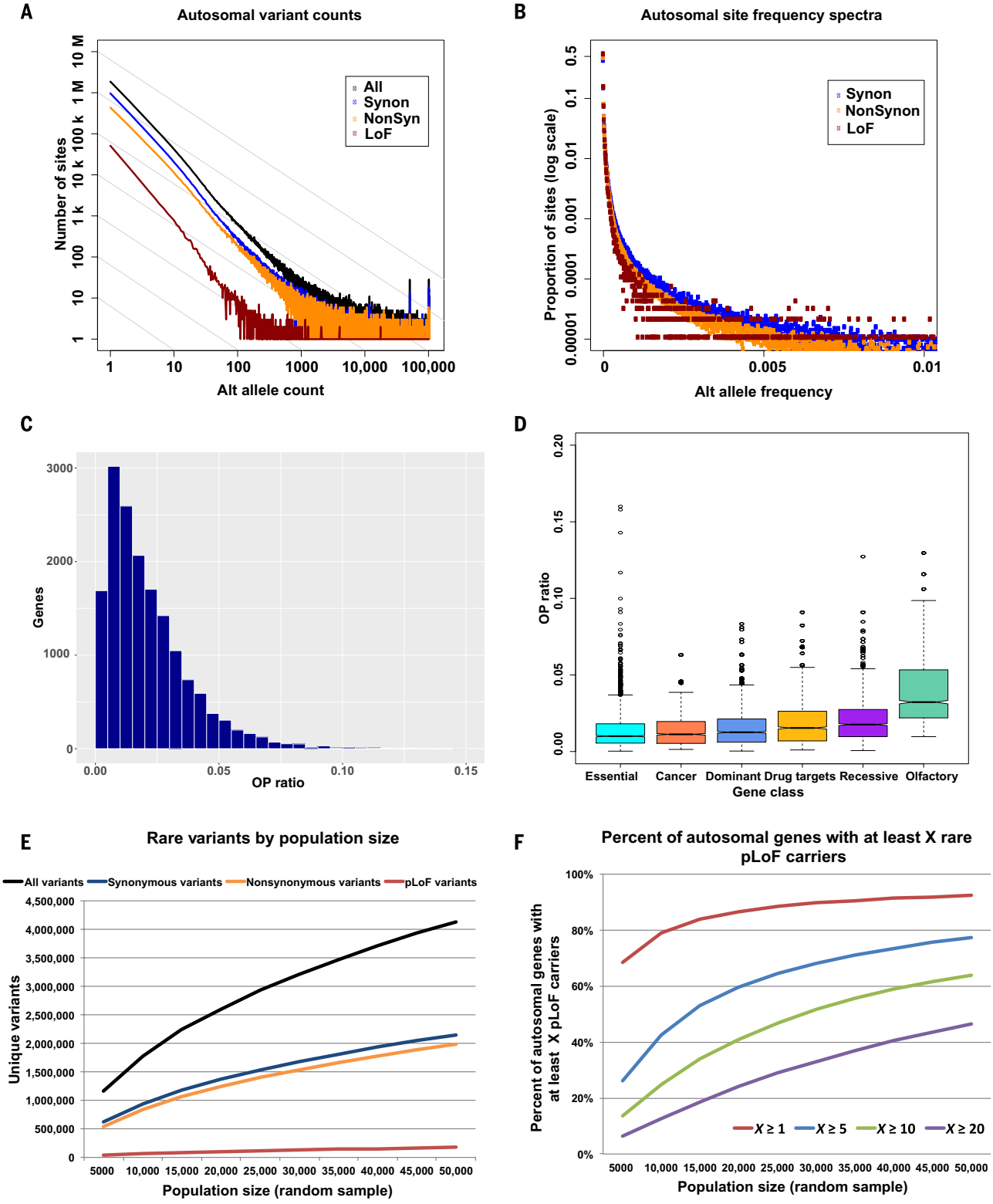


Fig. 1. Frequency and distribution of functional variants in 50,726 exome sequences. (**A**) Relationship between alternate allele count and SNV site number by functional class. (**B**) SNV site frequency spectra by functional class, demonstrating enrichment for rare alleles among more functionally deleterious variants. (**C**) Distribution of observed/predicted (OP) ratio of premature stop variants in 50,726 exome sequences. (**D**) Distribution of the OP ratio of premature stop variants in 50,726 exome sequences by gene class: essential, mouse essential genes (*73*); cancer, cancer predisposition genes (*74*); dominant, autosomal dominant disease genes curated from Online Mendelian Inheritance in Man (OMIM) (*75*, *76*); drug targets, genes encoding targets of 202 drugs (*77*); recessive, autosomal recessive disease genes curated from OMIM; olfactory, olfactory receptor genes. (**E**) Accrual of rare (alternate allele frequency < 1%) variants by functional class. (**F**) Percentage of autosomal genes with multiple predicted loss-of-function (pLoF) carriers as a function of sample size, estimated by randomly sampling the 50,726 sequenced individuals in increments of 5000, creating 10 samples for each increment.

Table 3. Number of genes affected by predicted loss-of-function variants with an allele frequency of ≤1% in 50,726 DiscovEHR participants.

Number of participants	Number of genes affected (%)		
	All, *N* (%)	Heterozygotes, *n* (%)	Homozygotes, *n* (%)
≥1	17,414 (92)	17,409 (92)	1313 (7)
≥5	14,608 (77)	14,598 (77)	312 (2)
≥10	12,105 (64)	12,093 (64)	161 (1)
≥20	8,815 (47)	8,803 (47)	81 (0.4)

factor (*30*). Homozygous truncating variants in *CSF2RB* have been identified in individuals with pulmonary surfactant metabolism dysfunction-5 [MIM# 614370], characterized by pulmonary alveolar proteinosis (PAP) in which there is excessive deposition of extracellular basophilic globular material (*31*, *32*). Review of structured clinical data (problem list entries, encounter diagnosis codes, procedures, medications, and family history) from EHRs of a single pLoF homozygote revealed diagnoses of chronic cough and pulmonary mycobacterial infection, which are common manifestations of PAP (*33*, *34*).

Among genes harboring only heterozygous pLoF variants, we observed experiment-wide significant associations with calcium (*CASR*), thyrotropin (*TG* and *TSHR*), hepatic transaminases (*GPT* and *GOT1*), alkaline phosphatase (*GPLD1* and *ASGR1*), bilirubin (*SLC01B3*), and hematological traits (*TET2*, *GMPR*, *TUBB1*, *ASXL1*, *HBB*, and *EGLN1*), in addition to lipid associations highlighted below. The association between pLoF variants in *EGLN1*, encoding egl-9 family hypoxia-inducible factor 1, and hemoglobin (β = 1.24 SD, $P = 5.4 \times 10^{-10}$) and hematocrit (β = 1.28 SD, $P = 1.4 \times 10^{-10}$) was driven by a single, previously unidentified frameshift variant, p.Pro165fs, carried by 22 individuals. Seven of these p.Pro165fs heterozygotes cosegregated with erythrocytosis in pedigrees reconstructed from exome sequence data, consistent with the gene's described role in familial erythrocytosis 3 [MIM# 609820] (*35*, *36*) (fig. S8). These examples illustrate the utility of pairing pLoF variant discovery with comprehensive EHR-derived clinical phenotypes to understand gene function in humans.

Exome-wide association discovery for serum lipids

To further explore phenotypic associations with coding variants in the DiscovEHR population, we performed an exome-wide association study for fasting lipid levels [total cholesterol, high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), and triglycerides], which are risk factors for ischemic vascular diseases, such as coronary artery disease and stroke. A total of 39,087 participants of European-American ancestry from the DiscovEHR cohort had recorded fasting lipid levels with a median of 6 measurements per individual. We performed a mixed linear model analysis of the association between rank inverse normal transformed residuals of median EHR-documented lipid levels (adjusted for lipid-lowering medication use, sex, age, and genetic estimates of ancestry) and 160,341 biallelic single variants with a minor allele frequency of >0.1%. Using an exome-wide significance criterion of $P < 1 \times 10^{-7}$, we identified 58 SNVs or indel variants (30 nonsynonymous or predicted RNA splice disrupting) in 17 loci with exome-wide significant associations with total cholesterol, 64 variants (29 nonsynonymous or splice) in 21 loci with exome-wide significant associations with HDL-C, 59 variants (27 nonsynonymous or splice) in 14 loci with exome-wide significant associations with LDL-C, and 66 variants (30 nonsynonymous or splice) in 14 loci with exome-wide significant associations with triglycerides (figs. S9 to S12 and tables S5 to S8).

Consistent with other reports (*14*, *37*, *38*), we observed an inverse association between allele frequency and effect size (Fig. 2A) and were able to find, for example, three independent exome-wide significant associations with lipid levels for rare single variants: rs138326449-A in *APOC3* (IVS2+1G>A, allele frequency of 0.2%), which was associated with lower triglyceride levels (β = −1.27 SD, $P = 1.4 \times 10^{-52}$) and higher HDL-C levels (β = 0.85 SD, $P = 4.3 \times 10^{-24}$); rs12713843-T in *APOB* (p.Arg1128His, allele frequency of 0.5%) associated with lower LDL-C levels (β = −0.33 SD, $P = 9.4 \times 10^{-10}$) and lower total cholesterol levels (β = −0.30 SD, $P = 2.0 \times 10^{-8}$); and rs72658867-A (allele frequency of 0.1%), an intronic variant in *LDLR* associated with lower LDL-C levels (β = −0.30 SD, 1.4×10^{-14}) and lower total cholesterol levels (β = −0.27 SD, $P = 7.1 \times 10^{-12}$), which corroborates a recent report of a similar association with LDL-C levels for this rare variant (*14*).

We also performed gene-based burden tests of association in mixed linear models for pLoFs and predicted deleterious nonsynonymous variants. This analysis led to the identification of previously unidentified rare alleles in known lipid-associated gene loci (table S9), which, in aggregate, achieved exome-wide levels of significance for gene-based burden testing ($P < 1 \times 10^{-6}$). Among these was an association between heterozygous pLoF or deleterious missense variants in 994 individuals in *CD36*, a broadly expressed membrane glycoprotein that serves as a receptor for various ligands, including oxidized lipoproteins and fatty acids (*39*), and HDL-C levels (β = 0.20 SD, $P = 3.4 \times 10^{-7}$). A role in HDL-C uptake in the liver has been proposed by studies of Cd36-deficient mice (*40*), and common variation at the *CD36* locus has been associated with HDL-C levels in African Americans (*41*, *42*). Our results provide further evidence of a role for CD36 in the modulation of HDL-C levels in individuals of European ancestry.

An association between *G6PC* and lipid levels was newly implicated by the burden test: 288 heterozygous carriers of 36 pLoF or predicted deleterious missense variants in *G6PC* had significantly higher triglyceride levels (β = 0.35 SD, $P = 5.2 \times 10^{-7}$; lipid values by variant in Fig. 2B). *G6PC* encodes glucose-6 phosphatase catalytic subunits. Homozygous and compound heterozygous mutations in *G6PC* are associated with glycogen storage disease type Ia [MIM# 232200], a canonically autosomal recessive disease that is characterized by lipid and glycogen accumulation in the liver and kidneys accompanied by hypoglycemia, lactic acidosis, hyperuricemia, and hyperlipidemia (*43*). Gene-based burden tests of association between *G6PC* and 709 ICD-9 (*International Classification of Disease, Ninth Edition*)–derived clinical disease phenotype groups with a case frequency of >1% (table S9) identified associations with gout [odds ratio (OR), 2.02; 95% confidence interval (CI), 1.39 to 2.93; $P = 0.0002$] and gouty arthropathy (OR, 2.58; 95% CI, 1.54 to 4.33; $P = 0.0003$), in addition to tension headache (OR, 2.99; 95% CI, 1.67 to 5.33; $P = 0.0002$). Association testing with median uric acid levels extracted from EHR data from 11,540 DiscovEHR participants, of whom 23 carried a pLoF or deleterious missense variant in *G6PC*, identified nominally significantly higher uric acid levels (β = 0.27 SD, $P = 0.01$). Thus, heterozygotes for protein-disrupting variants in *G6PC* appear to manifest an intermediate phenotype characterized by moderate levels of hypertriglyceridemia and increased risk for gout and gouty arthropathy.

Loss-of-function variants in lipid drug target genes and lipid levels

Human genetic variants that inactivate genes encoding drug targets may mimic the action of therapeutic antagonism of these targets, thereby providing an experiment of nature that may be used to infer the clinical effects of drug antagonists of that target. We evaluated associations between pLoF variants, aggregated by genes, in nine therapeutic targets of drugs for lipid modification and lipid levels extracted from the EHR (Fig. 3 and table S11). Of these drug target genes, six of nine harbored pLoF variants that were at least nominally associated with changes in lipid levels, recapitulating clinical effects of the therapeutic agent [15 total nominal ($P < 0.05$) associations among 36 tests]. Among all gene-based burden association tests between pLoFs and lipid levels ($n = 75{,}408$ tests), we observed 4335 nominal associations. Thus, under the null expectation of no association between pLoFs in lipid drug genes and lipid levels, we expected 2 nominally significant associations among 36 association tests, demonstrating ~7.5-fold enrichment for nominal associations with lipid drug target genes ($P = 4.3 \times 10^{-10}$ by exact binomial test).

Among currently approved therapeutics, these observations confirm associations between rare pLoF variants in *NPC1L1* (n = 137 heterozygotes), which encodes the target of ezetimibe, and *PCSK9* (n = 49 heterozygotes), which encodes the target of alirocumab and evolocumab and reduced LDL-C levels (*16*–*18*, *44*), mirroring the clinical effects of therapeutic antagonism of these genes; pLoFs in *PCSK9* were associated with the greatest reduction in LDL-C levels. We also observed highly statistically significant associations between heterozygous pLoF variants in *APOB* and reduced LDL-C and triglyceride levels among 58 pLoF carriers, recapitulating the effect of therapeutic antagonism with mipomersen, an antisense oligonucleotide to apo-B100 (*45*, *46*). Homozygous or compound heterozygous truncating mutations in *APOB* have been implicated in familial hypobetalipoproteinemia-1 [MIM# 615558], characterized by profound depression of apolipoprotein B (apoB)–containing lipoproteins, including LDL-C and triglyceride-rich lipoproteins, and hepatic triglyceride accumulation due to decreased secretion (*47*). Association testing with median alanine and aspartate aminotransferase levels revealed higher levels of both hepatic transaminases (β = 0.12 SD, P = 0.08 and β = 0.27, P = 0.02 for alanine and aspartate aminotransferase levels, respectively) in pLoF variant carriers compared to noncarriers. Consistent with autosomal codominant transmission of certain clinical features of hypobetalipoproteinemia, *APOB* pLoF heterozygotes in the DiscovEHR cohort manifest an intermediate phenotype characterized by moderate depression of LDL-C and triglyceride levels and elevated transaminase levels, suggesting hepatocyte injury. These associations mirror the clinical effects of mipomersen, including elevation in hepatic transaminases in some treated patients (*45*, *46*). In contrast, 29 individuals from the DiscovEHR cohort who were heterozygous for predicted loss-of-function mutations in *MTTP* did not have lipid levels that were significantly different from noncarriers, suggesting that *MTTP*-associated abetalipoproteinemia [MIM# 200100] segregates exclusively as a recessive trait in our study population.

We observed an association between pLoF variants in *CETP*, encoding the target of anacetrapib (currently in phase 3 clinical trials), and higher HDL-C (β = 0.82 SD, P =2.9 × 10^{-6}). Two of three genes encoding targets of therapeutic agents currently in phase 2 clinical trials for lipid modification (*APOC3* and *ANGPTL3*) harbored pLoFs that were associated with lipid profiles, recapitulating therapeutic effects. Nine heterozygotes for pLoF variants in *ACLY*, a target gene of the ACLY-antagonist bempedoic acid in phase 2 clinical trials for lipid lowering, had a trend toward lower LDL-C values (β = −0.67 SD, P = 0.07). These findings suggest that coding variation coupled to EHR-derived clinical phenotypes can recover associations that validate drug targets, anticipate on-target adverse effects, and potentially reveal previously unidentified targets for therapeutic development.

Prevalence of clinically returnable genetic findings in 50,726 exomes

Exome sequence data were analyzed to identify potentially pathogenic variants, according to the ClinVar "pathogenic" classification (*48*), in a subset of 76 genes [Geisinger-76 (G76)] that, when altered, lead to clinically actionable findings for 27 medical conditions (table S12). The G76 includes the 56 genes and 25 conditions identified in the American College of Medical Genetics and Genomics (ACMG) recommendations for identification and reporting of clinically actionable genetic findings (*49*), as well as an additional 17 genes associated with the same 25 conditions and 3 genes associated with 2 additional conditions. All genes and conditions were chosen on the basis of being associated with clinically actionable and highly penetrant monogenic disease. In addition to identifying variants documented to be pathogenic in ClinVar, we identified pLoF variants that were potentially pathogenic ("expected pathogenic"), as recommended by the ACMG guidelines (*49*).

In aggregate, 6653 individuals (13% of sequenced participants) harbored one or more such pathogenic or expected pathogenic variants in the G76 gene list: 5435 individuals with at least one variant

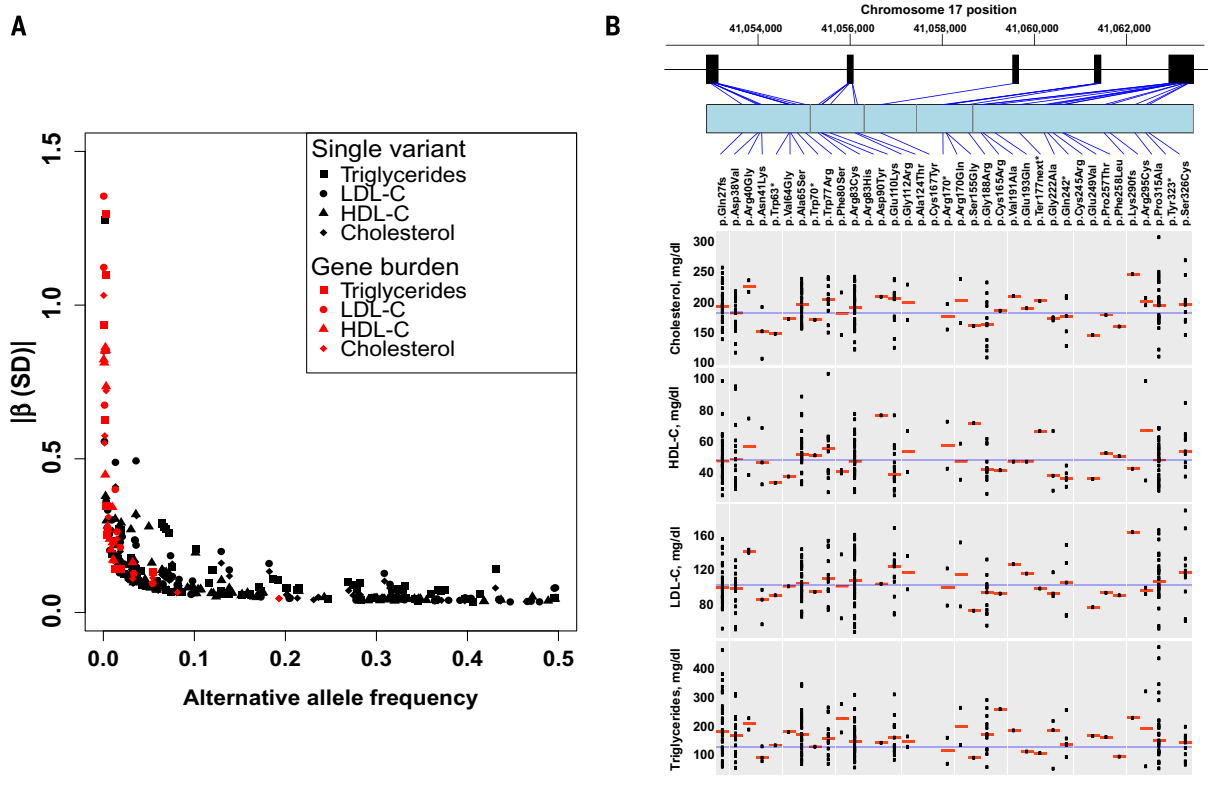


Fig. 2. Exome-wide association discovery for fasting lipid levels. (**A**) Relationship between allele frequency and effect size for single-variant and gene burden tests of association with lipid levels. Effect size is given as the absolute value of β, in SD units. Only single-variant and gene-based burden associations meeting exome-wide significance criteria (1 × 10^{-7} and 1 × 10^{-6} for single-variant and gene-based burden tests of association) are displayed. (**B**) Lipid values by variant for predicted loss of function and predicted deleterious missense variants in *G6PC*. Red lines indicate the median values for variant carriers. Each dot represents a trait value for a single carrier of the variants specified above each box. The blue line indicates the median value for all sequenced individuals not carrying a predicted loss of function or predicted deleterious missense variant in *G6PC*. To convert values for cholesterol to millimoles per liter, multiply by 0.0259. To convert the values for triglycerides to millimoles per liter, multiply by 0.0113.

in these genes with a pathogenic classification in ClinVar and 1218 additional participants with predicted pathogenic pLoF variants. A pilot set of 1415 sequence files then underwent clinical curation, applying clinical laboratory standards (*50*) for potential return to clinical care. Within this pilot set, 641 variants in the G76 were reviewed: 49 were considered either "pathogenic" (n = 32, 5.0%) or "likely pathogenic" (n = 17, 2.7%), and the remaining 592 (93.3%) were considered either variants of uncertain significance, likely benign, benign, or false positives. Of the variants classified in the pilot sample as pathogenic or likely pathogenic by a Clinical Laboratory Improvement Amendments–certified molecular diagnostic laboratory, 43 reportable variants (all either heterozygous for autosomal dominant conditions or hemizygous for X-linked conditions) occurred in 49 of the 1415 participants' samples; 3 individuals carried 2 such variants associated with two distinct diseases (table S13). Therefore, 3.5% of the DiscovEHR cohort participants are estimated to have a variant from the G76 that meets or exceeds current clinical standards for asserting pathogenicity, namely, >90% certainty of the variant being disease-causing (*50*). We investigated the expressivity of these variants by reviewing structured EHR-derived clinical phenotype data for variant carriers. Collectively, 32 of 49 individuals (65%) had clinical features in the EHR that were consistent with the associated disease: 7 (14%) individuals had a diagnosis code entry consistent with a formal diagnosis of the associated disease, and an additional 26 (53%) individuals had diagnosis codes or family history entries consistent with clinical features of the associated disease (table S13). In a companion publication, we describe the prevalence, familial segregation, and clinical impact of variants associated with familial hypercholesterolemia, highlighting the opportunity for early clinical intervention afforded by genetic diagnosis of this condition (*51*). These results demonstrate the potential for genomics-guided clinical care in a large unselected clinical population, establish an expectation for the burden of actionable genetic findings, and support the need for expert clinical review and adjudication of assertions of pathogenicity for potentially pathogenic variants, including those cataloged in mutation databases.

Conclusions

The findings reported here demonstrate the value of large-scale sequencing in a clinical population from an integrated health system and add to the knowledge base regarding human genetic variation provided by other large-population genetic surveys (*1*, *2*, *52*). Megaphenic variants and known pathogenic alleles of clinical relevance have been observed in the protein-coding regions of the genome (*3*, *49*, *53*). As purifying selection ensures that these variants are kept at very low frequencies, very large populations are needed to identify rare variants with large effects on clinical traits (*54*). The discovery of pLoF and other functionally impactful variants in nearly all genes, coupled with structured clinical data in EHRs, provides a rich resource to study the phenotypic effects of partial and complete gene knockouts in humans, as well as other forms of genetic variation. These data may inform discovery of previously unidentified biological mechanisms and therapeutic targets and facilitate interrogation of gene-centric hypotheses around specific phenotypic associations of potential clinical value. Although many of these variants are rare individually, in aggregate, they are not uncommon, and their identification and the biological insights gleaned are relevant to our understanding and treatment of both common and rare diseases.

The DiscovEHR collaboration represents a powerful platform for human genetics research and allows us to fill gaps in our knowledge regarding the role of genetic variation in determining health and disease and the implementation of genomic information in routine medical care. Large-scale sequencing initiatives in integrated health care systems, such as the DiscovEHR collaboration, hold great promise for genetic discoveries to advance precision medicine.

Materials and methods

The DiscovEHR collaboration cohort

The DiscovEHR collaboration study cohort is derived from individuals who consented to participate in Geisinger's MyCode Community Health Initiative. As described in detail elsewhere (*23*), the MyCode initiative leverages the resources of Geisinger as an integrated health care delivery

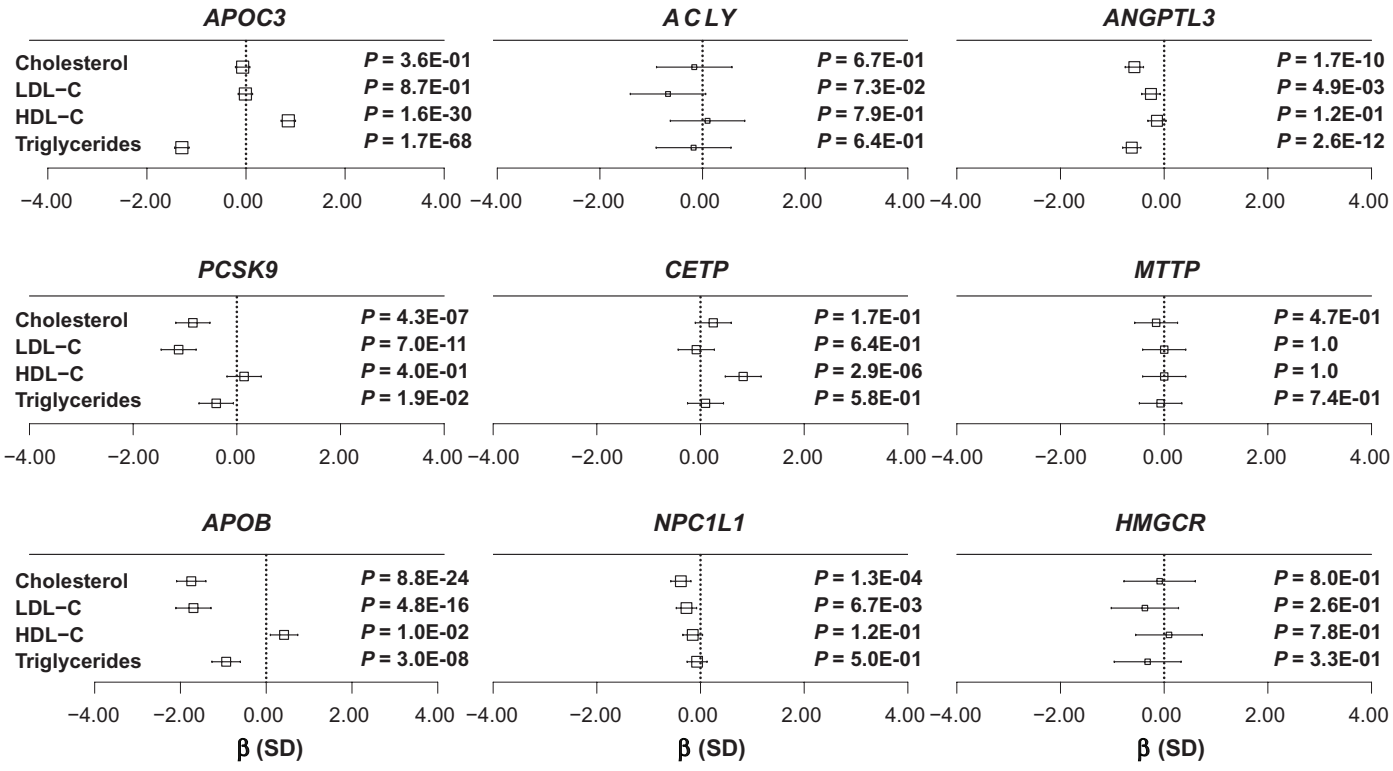


Fig. 3. Associations between predicted loss-of-function variants in lipid drug target genes and lipid levels. Boxes correspond to effect size, in SD units; whiskers denote 95% CIs for β. The size of the box is proportional to the logarithm (base 10) of predicted loss-of-function carriers.

system for implementation of precision medicine. All MyCode participants are enrolled through an opt-in informed consent process; participants agree to provide samples for broad, future research use, including genetic analysis, and to linking of samples and data to information in their GHS EHR; participants agree to be contacted in the future to receive additional information or to invite them to participate in additional research projects; and participants agree to receive clinically actionable findings, the sharing of that information with their clinical providers, and placement of the information in their EHR. Geisinger uses a "data-broker" system to protect confidentiality. Participants are informed that their samples or information might be shared with external research collaborators in a manner that would protect their privacy and confidentiality of their information. To date, more than 125,000 GHS patients have consented to enroll in MyCode, with an 86% enrollment rate for those invited to participate. Enrollment is ongoing.

The DiscovEHR cohort that provides the basis for the findings reported here includes the first 50,726 adult MyCode participants from whom DNA samples were obtained. This includes 6,672 individuals recruited from the Geisinger cardiac catherization laboratory and 2,785 individuals from the bariatric surgery clinic, with the remaining 41,269 individuals representing an otherwise unselected GHS outpatient population.

Sample preparation and sequencing

Sample preparation and whole-exome sequencing were performed as described (*55*). In brief, sample quantity was determined by fluorescence (Life Technologies) and quality assessed by running 100ng of sample on a 2% pre-cast agarose gel (Life Technologies). The DNA samples were normalized and one aliquot was sent for genotyping (Illumina, Human OmniExpress Exome Beadchip) and another sheared to an average fragment length of 150 base pairs using focused acoustic energy (Covaris LE220). The sheared genomic DNA was prepared for exome capture with a custom reagent kit from Kapa Biosystems using a fully-automated approach developed at the Regeneron Genetics Center. A unique 6 base pair barcode was added to each DNA fragment during library preparation to facilitate multiplexed exome capture and sequencing. Equal amounts of sample were pooled prior to exome capture with NimbleGen probes (SeqCap VCRome). Captured fragments were bound to streptavidin-conjugated beads and non-specific DNA fragments removed by a series of stringent washes according to the manufacturer's recommended protocol (Roche NimbleGen). The captured DNA was PCR amplified and quantified by qRT-PCR (Kapa Biosystems). The multiplexed samples were sequenced using 75 bp paired-end sequencing on an Illumina v4 HiSeq 2500 to a coverage depth sufficient to provide greater than 20x haploid read depth of over 85% of targeted bases in 96% of samples (approximately 80x mean haploid read depth of targeted bases).

Sequence alignment, variant identification, and genotype assignment

Upon completion of sequencing, raw sequence data from each Illumina Hiseq 2500 run was gathered in local buffer storage and uploaded to the DNAnexus platform (*56*) for automated analysis. After upload was complete, analysis began with the conversion of BCL files to FASTQ-formatted reads and assigned, via specific barcodes, to samples using the CASAVA software package (Illumina Inc., San Diego, CA). Sample-specific FASTQ files, representing all the reads generated for that sample, were then aligned to the GRCh37.p13 genome reference with BWA-mem (*57*). The resultant binary alignment file (BAM) for each sample contained the mapped reads' genomic coordinates, quality information, and the degree to which a particular read differed from the reference at its mapped location. Aligned reads in the BAM file were then evaluated to identify and flag duplicate reads with the Picard MarkDuplicates tool (http://picard.sourceforge.net), producing an alignment file (duplicatesMarked.BAM) with all potential duplicate reads marked for exclusion in later analyses.

Variant calls were produced using the Genome Analysis Toolkit (GATK) (*58*). GATK was used to conduct local realignment of the aligned, duplicate-marked reads of each sample around putative indels. GATK's HaplotypeCaller was then used to process the INDEL-realigned, duplicate-marked reads to identify all exonic positions at which a sample varied from the genome reference in the genomic VCF format (GVCF). Genotyping was accomplished using GATK's GenotypeGVCFs on each sample and a training set of 50 randomly selected samples, previously run at the Regeneron Genetics Center (RGC), outputting a single-sample VCF file identifying both SNVs and indels as compared to the reference. Additionally, each VCF file carried the zygosity of each variant, read counts of both reference & alternate alleles, genotype quality representing the confidence of the genotype call, the overall quality of the variant call at that position, and the QualityByDepth for every variant site.

Variant Quality Score Recalibration (VQSR), from GATK, was employed to evaluate the overall quality score of a sample's variants using training datasets (e.g., 1000 Genomes) to assess and recalculate each variant's score, increasing specificity. Metric statistics were captured for each sample to evaluate capture, alignment, and variant calling using Picard, bcftools (http://samtools.github.io/bcftools), and FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc).

Following completion of cohort sequencing, samples showing disagreement between genetically-determined and reported sex (n=143); low quality DNA sequence data indicated by high rates of heterozygosity or low sequence data coverage (less than 75% of targeted bases achieving 20X coverage) (n=181); or genetically-identified sample duplicates (n=222) were excluded (n=494 unique samples excluded). Following these exclusions, 51,298 exome sequences were available for downstream analysis, and we report here findings from exome sequences corresponding to 50,726 individuals who were 18 years of age or older at the time of initial consent. These samples were used to compile a project-level VCF (PVCF) for downstream analysis. The PVCF was created in a multi-step process using GATK's GenotypeGVCFs to jointly call genotypes across blocks of 200 samples, recalibrated with VQSR and aggregated into a single, cohort-wide PVCF using GATK's CombineVCFs. Care was taken to carry all homozygous reference, heterozygous, homozygous alternate, and no-call genotypes into the project-level VCF. For the purposes of downstream analyses, samples with QD < 5.0 and DP < 10 from the single sample pipeline had genotype information converted to 'No-Call', and variants falling more than 20 bp outside of the target region were excluded.

Sequence annotation and identification of functional variants

Sequence variants were annotated with snpEff (*59*) using the Ensembl75 gene definitions to determine their functional impact on transcripts and genes. In order to reduce the number of false-positive pLOF calls related to inaccurate transcript definitions, a "whiteList" set of 56,507 protein-coding transcripts that have an annotated start and stop codon (corresponding to 19,729 genes) were selected as a reference for functional annotations. These transcripts were also flagged to allow for downstream filtering for the following features: a) small introns (<15bp), b) small exons (<15bp), c) non-canonical splice sites (non-"GT/AG" splice sites).

The snpEff predictions that correspond to the "whiteList" filtered transcripts are then collapsed into a single most-deleterious functional impact prediction (i.e., the Regeneron Effect Prediction) by selecting the most deleterious functional effect class for each gene according to the hierarchy in table S1. Predicted loss of function mutations were defined as SNVs resulting in a premature stop codon, loss of a start or stop codon, or disruption of canonical splice dinucleotides; open-reading-frame shifting indels, or indels disrupting a start or stop codon, or indels disrupting of canonical splice dinucleotides (table S14). Predicted loss of function variants that correspond to the ancestral allele or that occur in the last 5% of all affected transcripts were excluded.

Principal components and ancestry estimation

We performed principal component (PC) analysis in PLINK2 (*60*) using the subset of overlapping variant sites (n=6,331) from DiscovEHR whole-exome sequence and the 1000 genomes Omni chip platform. This analysis was further restricted to common (MAF>5%) autosomal variant sites with high genotyping rate (>90%) in both Hardy-Weinberg ($p>1x10^{-8}$) and linkage equilibrium which did not map to the MHC region (n sites after filters = 3,974). Initial calculations were based on 1000 genomes v2 (*61*) samples and DiscovEHR individuals were projected onto these PC's.

To identify a subset of European individuals within DiscovEHR, we constructed a linear model

trained on PCs estimates from 1000 genomes known ancestry groups (EUR, ASN, AFR) using the first three PCs. Thresholds for each model (EUR=0.9, AFR=0.7, ASN=0.8) were applied to determine the best continental ancestry match for each DiscovEHR individual; we designated samples not meeting any of these thresholds as "Admixed". Within the DiscovEHR European population, we calculated a new set of PCs was for the maximum unrelated set of individuals using similar variant filtering criteria. Related individuals within DiscovEHR were subsequently projected onto these PCs. These European-only PCs calculated from unrelated DiscovEHR individuals were used in phenotype association analyses.

Relationship estimation

Pairwise identity-by-decent (IBD) estimates were calculated using PLINK2 (*60*) and were used to reconstruct pedigrees with PRIMUS (*26*). First, we used common variants (MAF >10%) in Hardy-Weinberg-Equilibrium (p-val > 0.000001) to calculate IBD proportions all pairs of samples, excluding individuals with >10% missing variant calls (–mind 0.1) and abnormally low inbreeding-coefficient (-0.15) calculated with the –het option in PLINK. We then removed samples with >100 relatives with pi_hat > 0.1875 if the proportion of relatives with pi_hat >0.1875 was less than 40% of the sample's total relationships determined by a pi_hat of 0.05, and removed all samples with >300 relatives. The remaining samples were grouped into family networks. Two individuals are in the same network if they were predicted to be second-degree relatives or closer. We then ran the IBD pipeline implemented in PRIMUS to calculate accurate IBD estimates among samples within each family network. This approach allowed for better-matched reference allele frequencies to calculate relationships within each family network.

Analysis of runs of homozygosity

Analysis of runs of homozygosity (ROH), which arise from shared parental ancestry in an individual's pedigree, is a powerful approach to estimate the extent of ancient kinship and recent parental relationship within a population. Typically, offspring of cousins have long ROH, commonly over 10 Mb. By contrast, almost all Europeans have ROH of ~2 Mb in length, reflecting shared ancestry from hundreds to thousands of years ago. By focusing on ROH of different lengths, it is therefore possible to infer aspects of demographic history at different time depths in the past (*52*). We used FROH measures to compare and contrast DiscovEHR to populations from 1000 genomes. These measures are genomic equivalents of the pedigree inbreeding coefficient, but do not suffer from problems of pedigree reconstruction. By varying the lengths of ROH that are counted, they may be tuned to assess parental kinship at different points in the past. We used FROH5, the fraction of the autosomal genome existing in ROH over 5Mb in length, reflective of a parental relationship in the last four to six generations, as our metric of autozygosity.

For a subset of DiscovEHR individuals for which Omni HumanOmniExpressExome-8v1-2 genotype data were available (N=34,246), genotypes were merged with 1092 individuals from 1000 genomes phase I. ROH were identified using PLINK2 (*60*). We performed LD-based SNP pruning in windows of 50kb with a step size of 5 variants and r-squared threshold of 0.2. On the pruned subset of variants (n=114,514), we applied the following parameters for calculating ROH: 5 MB window size; a minimum of 100 homozygous SNPs per ROH; a minimum of 50 SNPs per ROH window; one heterozygous and five missing calls per window; a maximum between-variant distance within a run of homozygosity of no more than 1Mb. ROH were identified separately for DiscovEHR individuals and for each 1000 genomes population.

We assessed three features of ROH: (i) number of homozygous segments (average and range, calculated across individuals within a population), (ii) summed segment length (average and range, calculated across individuals within a population) and (iii) FROH5, a genomic measure of individual autozygosity, defined as the proportion of the autosomal genome in ROH in runs of 5 Mb or greater in length)(*52*).

For DiscovEHR individuals, we note a mean FROH5 of 0.0006. For CEU individuals, we note a mean FROH5 of 0.0008. This is consistent with previous estimates for European and European-derived populations, where HapMap CEU individuals also had a mean FROH5 of 0.0008 and English individuals had a mean FROH5 of 0.0001 (*27*). We conclude that as a population as a whole, DiscovEHR individual have level of genomic autozygosity that is lower than CEU and only slightly higher than individuals from England.

Phenotype definitions

ICD-9 based diagnosis codes were collapsed to hierarchical clinical disease groups and corresponding controls using a modified version of the groupings proposed by Denny et al (*62*, *63*). ICD-9 based diagnoses required one or more of the following: a problem list entry of the diagnosis code or an encounter diagnosis code entered for two separate clinical encounters on separate calendar days. Median values for selected serially measured laboratory with serial outpatient measures (table S3) were calculated for all individuals with two or more measurements in the EHR following removal of likely spurious values that were > 3 standard deviations from the intra-individual median value. For the purposes of exome-wide association analysis of serum lipid levels, total cholesterol and LDL-C were adjusted for lipid-altering medication use by dividing by 0.8 and 0.7, respectively, to estimate pre-treatment lipid values based on the average reduction in LDL-C and total cholesterol for the average statin dose (*64*). HDL–C and triglyceride values were not adjusted for lipid-altering medication use. We then calculated trait residuals for all laboratory traits after adjustment for age, age^2, sex, and the first ten principal components of ancestry, and rank-inverse-normal transformed these residuals prior to association analysis.

Association analysis for serum lipid levels and other laboratory values

In single-marker exome-wide association analysis of lipid levels, we analyzed all bi-allelic variants with missingness rates < 1%, Hardy-Weinberg equilibrium p values > $1.0\text{x}10^{-6}$, and minor allele frequency > 0.1%. Genotypes were coded according to an additive genotypic model (0 for homozygous reference, 1 for heterozygous, and 2 for homozygous alternative). To account for population structure from ancestry and relatedness, we used mixed linear models of association to test for associations between single variants and lipid trait residuals, fitting a genetic relatedness matrix (constructed from 39,858 non-MHC markers in approximate linkage equilibrium with minor allele frequency > 0.1%) as a random-effects covariate.

We next used the same statistical testing framework to identify gene-based associations between variants, aggregated over the gene (*65*) with the traits enumerated above. We used three variant sets for this gene-based allele aggregation:

1. Predicted loss of function mutations.
2. Predicted loss of function mutations and non-synonymous variants that were predicted deleterious by consensus of 5/5 algorithms [SIFT (*66*), LRT (*67*), MutationTaster (*68*), PolyPhen2 HumDiv, PolyPhen2 HumVar (*69*)].
3. Predicted loss of function mutations and rare (alternative allele frequency < 1%) non-synonymous variants that were predicted deleterious by at least 1/5 algorithms.

Alleles were coded 0,1,2 for non-carriers, heterozygotes for at least one variant site but not homozygous at any variant site, and homozygotes for at least one variant site in each variant set, respectively. Exome-wide quantile-quantile plots and genomic control lambda values for single-marker and gene-based burden tests are provided in figs. S7 to S10. The tests all appeared to be well-calibrated. GCTA v1.2.4 (*70*) and R version 3.2.1 (R Project for Statistical Computing) were used for all statistical analyses.

Exploratory gene-based association analyses with 709 disease phenotypes with case frequency greater than 1% were performed using logistic regression (to provide point estimates of odds ratios), adjusted for age, age^2, sex, and the first ten principal components of ancestry, and BOLT-LMM (*71*) (to provide p-values), adjusted for age, age^2, sex, and genetic relatedness. Wald 95% confidence intervals were estimated for odds-ratios using standard error estimates back-calculated from p-values from the linear mixed models of association in BOLT-LMM. Alleles were coded as above. For association testing for pLoFs, aggregated by gene, and 80 clinical laboratory trait values, residualized as described above, we used BOLT-LMM.

Identification of expected and known pathogenic variants in genes associated with medically actionable diseases

To estimate the burden of potentially returnable genetic findings in all 50,726 sequenced participants, we extracted all the coding variants

identified in the ACMG 56 recommended gene list (*49*) and additional GHS 20 genes for returnable secondary findings. We cross-referenced these variants with the ClinVar dataset [updated December 2015] restricting to those with a Pathogenic classification and a minor allele frequency of less than 1% in the DiscovEHR population. We also cross-referenced the variants with the Human Gene Mutation Database [HGMD 2015.4] restricting to DM-Disease causing mutations of High-confidence variants only with a MAF <1%. We compiled the list of returnable variants according to the published guidelines for the genes where Expected Pathogenic (EP), comprising non-reported putative loss-of-function (pLoF), and/or Known Pathogenic (KP) variants are recommended for clinically actionable return of results.

For a subset of sequenced samples, we applied an orthogonal workflow for variant discovery, adjudication of asserted pathogenicity, and development of clinical reports. We used a custom pipeline to annotate and prioritize clinically important variants. This pipeline incorporates standard annotations, including variant effect, protein functional predictions, conservation, and allele frequencies, along with public and private databases of variants. Variants within the genes of interest were filtered to extract all known alleles of reported clinical relevance and rare, novel and likely disruptive variants (e.g. frameshift, nonsense, splicing, etc). After potentially important variants were identified, they were subjected to manual assessment.

Manual assessments were performed as previously described (*72*), and were consistent with the process developed through a national effort to establish standardized clinical classification criteria (*50*). Briefly, variants were assessed using a comprehensive evidence review, which included published case-control, genetic and functional data as well as population frequency and predictive assessment data. Classifications were subsequently assigned according to a 5-tier system (pathogenic, likely pathogenic, uncertain significance, likely benign, benign). For autosomal dominant disorders, reported variants were limited to those classified as likely pathogenic or pathogenic for the diseases of interest. For autosomal recessive and X-linked disorders, (i.e. *MUTYH*-associated polyposis), likely pathogenic or pathogenic variants were only reported when present in a homozygous, compound heterozygous, or hemizygous state. All reported variants were confirmed via Sanger sequencing.

REFERENCES AND NOTES

1. 1000 Genomes Project Consortium, G. R. Abecasis *et al*., A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010). doi: 10.1038/nature09534; pmid: 20981092
2. J. A. Tennessen *et al*., Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012). doi: 10.1126/science.1219240; pmid: 22604720
3. J. X. Chong *et al*., The genetic basis of Mendelian phenotypes: Discoveries, challenges, and opportunities. *Am. J. Hum. Genet.* **97**, 199–215 (2015). doi: 10.1016/j.ajhg.2015.06.009; pmid: 26166479
4. Y. Yang *et al*., Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312**, 1870–1879 (2014). doi: 10.1001/jama.2014.14601; pmid: 25326635
5. R. Do *et al*., Exome sequencing identifies rare *LDLR* and *APOA5* alleles conferring risk for myocardial infarction. *Nature* **518**, 102–106 (2015). doi: 10.1038/nature13917; pmid: 25487149
6. H. Holm *et al*., A rare variant in *MYH6* is associated with high risk of sick sinus syndrome. *Nat. Genet.* **43**, 316–320 (2011). doi: 10.1038/ng.781; pmid: 21378987
7. S. Steinberg *et al*., Loss-of-function variants in *ABCA7* confer risk of Alzheimer's disease. *Nat. Genet.* **47**, 445–447 (2015). doi: 10.1038/ng.3246; pmid: 25807283
8. D. G. MacArthur *et al*., A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012). doi: 10.1126/science.1215040; pmid: 22344438
9. P. Sulem *et al*., Identification of a large set of rare complete human knockouts. *Nat. Genet.* **47**, 448–452 (2015). doi: 10.1038/ng.3243; pmid: 25807282
10. E. T. Lim *et al*., Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLOS Genet.* **10**, e1004494 (2014). doi: 10.1371/journal.pgen.1004494; pmid: 25078778
11. V. M. Narasimhan *et al*., Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**, 474–477 (2016). doi: 10.1126/science.aac8624; pmid: 26940866
12. A. H. Li *et al*., Analysis of loss-of-function variants and 20 risk factor phenotypes in 8,554 individuals identifies loci influencing chronic disease. *Nat. Genet.* **47**, 640–642 (2015). doi: 10.1038/ng.3270; pmid: 25915599
13. D. F. Gudbjartsson *et al*., Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015). doi: 10.1038/ng.3247; pmid: 25807286
14. UK10K Consortium, K. Walter *et al*., The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015). doi: 10.1038/nature14962; pmid: 26367797
15. J. Cohen *et al*., Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in *PCSK9*. *Nat. Genet.* **37**, 161–165 (2005). doi: 10.1038/ng1509; pmid: 15654334
16. J. C. Cohen, E. Boerwinkle, T. H. Mosley Jr., H. H. Hobbs, Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006). doi: 10.1056/NEJMoa054013; pmid: 16554528
17. S. Kathiresan, A *PCSK9* missense variant associated with a reduced risk of early-onset myocardial infarction. *N. Engl. J. Med.* **358**, 2299–2300 (2008). doi: 10.1056/NEJMc0707445; pmid: 18499582
18. M. Benn, B. G. Nordestgaard, P. Grande, P. Schnohr, A. Tybjærg-Hansen, PCSK9 R46L, low-density lipoprotein cholesterol levels, and risk of ischemic heart disease: 3 independent studies and meta-analyses. *J. Am. Coll. Cardiol.* **55**, 2833–2842 (2010). doi: 10.1016/j.jacc.2010.02.044; pmid: 20579540
19. T. I. Pollin *et al*., A null mutation in human *APOC3* confers a favorable plasma lipid profile and apparent cardioprotection. *Science* **322**, 1702–1705 (2008). doi: 10.1126/science.1161524; pmid: 19074352
20. TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute, J. Crosby *et al*., Loss-of-function mutations in *APOC3*, triglycerides, and coronary disease. *N. Engl. J. Med.* **371**, 22–31 (2014). doi: 10.1056/NEJMoa1307095; pmid: 24941081
21. A. B. Jørgensen, R. Frikke-Schmidt, B. G. Nordestgaard, A. Tybjærg-Hansen, Loss-of-function mutations in *APOC3* and risk of ischemic vascular disease. *N. Engl. J. Med.* **371**, 32–41 (2014). doi: 10.1056/NEJMoa1308027; pmid: 24941082
22. N. J. Timpson *et al*., A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nat. Commun.* **5**, 4871 (2014). doi: 10.1038/ncomms5871; pmid: 25225788
23. D. J. Carey *et al*., The Geisinger MyCode community health initiative: An electronic health record–linked biobank for precision medicine research. *Genet. Med.* **18**, 906–913 (2016). doi: 10.1038/gim.2015.187; pmid: 26866580
24. W. Fu *et al*., Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013). doi: 10.1038/nature11690; pmid: 23201682
25. W. Fu, R. M. Gittelman, M. J. Bamshad, J. M. Akey, Characteristics of neutral and deleterious protein-coding variation among individuals and populations. *Am. J. Hum. Genet.* **95**, 421–436 (2014). doi: 10.1016/j.ajhg.2014.09.006; pmid: 25279984
26. J. Staples *et al*., PRIMUS: Rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *Am. J. Hum. Genet.* **95**, 553–564 (2014). doi: 10.1016/j.ajhg.2014.10.005; pmid: 25439724
27. C. T. O'Dushlaine *et al*., Population structure and genome-wide patterns of variation in Ireland and Britain. *Eur. J. Hum. Genet.* **18**, 1248–1254 (2010). doi: 10.1038/ejhg.2010.87; pmid: 20571510
28. N. A. O'Leary *et al*., Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016). doi: 10.1093/nar/gkv1189; pmid: 26553804
29. E. Khurana *et al*., Integrative annotation of variants from 1092 humans: Application to cancer genomics. *Science* **342**, 1235587 (2013). doi: 10.1126/science.1235587; pmid: 24092746
30. T. B. van Dijk *et al*., Cloning and characterization of the human interleukin-3 (IL-3)/IL-5/ granulocyte-macrophage colony-stimulating factor receptor βc gene: Regulation by Ets family members. *Blood* **92**, 3636–3646 (1998). pmid: 9808557
31. T. Suzuki *et al*., Hereditary pulmonary alveolar proteinosis caused by recessive CSF2RB mutations. *Eur. Respir. J.* **37**, 201–204 (2011). doi: 10.1183/09031936.00090610; pmid: 21205713
32. T. Tanaka *et al*., Adult-onset hereditary pulmonary alveolar proteinosis caused by a single-base deletion in *CSF2RB*. *J. Med. Genet.* **48**, 205–209 (2011). doi: 10.1136/jmg.2010.082586; pmid: 21075760
33. U. B. Prakash, S. S. Barham, H. A. Carpenter, D. E. Dines, H. M. Marsh, Pulmonary alveolar phospholipoproteinosis: Experience with 34 cases and a review. *Mayo Clin. Proc.* **62**, 499–518 (1987). doi: 10.1016/S0025-6196(12)65477-9; pmid: 3553760
34. L. A. Witty, V. F. Tapson, C. A. Piantadosi, Isolation of mycobacteria in patients with pulmonary alveolar proteinosis. *Medicine* **73**, 103–109 (1994). doi: 10.1097/00005792-199403000-00003; pmid: 8152364
35. C. Ladroue *et al*., *PHD2* mutation and congenital erythrocytosis with paraganglioma. *N. Engl. J. Med.* **359**, 2685–2692 (2008). doi: 10.1056/NEJMoa0806277; pmid: 19092153
36. M. J. Percy *et al*., A family with erythrocytosis establishes a role for prolyl hydroxylase domain protein 2 in oxygen homeostasis. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 654–659 (2006). doi: 10.1073/pnas.0508423103; pmid: 16407130
37. G. M. Peloso *et al*., Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am. J. Hum. Genet.* **94**, 223–232 (2014). doi: 10.1016/j.ajhg.2014.01.009; pmid: 24507774
38. L. A. Lange *et al*., Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am. J. Hum. Genet.* **94**, 233–245 (2014). doi: 10.1016/j.ajhg.2014.01.010; pmid: 24507775
39. R. F. Thorne, N. M. Mhaidat, K. J. Ralston, G. F. Burns, CD36 is a receptor for oxidized high density lipoprotein: Implications for the development of atherosclerosis. *FEBS Lett.* **581**, 1227–1232 (2007). doi: 10.1016/j.febslet.2007.02.043; pmid: 17346709
40. M. Brundert *et al*., Scavenger receptor CD36 mediates uptake of high density lipoproteins in mice and by cultured cells. *J. Lipid Res.* **52**, 745–758 (2011). doi: 10.1194/jlr.M011981; pmid: 21217164
41. M. A. Coram *et al*., Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations. *Am. J. Hum. Genet.* **92**, 904–916 (2013). doi: 10.1016/j.ajhg.2013.04.025; pmid: 23726366
42. C. C. Elbers *et al*., Gene-centric meta-analysis of lipid traits in African, East Asian and Hispanic populations. *PLOS ONE* **7**, e50198 (2012). doi: 10.1371/journal.pone.0050198; pmid: 23236364
43. J. Y. Chou, D. Matern, B. C. Mansfield, Y. T. Chen, Type I glycogen storage diseases: Disorders of the glucose-6-phosphatase complex. *Curr. Mol. Med.* **2**, 121–143 (2002). doi: 10.2174/1566524024605798; pmid: 11949931
44. The Myocardial Infarction Genetics Consortium Investigators, N. O. Stitziel *et al*., Inactivating mutations in *NPC1L1* and protection from coronary heart disease. *N. Engl. J. Med.* **371**, 2072–2082 (2014). doi: 10.1056/NEJMoa1405386; pmid: 25390462
45. G. S. Thomas *et al*., Mipomersen, an apolipoprotein B synthesis inhibitor, reduces atherogenic lipoproteins in patients with severe hypercholesterolemia at high cardiovascular risk: A randomized, double-blind, placebo-controlled trial. *J. Am. Coll.*

Cardiol. **62**, 2178–2184 (2013). doi: 10.1016/j.jacc.2013.07.081; pmid: 24013058
46. F. J. Raal *et al.*, Mipomersen, an apolipoprotein B synthesis inhibitor, for lowering of LDL cholesterol concentrations in patients with homozygous familial hypercholesterolaemia: A randomised, double-blind, placebo-controlled trial. *Lancet* **375**, 998–1006 (2010). doi: 10.1016/S0140-6736(10)60284-X; pmid: 20227758
47. F. K. Welty, Hypobetalipoproteinemia and abetalipoproteinemia. *Curr. Opin. Lipidol.* **25**, 161–168 (2014). doi: 10.1097/MOL.0000000000000072; pmid: 24751931
48. M. J. Landrum *et al.*, ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014). doi: 10.1093/nar/gkt1113; pmid: 24234437
49. R. C. Green *et al.*, ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **15**, 565–574 (2013). doi: 10.1038/gim.2013.73; pmid: 23788249
50. S. Richards *et al.*, 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**, 405–424 (2015). doi: 10.1038/gim.2015.30; pmid: 25741868
51. N. S. Abul-Husn *et al.*, Genetic identification of familial hypercholesterolemia within a single U.S. health system. *Science* **354**, aaf7000 (2016). doi: 10.1126/science.aaf7000
52. 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012). doi: 10.1038/nature11632; pmid: 23128226
53. M. Choi *et al.*, Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 19096–19101 (2009). doi: 10.1073/pnas.0910672106; pmid: 19861545
54. O. Zuk *et al.*, Searching for missing heritability: Designing rare variant association studies. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E455–E464 (2014). doi: 10.1073/pnas.1322563111; pmid: 24443550
55. F. E. Dewey *et al.*, Inactivating variants in *ANGPTL4* and risk of coronary artery disease. *N. Engl. J. Med.* **374**, 1123–1133 (2016). doi: 10.1056/NEJMoa1510926; pmid: 26933753
56. J. G. Reid *et al.*, Launching genomics into the cloud: Deployment of *Mercury*, a next generation sequence analysis pipeline. *BMC Bioinformatics* **15**, 30 (2014). doi: 10.1186/1471-2105-15-30; pmid: 24475911
57. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). doi: 10.1093/bioinformatics/btp324; pmid: 19451168
58. A. McKenna *et al.*, 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; pmid: 20644199
59. P. Cingolani *et al.*, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster strain* w^{1118}; *iso-2*; *iso-3*. *Fly* **6**, 80–92 (2012). doi: 10.4161/fly.19695; pmid: 22728672
60. C. C. Chang *et al.*, Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015). doi: 10.1186/s13742-015-0047-8; pmid: 25722852
61. 100 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). doi: 10.1038/nature15393; pmid: 26432245
62. J. C. Denny *et al.*, 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; pmid: 24270849
63. J. C. Denny *et al.*, PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* **26**, 1205–1210 (2010). doi: 10.1093/bioinformatics/btq126; pmid: 20335276
64. C. Baigent *et al.*, Efficacy and safety of cholesterol-lowering treatment: Prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. *Lancet* **366**, 1267–1278 (2005). doi: 10.1016/S0140-6736(05)67394-1; pmid: 16214597
65. B. Li, S. M. Leal, Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008). doi: 10.1016/j.ajhg.2008.06.024; pmid: 18691683
66. P. Kumar, S. Henikoff, P. C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009). doi: 10.1038/nprot.2009.86; pmid: 19561590
67. S. Chun, J. C. Fay, Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009). doi: 10.1101/gr.092619.109; pmid: 19602639
68. J. M. Schwarz, D. N. Cooper, M. Schuelke, D. Seelow, MutationTaster2: Mutation prediction for the deep-sequencing age. *Nat. Methods* **11**, 361–362 (2014). doi: 10.1038/nmeth.2890; pmid: 24681721
69. I. A. Adzhubei *et al.*, A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010). doi: 10.1038/nmeth0410-248; pmid: 20354512
70. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011). doi: 10.1016/j.ajhg.2010.11.011; pmid: 21167468
71. P. R. Loh *et al.*, Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015). doi: 10.1038/ng.3190; pmid: 25642633
72. H. Duzkale *et al.*, A systematic approach to assessing the clinical significance of genetic variants. *Clin. Genet.* **84**, 453–463 (2013). doi: 10.1111/cge.2013.84.issue-5; pmid: 24033266
73. B. Georgi, B. F. Voight, M. Bućan, From mouse to human: Evolutionary genomics analysis of human orthologs of essential genes. *PLOS Genet.* **9**, e1003484 (2013). doi: 10.1371/journal.pgen.1003484; pmid: 23675308
74. N. Rahman, Realizing the promise of cancer predisposition genes. *Nature* **505**, 302–308 (2014). doi: 10.1038/nature12981; pmid: 24429628
75. R. Blekhman *et al.*, Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.* **18**, 883–889 (2008). doi: 10.1016/j.cub.2008.04.074; pmid: 18571414
76. J. S. Berg *et al.*, An informatics approach to analyzing the incidentalome. *Genet. Med.* **15**, 36–44 (2013). doi: 10.1038/gim.2012.112; pmid: 22995991
77. M. R. Nelson *et al.*, An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100–104 (2012). doi: 10.1126/science.1217876; pmid: 22604722

ACKNOWLEDGMENTS

We thank the MyCode Community Health Initiative participants for their permission to use their health and genomics information in the DiscovEHR collaboration. The study was funded by Regeneron Pharmaceuticals. In addition to employment by Regeneron Pharmaceuticals, G.D.Y. is a cofounder, member of the board of directors, and stockholder in Regeneron Pharmaceuticals. J.G.R., O.G., and L.H. are listed inventors on three related provisional patent applications filed by Regeneron Pharmaceuticals (62/314,684; 62/362,660; and 62/404,912), which disclose computer systems and methods of generating, storing, and viewing genetic variant data, phenotype data, and genetic variant phenotype association results. 62/314,684 discloses methods of generating genetic variant association results, systems for viewing genetic variant data, phenotypic data and association results, and systems for generating pedigree data from genetic data. 62/362,660 and 62/404,912 further disclose additional supporting data for the methods. The data reported in this paper are tabulated in tables S1 to S14 in Excel format, and variant sites and frequencies with basic annotations are hosted in the following database and webserver: www.discovehrshare.com. Additional information for reproducing the results described in the article is available upon reasonable request and subject to a data use agreement.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/354/6319/aaf6814/suppl/DC1
Figs. S1 to S12
Tables S1 to S14

14 March 2016; accepted 16 November 2016
10.1126/science.aaf6814

Science
AAAS

Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study

Frederick E. Dewey, Michael F. Murray, John D. Overton, Lukas Habegger, Joseph B. Leader, Samantha N. Fetterolf, Colm O'Dushlaine, Cristopher V. Van Hout, Jeffrey Staples, Claudia Gonzaga-Jauregui, Raghu Metpally, Sarah A. Pendergrass, Monica A. Giovanni, H. Lester Kirchner, Suganthi Balasubramanian, Noura S. Abul-Husn, Dustin N. Hartzel, Daniel R. Lavage, Korey A. Kost, Jonathan S. Packer, Alexander E. Lopez, John Penn, Semanti Mukherjee, Nehal Gosalia, Manoj Kanagaraj, Alexander H. Li, Lyndon J. Mitnaul, Lance J. Adams, Thomas N. Person, Kavita Praveen, Anthony Marcketta, Matthew S. Lebo, Christina A. Austin-Tse, Heather M. Mason-Suares, Shannon Bruse, Scott Mellis, Robert Phillips, Neil Stahl, Andrew Murphy, Aris Economides, Kimberly A. Skelding, Christopher D. Still, James R. Elmore, Ingrid B. Borecki, George D. Yancopoulos, F. Daniel Davis, William A. Faucett, Omri Gottesman, Marylyn D. Ritchie, Alan R. Shuldiner, Jeffrey G. Reid, David H. Ledbetter, Aris Baras and David J. Carey (December 22, 2016)
Science **354** (6319), . [doi: 10.1126/science.aaf6814]

Editor's Summary

Unleashing the power of precision medicine

Precision medicine promises the ability to identify risks and treat patients on the basis of pathogenic genetic variation. Two studies combined exome sequencing results for over 50,000 people with their electronic health records. Dewey *et al.* found that ~3.5% of individuals in their cohort had clinically actionable genetic variants. Many of these variants affected blood lipid levels that could influence cardiovascular health. Abul-Husn *et al.* extended these findings to investigate the genetics and treatment of familial hypercholesterolemia, a risk factor for cardiovascular disease, within their patient pool. Genetic screening helped identify at-risk patients who could benefit from increased treatment.

Science, this issue p. 10.1126/science.aaf6814, p. 10.1126/science.aaf7000

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.