nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-23556-4 OPEN

Determinants of penetrance and variable expressivity in monogenic metabolic conditions across 77,184 exomes

Hundreds of thousands of genetic variants have been reported to cause severe monogenic diseases, but the probability that a variant carrier develops the disease (termed penetrance) is unknown for virtually all of them. Additionally, the clinical utility of common polygenetic variation remains uncertain. Using exome sequencing from 77,184 adult individuals (38,618 multi-ancestral individuals from a type 2 diabetes case-control study and 38,566 participants from the UK Biobank, for whom genotype array data were also available), we apply clinical standard-of-care gene variant curation for eight monogenic metabolic conditions. Rare variants causing monogenic diabetes and dyslipidemias display effect sizes significantly larger than the top 1% of the corresponding polygenic scores. Nevertheless, penetrance estimates for monogenic variant carriers average 60% or lower for most conditions. We assess epidemiologic and genetic factors contributing to risk prediction in monogenic variant carriers, demonstrating that inclusion of polygenic variation significantly improves biomarker estimation for two monogenic dyslipidemias.

A list of authors and their affiliations appears at the end of the paper.

Healthcare providers and researchers are increasingly faced with interpreting genetic sequence data collected from individuals who are asymptomatic or for whom limited clinical information is available. Standard clinical practice for reporting whole exome and genome sequencing results may involve risk assessment for genetic variation causing conditions of known relevance to the individual and also potentially impactful variants unrelated to the primary indication for testing (termed "secondary genetic findings," for example the American College of Medical Genetics and Genomics (ACMG) list of 59 medically actionable genes)[1–3]. Thus, predicting the risk conferred by genetic findings in individuals who are not known to have the relevant conditions is of critical importance, but remains a challenge[4]. Furthermore, the scope of genetic variation interpreted in current clinical genetics practice is predominantly limited to rare monogenic "Mendelian" disease variants with large predicted effect sizes, leaving the vast majority of the genome, including common variants, unassessed. Recent studies have suggested that a high burden of common genetic variation may confer increased disease risk equivalent in magnitude to carrying rare monogenic variants[5]; however, this equivalency has also been called into question[6], and it remains uncertain whether and how to integrate polygenic scores capturing common genetic variation into medical care[7].

Clinical application of genomic sequence data requires identification of medically significant genetic variants and estimation of their impact. In recent years, detailed guidelines from the ACMG and the Association for Molecular Pathology (AMP)[8] have provided standards for reporting clinically significant variants, which have been implemented by ~95% of clinical laboratories internationally[9]. Nevertheless, the probability that carriers of such variants will manifest the given condition (termed "penetrance") is unknown or uncertain for the vast majority of reported pathogenic variants[4]. Furthermore, individuals with the same genotype may exhibit variable degrees of phenotype expression (termed "variable expressivity")[10,11]. Estimates of penetrance and expressivity traditionally have been derived from studies focusing on individuals with a given condition and their family members; this approach suffers from ascertainment bias, since the proband, who came to clinical attention due to having the condition, may share other genetic and/or environmental factors influencing manifestation of the condition with their family members[11,12]. Interpretation of rare variants identified by sequencing is further complicated by limited or no data available from any source, including families, to assess penetrance[4].

Large-scale population-based and cohort studies with both sequence and phenotype data offer an opportunity to estimate penetrance and expressivity with less upward bias compared to family or case-control studies. In fact, population-based studies may have a healthy-participant bias, which could provide downwardly biased estimates of penetrance[13]. Recent studies attempting to connect large-scale genetic and phenotypic data have noted reduced penetrance estimates compared to those previously reported; however, these recent studies were limited by sample size and/or application of less stringent curation of genetic variants than the current clinical standard of care ACMG/AMP guideline approach[6,13–16]. In addition, further characterization of additional epidemiologic and genetic factors, such as phenotypic ascertainment and polygenic risk, is needed for accurate prediction of penetrance and expressivity for rare monogenic variants.

Here we present analyses performed in two separate datasets: 38,618 exomes from individuals ascertained as part of multi-ancestral type 2 diabetes (T2D) case-control studies, and 38,566 exomes from individual volunteers in the UK Biobank (UKB). Our analyses focus on traits with complex genetic architectures, involving rare and common genetic contribution, and well-defined biomarkers. These include diabetes (maturity-onset diabetes of the young (MODY), neonatal diabetes, autosomal dominant lipodystrophy) and disorders of LDL cholesterol, HDL cholesterol, triglycerides, and obesity. In addition to performing stringent curation using the ACMG/AMP criteria[8] to generate a set of clinically significant genetic variants, we also calculate polygenic scores in the UKB dataset to assess the cumulative impact of common variation on the same phenotypes. These data allow us to make a direct comparison between monogenic and polygenic risk, and to assess the contribution of polygenic risk to expressivity for carriers of monogenic variants.

Results

Identification of high confidence clinically significant variants enhances risk stratification. We studied two distinct datasets for which both individual-level exome sequence and phenotypic data were available ($N = 77{,}184$): a compilation of multi-ancestral case-control studies for T2D, involving 22,875 T2D (or prediabetes) cases (see "Methods") and 15,743 controls from the T2D-GENES and AMP-T2D consortia[17], (referred to subsequently as AMP-T2D-GENES); and 38,566 unrelated individuals of European origin from the UKB[18] (see "Methods", Supplementary Table 1, Supplementary Data 1). Our analyses focused on 26 genes offered by clinical laboratories in the United States for evaluation of monogenic forms of diabetes or diabetes-related traits through autosomal dominant modes of inheritance: MODY most commonly offered in panel testing (*GCK, HNF1A, HNF1B, HNF4A, PDX1*), an extended set of purported MODY genes less frequently offered in panel testing (*AKT2, KLF11, APPL1, ABCC8, KCNJ11, NEUROD1, CEL, INS*), neonatal diabetes (*ABCC8, GATA4, GATA6, HNF1B, INS, KCNJ11*), lipodystrophy (*AKT2, LMNA, PLIN1, PPARG*), elevated LDL cholesterol (*LDLR, APOB*), low serum LDL cholesterol (*APOB, PCSK9*), elevated serum HDL cholesterol (*CETP*), hypertriglyceridemia (*APOA5, LPL*), and monogenic obesity (*MC4R*).

We performed stringent variant curation using the clinical gold standard ACMG/AMP criteria, blinded to carrier phenotypic data for two classes of variants: 276 variants previously reported to be clinically significant (designated "pathogenic" or "likely pathogenic") in the ClinVar database[19] or designated as disease-causing in review articles[20–22]; and 218 predicted loss of function (pLoF) variants in genes with supported loss-of-function mechanism of action, which underwent curation including manual inspection of sequence reads by two independent reviewers (see "Methods"). Our approach was intended to capture high-confidence clinically significant variants, although notably excluded missense variants beyond those in the ClinVar database because of the low prior probability of disease relevance and the challenges of inferring pathogenicity for this variant class. In total across the AMP-T2D-GENES and UKB study exomes, 238 variants, representing 51% of all 463 variants curated, were determined by ACMG/AMP criteria to be clinically significant and were found in 626 carriers (Fig. 1, Supplementary Table 2, Supplementary Data 2). Across the conditions, the clinically significant variants were observed in all represented ancestral groups (Supplementary Fig. 1).

We next assessed the impact of clinically significant monogenic variants on corresponding biomarkers, restricting analyses to conditions with at least ten carriers of variants in relevant genes (Supplementary Table 2). Monogenic variant carriers for dyslipidemias had significantly more extreme lipid trait values compared to non-carriers, with shifts on average of ~55 mg/dL for both high and low LDL cholesterol conditions, ~130 mg/dL for high triglycerides, and ~16 md/dL for high HDL cholesterol ($P < 10^{-5}$ for all; adjusted for age, sex, and 10 PCs; Table 1). For monogenic obesity, *MC4R* variant carriers had ~2 kg/m^2 higher mean body mass index (BMI) than non-carriers in both datasets, however, this difference reached significance only in UKB

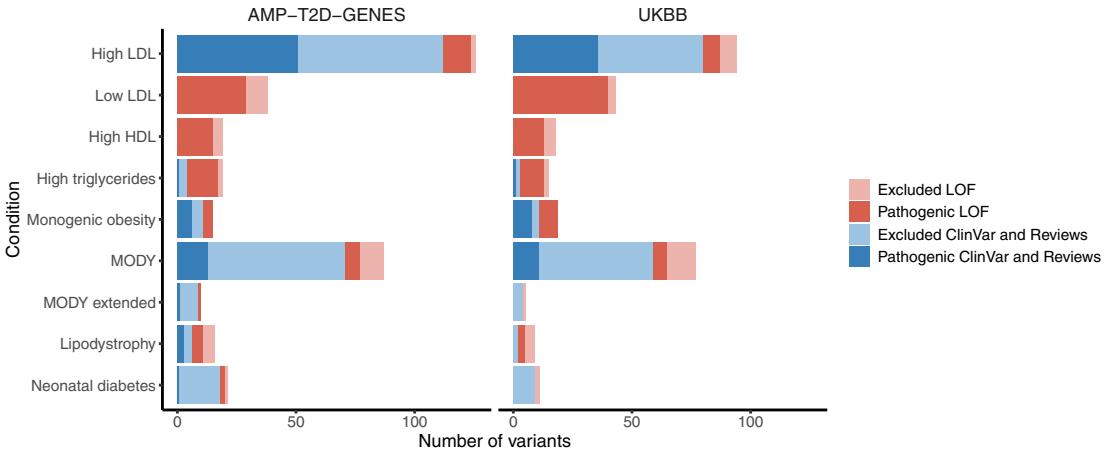

Fig. 1 Curation of ClinVar and pLoF variants across the monogenic conditions. Total number of curated ClinVar/Review (blue) and pLoF (red) variants with carriers in AMP-T2D-GENES (left panel) and UKB (right panel). Darker color shades indicate variants determined to be clinically significant (pathogenic, likely pathogenic, or pLoF) and lighter shades indicate variants excluded during curation from further analysis.

Table 1 Impact of clinically significant variants on traits.

		AMP-T2D-GENES ($N = 38,618$)			UK Biobank ($N = 38,566$)		
Condition (proxy measure)	**Gene**	***N* carrier**	**Beta (se)**	***P* value***	***N* carrier**	**Beta (se)**	***P* value***
High LDL (LDL mg/dL)	composite	55	56.0 (5.2)	3.9×10^{-24}	83	54.2 (3.9)	1.6×10^{-44}
	APOB	11	31.5 (12.0)	8.9×10^{-3}	26	52.2 (6.8)	2.2×10^{-14}
	LDLR	44	65.3 (6.3)	9.0×10^{-25}	57	55.1 (4.7)	1.1×10^{-31}
Low LDL (LDL mg/dL)	composite	35	−56.1 (7.1)	4.4×10^{-15}	90	−56.4 (3.7)	6.9×10^{-52}
	APOB	8	−79.8 (14.7)	5.9×10^{-8}	48	−74.5 (5.1)	6.7×10^{-48}
	PCSK9	27	−48.7 (8.2)	2.6×10^{-9}	42	−36.1 (5.4)	2.7×10^{-11}
High HDL (HDL mg/dL)	*CETP*	21	16.5 (3.0)	3.6×10^{-8}	20	16.8 (2.4)	2.3×10^{-12}
High triglycerides (TG mg/dL)	composite	20	130.0 (27.3)	2.8×10^{-6}	54	126.0 (12.2)	2.4×10^{-16}
	APOA5	15	122.4 (29.7)	2.6×10^{-5}	38	145.5 (13.6)	2.4×10^{-14}
	LPL	5	152.8 (54.6)	2.5×10^{-2}	16	79.3 (22.4)	9.4×10^{-4}
Monogenic obesity (BMI kg/m^2)	MC4R	28	1.5 (1.0)	6.3×10^{-2}	31	2.2 (0.8)	6.3×10^{-3}

Condition	**Gene**	***N* carrier**	**OR**	***P* value***	***N* carrier**	**OR**	***P* value***
MODY (diabetes)	composite	22	7.8 (4.2–14.6)	6.5×10^{-5}	16	21 (12.5–35.2)	2.6×10^{-8}
	GCK	7	37.4 (6.3–222.0)	1.3×10^{-3}	10	40.5 (20.3–80.7)	3.1×10^{-8}
	HNF1A	11	4.8 (2.2–10.4)	1.7×10^{-2}	5	9.0 (3.51–22.9)	2.3×10^{-2}
MODY (T2D and prediabetes)	composite	22	4.8 (2.6–8.8)	2.5×10^{-3}	16	21.5 (11.5–40.4)	9.1×10^{-9}
	GCK	7	17.8 (3.4–94.0)	8.2×10^{-3}	10	132.0 (28.7–611.0)	1.4×10^{-9}
	HNF1A	11	3.1 (1.5–6.6)	8.9×10^{-2}	5	5.1 (2.0–12.9)	6.1×10^{-2}

Composite = individuals carrying variants in any of the genes analyzed for that condition. Note that MODY composite gene set included *GCK*, *HNF1A*, *HNF1B*, *HNF4A*, and *PDX1*.
*Comparison of variant carriers to non-carries using EPACTS burden two-sided testing, adjusted for age, sex, 10 PCs. No adjustment has been made for multiple comparisons.

($P = 0.063$ AMP-T2D-GENES, $P = 0.006$ UKB). Despite differences in the study populations and designs in AMP-T2D-GENES and UKB, the effect sizes of clinically significant variants on relevant biomarkers were remarkably consistent across the two studies for dyslipidemia and obesity gene sets, once the former was adjusted for lipid medication use (Table 1, Supplementary Data 3). MODY variant carriers had significantly increased odds of having diabetes compared to non-carriers in both studies (OR > 7, $P < 10^{-4}$; Table 1, Supplementary Data 3); differences in risk estimates between the two studies were likely influenced by ascertainment practices in AMP-T2D-GENES, as it was a T2D case-control study and several sub-studies intentionally excluded diabetes cases with clinical features suggestive of MODY[17] (Supplementary Data 1).

We also performed the same effect size estimates noted above, but for the variants filtered out during our curation process. We reclassified 7% (21/276) of curated variants from review articles and from ClinVar (which had been designated as clinically significant by at least one submitting source) to "benign" or "likely benign." Likewise, 27% (59/218) of the pLoF variants were downgraded by our manual review of sequence reads. Together, these ClinVar, review, and pLoF variants that were downgraded by our curation (77/463, 17%) had markedly reduced effect sizes compared to variants we curated as clinically significant (Supplementary Data 4)[23–26]. These findings support our curation process and highlight the need for caution in relying on available variant designations without additional review.

Monogenic variant effect sizes are significantly larger than the top 1% of polygenic risk scores. We next directly compared the effect of monogenic variation to common genetic variation on the same corresponding biomarkers in UKB participants. We employed published polygenic scores capturing millions of

common genetic variants across the genome, termed global extended polygenic scores (gePS)[27] (see "Methods"). Since the gePS predicts lifetime risk of developing a disease, and the population mean age in UKB was 58 years, it was possible that estimates by gePS would be under-estimates not capturing individuals who would later in life develop a given condition. We therefore performed gePS analyses restricted to individuals age ≥ 60 year (mean age 65 years) so as to have a fairer comparison with monogenic conditions, which are typically diagnosed at a younger age.

Individuals with the top 1% of gePS had more extreme lipid levels or diabetes risk compared to those with average gePS (25–75% tiles) (Supplementary Table 3); however, the carriers of clinically significant monogenic variants for these same conditions had even more severe values compared to those top 1% respective gePS's ($P < 0.05$ for each condition, Fig. 2, Supplementary Table 3). For obesity, the difference in BMI between *MC4R* monogenic variant carriers and the top 1% BMI gePS was not significant (Fig. 2).

Monogenic metabolic conditions display highly variable penetrance estimates. While in aggregate clinically significant monogenic variants had marked effect sizes, individual-level trait values in carriers varied considerably (Fig. 3A). In both datasets, penetrance estimates based on standard disease cut-offs (see "Methods") were estimated to be 60% or lower in both studies for all monogenic metabolic conditions except *APOB* low HDL cholesterol and monogenic diabetes (Fig. 3B, Supplementary Data 1). Penetrance estimates for continuous traits will depend on the chosen threshold level, and it is notable that there was greater variability between studies than was seen with the analysis of effect sizes. Nevertheless, we clearly saw evidence of incomplete penetrance for all gene-conditions with the only exception of *GCK*-MODY; in both datasets 100% (17/17) of carriers of clinically significant *GCK* variants developed diabetes or prediabetes (penetrance estimates of 100%, 95% CI: 59.0–100% in AMP-T2D-GENES and 69.2–100% in UKB) (Fig. 3B, Supplementary Data 3, 5).

Genetic vs phenotypic ascertainment of MODY suggests broad phenotypic spectrum. We performed deeper phenotyping of MODY variant carriers in the two datasets to determine whether these genetically ascertained individuals manifested clinical features suggestive of MODY, as typically seen in phenotypically ascertained MODY cases. Monogenic diabetes, and particularly MODY (the most common form) can often be misdiagnosed as type 1 or type 2; however, MODY has subtle phenotypic differences from these other forms of diabetes and also, importantly, distinct gene-specific therapeutic strategies[28].

Focusing on the MODY genes most commonly offered in commercial panels available in the United States (*HNF1A*, *GCK*, *HNF4A*, *HNF1B*, and *PDX1*)[29], 86.4%, 95% CI 65.1–97.1%, of carriers of clinically significant variants had evidence of prediabetes or diabetes in AMP-T2D-GENES and 81.2%, 95% CI 54.4–96.0%, in UKB (Supplementary Data 5, Supplementary Fig. 2). *GCK*-MODY is characterized by non-progressive asymptomatic mild hyperglycemia that is present from birth and may remain in the prediabetes state rather than progress to diabetes.[30] As noted, there was 100% penetrance for carriers of clinically significant *GCK* variants developing diabetes or prediabetes; in addition, all those with glycated hemoglobin (HbA1c) values available ($N = 13$) had levels consistent with *GCK*-MODY, ranging from 5.7 to 7.2% (HbA1c in *GCK*-MODY is typically 5.6–7.6%[31]) (Supplementary Data 5). Penetrance estimates for diabetes in *HNF1A*-MODY from our two datasets (81% in AMP-T2D-GENES, 95% CI 48.2–97.7% and 40% in UKB, 95% CI 5.27–85.3% diagnosed with diabetes by 56 years) were lower than what has previously been reported in the literature (e.g., 97%, 95% CI 96–98% by 50 years[32]) (Supplementary Data 5).

Clinical features classically associated with MODY (BMI ≤ 30 and triglycerides ≤150[33,34]) were only observed in 50% (11/22) of MODY variant-carrying individuals in AMP-T2D-GENES and 75% (12/16) in UKB. Similarly, an expected young age of diagnosis (age ≤ 35 years), was only observed in 20% (3/15) of those with available data across both datasets (Supplementary Data 5). Thus, at least 63% of all MODY variant carriers did not have expected clinical features. Since participants in AMP-T2D-GENES were selected to be T2D cases or controls, and specific exclusion criteria were employed by several studies to remove possible monogenic diabetes cases (Supplementary Data 1)[17], these ascertainment practices could have introduced bias away from classical MODY features in MODY variant carriers. Nevertheless, when all MODY carriers were compared to others with diabetes in each study, they had significantly lower mean BMI and serum triglycerides (BMI: AMP-T2D-GENES: 26.6 vs 28.7 kg/m^2, $P = 0.027$; UKB: 25.8 vs 31.7 kg/m^2, $P = 0.004$; triglycerides: AMP-T2D-GENES: 136 vs 182 mg/dL, $P = 0.032$; UKB: 97 vs 186 mg/dL, $P = 0.004$; adjusted for age, sex, and 10 PCs). Thus, in aggregate, MODY variant carriers displayed expected clinical features, but on an individual level, genetically ascertained individuals revealed a broader spectrum of disease phenotype.

Phenotypic ascertainment strongly impacts estimates of expressivity. It is well-appreciated that phenotypic ascertainment of individuals can upwardly bias estimates of expressivity[13,35], and we sought to better define this impact by studying conditions of high and low LDL cholesterol levels, where we had information on phenotypic ascertainment within a specific AMP-T2D-GENES cohort. A set of 535 individuals selected for extreme LDL cholesterol (>98th or <2nd percentile), without knowledge of their monogenic condition carrier status, were sequenced as part of the Exome Sequencing Project (ESP) cohort in AMP-T2D-GENES[36] and not included in the prior analyses. Within this ascertained sample, we identified 18 carriers of clinically significant monogenic high LDL cholesterol variants in *APOB* and *LDLR* (mean LDL 329 mg/dL) and 15 carriers in low LDL cholesterol variants in *APOB* and *PCSK9* (mean LDL 49.2 mg/dL). As expected, compared to carriers of variants for the same LDL cholesterol conditions, but not ascertained on LDL phenotype, the two ascertained groups had more extreme LDL cholesterol levels (mean LDL cholesterol values 198 mg/dL, $P = 4 \times 10^{-4}$ and 77 mg/dL, $P = 0.06$, respectively, Fig. 4, Supplementary Table 4).

Five variants (High LDL: *LDLR* p.Glu101Lys, *LDLR* p.Asp266Glu, *LDLR* p.Gly592Glu, *APOB* p.Arg3527Gln; Low LDL: *PCSK9* p.Tyr142Ter) were carried by individuals both in the phenotypically ascertained group and in the rest of the AMP-T2D-GENES cohort. These variants showed the same pattern of significantly more extreme LDL cholesterol values in the phenotypically ascertained compared to genetically ascertained individuals ($P < 0.05$; all analyses adjusted for age, sex, ancestry, and diabetes status; Fig. 4; Supplementary Table 4). These marked differences in LDL cholesterol values between the phenotypic vs genetic ascertained carriers, even among those carrying exactly the same LDL cholesterol variant, could not be explained by the use of lipid-lowering medication, assay use, or biased selection of the LDL cholesterol values among those available (e.g., selection of maximum LDL cholesterol value ever for phenotypically ascertained participants)[36].

In fact, the mean absolute impact of phenotype ascertainment on serum LDL cholesterol levels among individuals with monogenic LDL-raising or lowering variants (27.8–131.0 mg/dL, Supplementary Table 4, Fig. 4) was thus similar or greater than the mean impact of carrying these same variants compared to non-carriers (31.5–65.3 mg/dL, Table 1, Fig. 4). Such a substantial

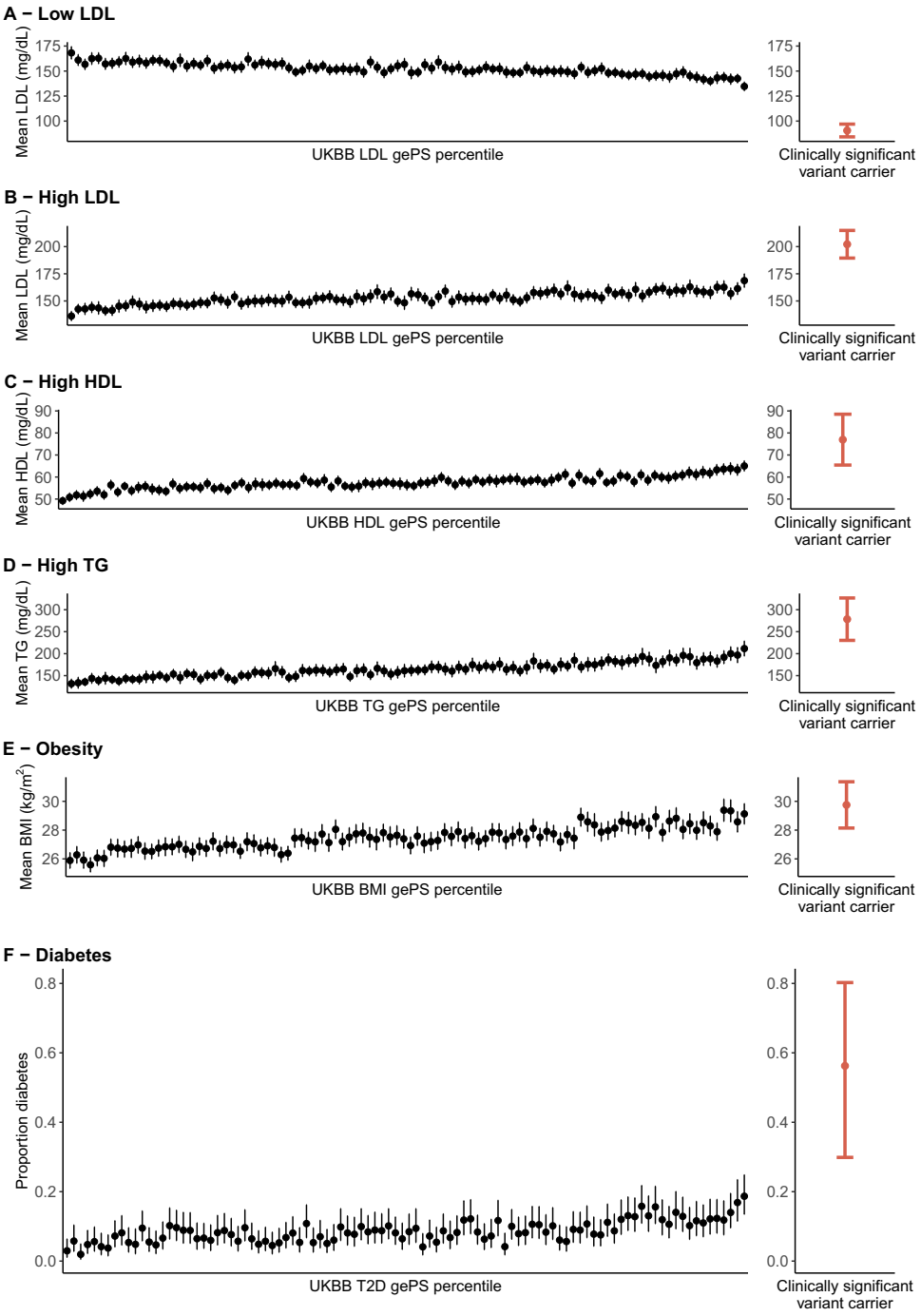

Fig. 2 Carriers of rare clinically significant monogenic variants for lipid conditions and monogenic diabetes have more extreme effect size estimates than individuals with the top 1% of global extended polygenic scores (gePS). In all plots data is from the UK Biobank participants. The left panels show the distribution of the phenotype in each percentile of the gePS for the relevant condition (black, *N* mean 364 individuals per centile), and the right panel shows the phenotype distribution in carriers of rare clinically significant monogenic variants for the corresponding condition (red); low LDL cholesterol (*APOB, PCSK9; N* = 90), high LDL cholesterol (*LDLR, APOB; N* = 83), high HDL cholesterol (*CETP; N* = 20), high triglycerides (*APOA5, LPL; N* = 54), monogenic obesity (*MC4R; N* = 31), and MODY (*GCK, HNF1A, PDX1; N* = 16). **A**–**E** Mean and 95% CI of each phenotype are indicated by the point and error bars, respectively. The same gePS calculated for risk of increasing LDL levels was used for (**A** and **B**); however, the inverse of this gePS was used for (**B**) to illustrate that higher gePS indicates risk of lower LDL cholesterol. **F** The proportion of individuals with diabetes and 95% CI computed with the Clopper–Pearson method are shown as points and error bars, respectively. Individuals in the gePS analysis were restricted to those age ≥ 60 years. LDL cholesterol and triglyceride values were adjusted for lipid-lowering medication use (see "Methods").

effect from phenotypic ascertainment reflects the large variation in expressivity at the single-variant level and underscores the importance of considering phenotypic ascertainment bias in monogenic risk prediction.

Polygenic risk may increase expressivity of monogenic variants. The variability in phenotype expressivity that we observed across all monogenic conditions (Fig. 3A) suggests that additional environmental and/or genetic factors contribute to expressivity

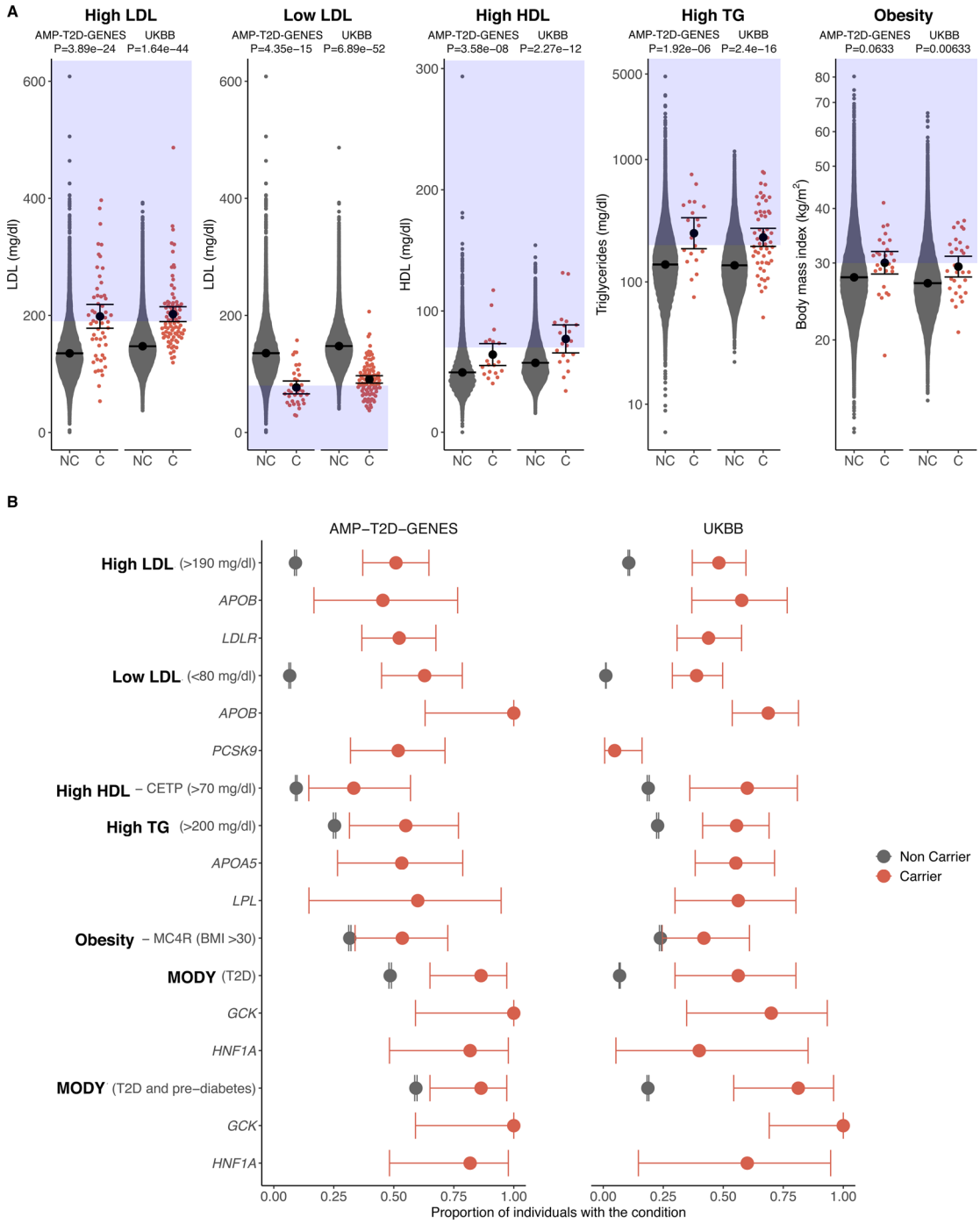

Fig. 3 Phenotype distributions and penetrance estimates of clinically significant variant carriers. In all plots, clinically significant variant carriers are shown in red and non-carriers are shown in grey. The left panel of each plot shows AMP-T2D-GENES participants (T2D case/control study) and the right panel shows UK Biobank participants (population-based study). See Supplementary Data 3 for individual counts. **A** Mean and 95% CI are represented by the black circle and black lines, respectively. Relevant lipid levels (mg/dl) or body mass index (kg/m^2) are shown for carriers (C) and non-carriers (NC) of clinically significant variants for the five monogenic conditions. The blue boxes indicate the phenotype values that meet a clinical threshold for diagnosis of each of the conditions, and *P* values were obtained by two-tailed burden analysis in EPACTS (see "Methods"). No adjustment has been made for multiple testing. **B** Dots are the proportion of individuals that have the condition based on the clinical diagnosis threshold for each condition; for MODY, we show the proportion of individuals meeting T2D as well as T2D and prediabetes criteria (see "Methods"). Error bars reflect 95% CI computed with the Clopper–Pearson method.

beyond the given monogenic variant. We assessed whether common genetic variation alters expressivity in UKB participants carrying monogenic disease variants.

Among carriers of high HDL cholesterol, low LDL cholesterol, high triglycerides, and monogenic obesity variants, we found that a higher gePS for each condition was associated with a more severe phenotype (e.g., among carriers of monogenic high HDL cholesterol variants, having an increased HDL gePS was associated with even higher HDL cholesterol). However, these trends were only significant for high HDL cholesterol (gePS one

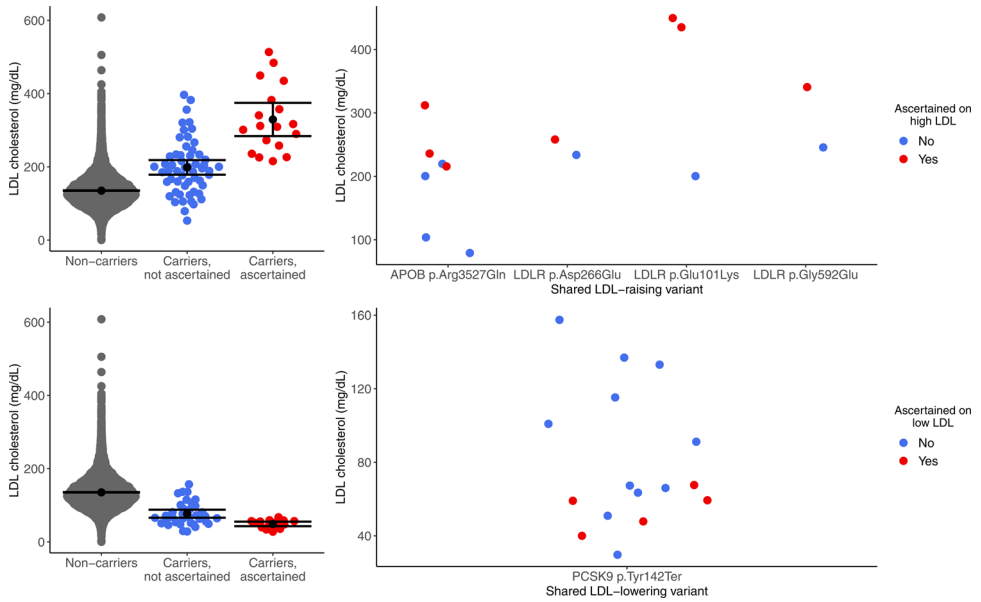

Fig. 4 Ascertainment bias significantly impacts expressivity of clinically significant variants for LDL cholesterol conditions. LDL cholesterol levels are shown for carriers and non-carriers of LDL cholesterol raising (top panels) or lowering (bottom panels) clinically significant variants in AMP-T2D-GENES. The variants carriers are stratified by whether they were identified in individuals phenotypically ascertained for extreme serum LDL cholesterol levels (Yes, Red) or in a separate unascertained population (No, Blue) (see "Methods"). The left panels show all clinically significant variant carriers. The right panels show carriers of the single variants that were present in both ascertained and unascertained individuals. Top left, LDL-raising variant Non-carriers $N =$ 19,131, Carriers not ascertained on LDL cholesterol level $N = 55$, Carriers ascertained on LDL cholesterol level $N = 18$. Bottom left, LDL-lowering variant Non-carriers $N = 19{,}151$, Carriers not ascertained $N = 35$, Carriers ascertained $N = 15$. Mean and 95% CI are represented by the black circle and black lines, respectively. LDL cholesterol values are adjusted for lipid-lowering medication use as per methods. See also Supplementary Table 4.

SD: beta 17.52 mg/dL, $P = 0.012$) and high triglycerides (gePS one SD: beta 80.57 mg/dL, $P = 0.014$) (Fig. 5, Supplementary Table 5). Notably, despite our large study size, power in this analysis was limited, and we estimate that at least 98 carriers of clinically significant variants for a given monogenic condition would be needed for 80% power to detect a correlation of 0.25 (the minimum noted for the above traits) between a given trait and gePS at significance level $\alpha = 0.05$. Therefore, for a monogenic condition with prevalence of 1 in 10,000 individuals, a population-based study with sample size on the order of one million individuals would be required to categorically determine the impact of polygenic risk.

We also assessed the interaction between gePS and monogenic risk in both monogenic carriers and non-carriers in the UKB, and observed significant positive interactions for the same two conditions, high HDL cholesterol ($P = 0.001$) and high triglyceride levels ($P = 0.01$); however, given the complexities of interaction analyses, additional work will also be needed in larger cohorts before we can conclude that gePS contributes to phenotype expression differently in carriers and non-carriers[37].

Discussion

Until recently, the impact of clinically significant monogenic variants on predicting phenotype expression has been predominantly studied in individuals or families ascertained on phenotype[12]. Our analysis employed population-based studies to provide less upwardly biased estimates of penetrance and expressivity, and to quantify the impact of phenotypic ascertainment and polygenic risk. We were able to directly compare monogenic and polygenic risk for each condition, and also assess the additional contribution of polygenic risk to expressivity for carriers of monogenic variants.

We applied the current gold standard ACMG/AMP clinical variant classification criteria[8] to ensure relevance to current clinical practice and demonstrated resultant improvement in risk stratification (Supplementary Data 3, 4). Gene variant curation was blinded to participant phenotypes and assessed variants expected to cause multiple metabolic conditions in 77,184 exomes of adults (age ≥ 40 years) from the AMP-T2D-GENES consortium and the UK Biobank. Our current analysis adds to a growing set of studies aimed at re-evaluating penetrance estimates using population-based studies[6,8,13–16,38], with our study notable for its large sample size, use of clinical standard-of-care ACMG/AMP criteria to curate genetic variants, and investigation of multiple monogenic metabolic conditions.

Carriers of the highly curated clinically significant variants for monogenic dyslipidemias and MODY had significantly more extreme trait effect sizes compared to non-carriers (betas 16.5–130.0 mg/dL for dyslipidemias, OR > 7 for diabetes risk, $P < 10^{-4}$, Table 1). Despite differences in study populations and designs, the effect estimates for rare monogenic variation for all conditions aside from monogenic diabetes (which was subject to ascertainment bias in AMP-T2D-GENES) were remarkably consistent between the two studies, supporting the integrity of our variant curation. We also assessed the impact of common genetic variation with polygenic scores. There has recently been a great deal of interest around the potential clinical contribution of such scores, especially gePS, and particularly in comparison to monogenic variant risk[5]. We show here that with the exception of monogenic obesity, polygenic risk at the top 1% of the risk distribution is not equivalent to monogenic risk, consistent with recent observations,[6] but in contrast with others[5]. In their current state and for the conditions we studied, the risk conferred by polygenic scores on their own was still substantially less than clinically significant monogenic variants; the only exception to this was *MC4R* obesity variants, which are known to have low predictive value for obesity risk[39]. There will likely be further development of polygenic scores with improved disease prediction in the coming years and with improved capture of SNP-

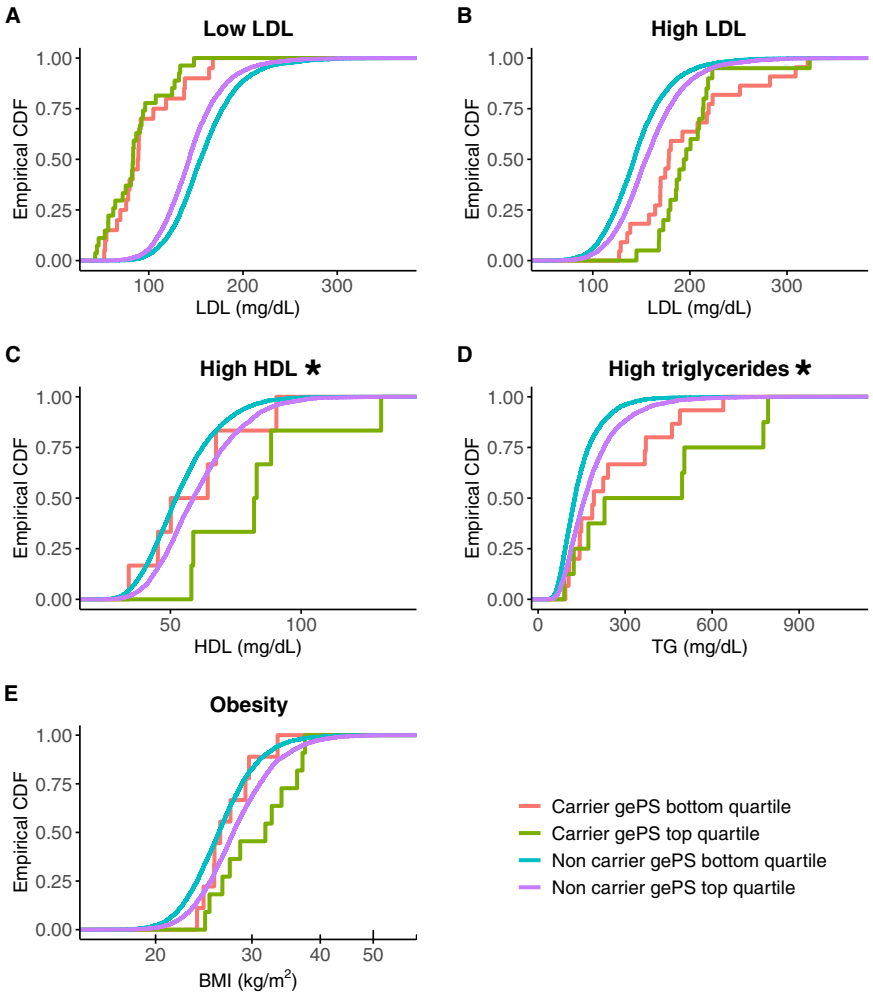

Fig. 5 The combination of clinically significant monogenic variants and corresponding polygenic scores significantly improves prediction for high HDL cholesterol and high triglyceride conditions. In all plots, an empirical cumulative distribution function (CDF) of each phenotype is shown for clinically significant variant carriers and non-carriers in the UKB for each monogenic condition stratified by bottom/top quartiles of the corresponding gePS. The monogenic conditions are (**A**) low LDL cholesterol (*APOB, PCSK9*), (**B**) high LDL cholesterol (*LDLR, APOB*), (**C**) high HDL cholesterol (*CETP*), (**D**) high triglycerides (*APOA5, LPL*), and (**E**) monogenic obesity (*MC4R*). The same gePS calculated for risk of increasing LDL cholesterol levels was used for (**A** and **B**), however, the inverse of the gePS was used for (**A**) to illustrate that higher gePS indicates risk of lower LDL cholesterol. The impact of higher gePS was testing in carrier-only linear regression analysis; asterisks indicate two-sided $P < 0.05$ unadjusted for multiple testing (High HDL $P = 0.012$, High Triglycerides $P = 0.014$). See also Supplementary Table 5.

based heritability, it may well be possible to identify monogenic risk equivalents. For example, we estimate that using a polygenic score for HDL cholesterol capturing 15.7% heritability (a maximum SNP-based heritability predicted by various analyses[40–42]), 0.13% of the population with the highest polygenic scores would have mean HDL cholesterol values equivalent to the mean HDL cholesterol value we observed in carriers of monogenic high HDL in this study.

We observed a wide range of expressivity among clinically significant monogenic variant carriers across all traits (Fig. 2A), and consequently estimates of penetrance were 60% or lower for all conditions except *APOB* low LDL cholesterol and monogenic diabetes. Given the dependence of penetrance estimates for continuous traits on a chosen threshold and the great degree of variability between studies observed for penetrance estimates, the most important message from our findings is not an exact penetrance estimate per se, but rather the wide range of expressivity observed for carriers of highly curated monogenic variants. We observed particularly low penetrance of *MC4R* for obesity (<55% for BMI ≥ 30 kg/m^2), consistent with previous findings[6,39,43] (Fig. 2B) and particularly high penetrance for *GCK*-MODY (100% for diabetes or prediabetes in both studies, 95% CI's 59–100% in AMP-T2D-GENES, 69–100% in UKB). The range of penetrance estimates across genes and conditions may relate to ability to measure the direct biomarker(s) impacted by a given gene, the extent to which there are redundant mechanisms available in a given pathway to overcome a genetic defect[44], and the extent to which additional factors, such as other genetic and environmental factors (e.g., diet), impact the trait[11]. The finding of 100% penetrance for diabetes or prediabetes seen in the 17 carriers of *GCK*-MODY across both datasets is particularly intriguing. *GCK* encodes glucokinase, which acts as the cell's glucose sensor as it facilitates phosphorylation of glucose to glucose-6-phosphate in the pancreatic beta cell, which is the first and rate-limiting step in glucose metabolism[45]. The complete penetrance we have observed may be due to the ability to directly measure glucose as a relevant biomarker, as well as the essential role of *GCK* in glucose homeostasis, with suspected non-redundancy in functioning as a glucose sensor[45].

We also characterized the impact of phenotypic ascertainment bias on expressivity of clinically significant variants, showing that

in individuals with the same LDL cholesterol-raising or -lowering variants there were significant differences in biomarker levels depending on the mode of ascertainment (genetic vs phenotypic) (Fig. 3) and that the magnitude of this difference on LDL cholesterol levels (29–129 mg/dL) was similar or greater than the mean effect size of such variants (31.5–65.3 mg/dL, Table 1). This substantial impact of ascertainment bias was seen at the individual variant level, consistent with other similar observations of LDL cholesterol levels in *LDLR* and *APOB* carriers in a different study population[35,46] and *HNF4A* p.Arg114Trp in diabetes risk[13] (*HNF4A* p.Arg114Trp was present in the present datasets, but filtered out due to its designation as a variant of uncertain significant (VUS), reflecting its known low penetrance). The extent of ascertainment bias that we and others have identified highlights an important genetic counseling consideration, particularly with respect to interpretations of genomic sequencing data with limited clinical context available: interpretation of the same test result will likely have different prognostic implications depending on whether the individual tested or family members carry the phenotype of interest (e.g., hyperlipidemia) vs if a variant is identified secondarily; a Bayesian framework that takes into account pre-test probability might therefore be useful[47]. In addition, the variable expressivity seen at the single-variant level in multiple instances further supports additional risk factor modulation from other genetic and environmental exposures.

With regard to additional genetic factors impacting expressivity, we assessed the impact of more common polygenic variation on carriers of monogenic variants and found significant contributions for both high HDL cholesterol and high triglyceride levels ($P < 0.05$). These results add to a growing body of research supporting a significant polygenic contribution to monogenic risk across a number of conditions, including height, breast cancer, and coronary artery disease[6,38,48–50]. These studies, like ours, suggest that polygenic scores could be used clinically to improve risk estimation of monogenic disease carriers; however, power is limited in population-based studies given how rare carriers typically are, and it will be important to investigate in even larger datasets for refining risk estimates. We estimate that for a monogenic condition with prevalence of 1 in 10,000 individuals, population-based analyses well-powered to capture the contribution of polygenic risk to individuals with the monogenic condition would require on the order of one million individuals.

One limitation of this study is that our selection of variants for curation did not include all possible missense variants, but rather was confined to those reported in ClinVar or subject area reviews. This approach was designed to streamline the variant curation process and restrict our analyses to highly-confident pathogenic variants, but also meant that we were unable to generate estimates of the prevalence of monogenic condition in the two datasets. As discussed previously, there is also the potential for residual bias within the datasets. In the case of AMP-T2D-GENES, ascertainment of participants could have impacted penetrance of monogenic diabetes and expressivity of the metabolic phenotypes (Supplementary Data 1). In the UKB, a healthy-participant bias[51] would be expected to reduce estimates of penetrance. In addition, the age cut off of 40 years applied to both studies could introduce a survivor bias, such that carriers of highly penetrant variants causing lethal conditions could have died before age 40, precluding their enrollment; such a survivor bias could cause a downward bias of effect size estimates, but would be expected to impact a minority of the conditions we studied, such as high LDL cholesterol and high triglycerides (due to increased risk of early coronary artery disease). Furthermore, despite our large dataset of exomes, the likelihood of observing any specific rare pathogenic variant is still low; this raises the possibility of bias toward lower penetrance of clinically significant variants, since allele frequency is a major predictor of pathogenicity[52], and rarer variants with potentially greater penetrance are less likely to be observed. While the present study includes diverse ancestral representation for estimates of effect size for clinically significant monogenic variants, analyses involving polygenic scores were limited by availability of SNP data, and thus restricted to the available UKB exome data, of which the overwhelming majority were individuals of European ancestry. It will be important for future research to extend this work to populations of non-European ancestry. Finally, analyses to assess penetrance and expressivity were limited to single phenotypic measures, which are less ideal than multiple longitudinal measures, and while we attempted to correct for large factors impacting measures (e.g., use of lipid-lowering medication for serum lipid measures), there may have been other relevant factors that were not taken into account. Strengths of this study include the large number of participants with both phenotype and exome data, and the strict variant curation methodology applied. Our analysis of 276 variants designated by ClinVar as pathogenic or likely pathogenic highlights the need for careful curation of variants in clinical practice, with 57% reclassified to "benign," "likely benign," or "variant of uncertain significance" with application of ACMG/AMP criteria (Fig. 1). Of note, however, the ClinVar variants we curated included those submitted to the database before establishment of current standards for curation[8]. With time, we can expect that the ClinVar database will become a more reliable resource for ascertaining clinically significant variants, as more submitters utilize standardized curation practices and additionally as condition-specific standards and curation are provided by ClinGen Expert Panels, including the Monogenic Diabetes Expert Panel in which several of the co-authors participate[53].

Our study emphasizes the critical need for careful interpretation of monogenic variation, highlighting the roles of variant curation, phenotypic ascertainment, and polygenic risk in the estimates of penetrance and expressivity. In the coming years, access to larger sequencing studies will allow assessment of increasingly rare variants; however, deep phenotyping of such datasets, for example information on medication use and age of disease onset, will to be needed in parallel to better define genetic risk estimates. Improved understanding of monogenic variant expressivity will also likely require broader incorporation of genetic variation across the allelic frequency spectrum and integration of environmental factors. Such advances will facilitate modeling of disease risk and ultimately guide individualized patient genetic counseling and management recommendations.

Methods

Study populations and phenotype curation

AMP-T2D-GENES. The complete AMP-T2D-GENES cohort consists of 20,791 cases and 24,440 controls selected from multiple distinct multi-ancestry studies[17]. The present study includes a subset of 22,875 T2D or prediabetes and 15,743 controls from studies who consented for the data to be used in this analysis, which included Genetics of Type 2 Diabetes (GoT2D), the Exome Sequencing Project (ESP), Lundbeck Foundation Centre for Applied Medical Genomics in Personalised Disease Prediction, Prevention and Care (LuCamp), Slim Initiative in Genomic Medicine for the Americas (SIGMA), and T2D-GENES (Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples). General study characteristics are provided in Supplementary Table 1 with more details, including exclusion criteria available in Supplementary Data 1, which is adapted from Flannick et al., 2019[17]. All samples were approved for use by their home institution's institutional review board or ethics committee. Analysis of the data was approved by the Mass General Brigham (formerly Partners) institutional review board in Boston, Massachusetts (protocol # 2017P000445/PHS) and were limited to those participants in each cohort with available DNA who consented to genetic studies.

Phenotype information related to diabetes status was collected by each case-control or cohort study, as previously described in Flannick et al[17]. In addition, we defined prediabetes as any individual with HbA1c ≥ 5.7%, fasting blood glucose ≥ 100 mg/dL, or oral glucose tolerance test (OGTT) 2 h blood glucose ≥ 140 mg/dL. In individuals who were reported to be on lipid-lowering medication, serum LDL

cholesterol and triglyceride levels were adjusted for statin use based on previous studies estimating the impact[54,55]: we divided LDL by 0.7 and triglycerides by 0.85 as has been previously been implemented[56]. Self-reported ancestry was used, as this was previously shown to correlate well with principal component analysis (PCA) defined ancestry and specific exceptions were dropped from analyses[17]. Analyses described below used a dataset restricted to individuals in the "unrelated analysis set" (see Flannick et al.[17] methods). To provide consistency with the UKB dataset, individuals younger than age 40 were also excluded. Individuals recruited to the Pakistan Genomic Resource cohort were excluded for all analyses involving lipid levels or BMI.

UK Biobank. UK Biobank (UKB) is a prospective cohort of ~500,000 recruited individuals from the general population aged 40–69 years in 2006–2010 from across the United Kingdom, with genotype, phenotype, and linked healthcare record data[57]. All participants provided electronic informed consent at their initial visit. Analysis of the data was approved by the Mass General Brigham (formerly Partners) institutional review board in Boston, Massachusetts, and was performed under UK Biobank application 27892.

Direct LDL cholesterol (mmol/L), direct HDL cholesterol (mmol/L), triglyceride (mmol/L), BMI (kg/m2) (field codes: 30780, 30760, 30870, 21001) data were extracted for all individuals. Lipid measurements were converted from mmol/L to mg/dL. The mean for all visits was used in subsequent analyses. The "Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones" fields (6177 and 6153) was used to determine lipid-lowering medication, where an individual was considered to be on lipid-lowering medication if it was recorded at any of the visits. LDL and triglyceride values were adjusted for use of lipid-lowering medication, as described above.

Glycated hemoglobin (HbA1c; field code 30750) was taken as the maximum observed across visits. Since monogenic diabetes may be misdiagnosed as type 1 or type 2 diabetes, we used an inclusive definition of diabetes: possible and probable type 1 or type 2 diabetes was determined in a manner similar to previously described methods[58]. We also considered individuals as having diabetes if they had ICD10 codes E10-E14 (fields: 41202), and recorded diabetes medication use (fields: 6177, 6153), diabetes ever diagnosed by a doctor (field: 2443), nurse interview codes indicating diabetes (fields: 1220—any diabetes, 1222—T1D, 1223—T2D), or HbA1c ≥ 6.5%. Prediabetes was defined as any individual with HbA1c ≥5.7%. We also extracted data for the first recorded age of diabetes diagnosis (fields: 20009, 2976), age, and sex.

This dataset was filtered to only unrelated individuals with European ancestry to facilitate comparisons of biomarkers in analyses using polygenic risk scores. Filtering to unrelated individuals was done using the column "used.in.pca.calculation" in the UKB genotype data sample QC document (ukb_sqc_v2.txt) as a proxy. This column indicates samples which UKB used in a principal component analysis (PCA), and this analysis was only performed on unrelated, high quality samples. To filter to European ancestry only, samples were first projected onto 1000 Genomes phase 3[59] PCA coordinate space. Then Aberrant R package[60] clustering was used to identify individuals falling within 1000 Genomes project EUR PC1 and PC2 limits (lambda = 4.5). Individuals that self-reported as non-European ethnicity were also filtered. There were 38,566 individuals remaining after all filtering and intersection with individuals that also have exome sequence data released in the first tranche (Category 170).

Generation of gene list. We sought to include genes that would be ordered in the United States in clinical practice to diagnose conditions of monogenic diabetes, lipodystrophy, obesity, and lipid disorders. We searched the Genetic Testing Registry (https://www.ncbi.nlm.nih.gov/gtr/) and Concert Genetics (https://app.concertgenetics.com/), last accessed March 14th, 2018, for lists of available commercial gene panels for clinical genetic testing for these diseases available in the United States. We filtered this list of genes to those with an autosomal dominant mode of inheritance, as determined by the Online Mendelian Inheritance in Man® (OMIM, https://www.omim.org/). For the genes in OMIM where mode of inheritance was not specified, the genes were researched in ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) and related literature. In total there were 26 autosomal dominant genes across the conditions. We further excluded any gene where there was no ClinVar submission (April 2019 ClinVar submission summary) of pathogenic or likely pathogenic for the phenotype of interest that also included clinical testing as a collection method, leaving 20 genes. We determined phenotype overlap by manual review of the "SubmittedPhenotypeInfo" and "ReportedPhenotypeInfo" fields in the submission summary where present and "ExplanationOfInterpretation" or submitted PubMed articles when phenotype info was not reported in the other fields. For MODY, most commercial panels available in March, 2018 included *HNF1A*, *HNF4A*, *GCK*, *HNF1B*, and *PDX1*, with larger panels less widely available. We therefore separated the MODY genes into two categories: "MODY" including those five genes and "MODY extended" including eight additional genes.

Determination of genes with LoF mechanism. The pLoF curation was restricted to genes alleged to cause disease with a LoF mechanism based on reporting in ClinVar or a PubMed publication of an LoF variant in an individual with the phenotype of interest.

Two genes were determined to be related to both high LDL (familial hypercholesterolemia) and low LDL (familial hypobetalipoproteinemia): *APOB* and *PCSK9*. Gain-of-function missense mutations in both genes result in increased LDL levels, while LoF mutations cause lower LDL levels[61–63]. Therefore, only missense ClinVar variants in *APOB* and *PCSK9* were assessed in the curation process for high LDL, and LoF variants were considered for low LDL.

Exome data variant filtering and annotation. All filtering and annotation described below was performed using Hail 0.2.54 (https://hail.is).

AMP-T2D-GENES. Exome sequencing and quality control were described previously[17]. We applied additional genotype filters to retain only high-quality genotypes: genotype quality ≥ 20, depth ≥10, and minor allele balance > 0.25 for heterozygous genotypes. Variants were annotated using Ensembl's Variant Effect Predictor (VEP) v85[64] with the Loss-of-function Transcript Effect Estimator (LOFTEE) plugin[65]. The dataset was then filtered to only variants with a consequence on any of the genes of interest. The filtered VCF was used in analyses described below that involve EPACTS.

We determined which variants in our dataset have been submitted to ClinVar by cross-referencing this filtered variant list with the ClinVar VCF (April 2019) (further curation described below). A list of predicted loss-of-function (pLoF) variants, including stop gained, frameshift or essential splice site (splice donor or splice acceptor), was generated by filtering to variants with a LOFTEE high-confidence (HC) annotation on any transcript. Finally, we used transcript expression-aware annotation[66] to add pext (proportion expression across transcripts) values for the worst consequence annotation to each variant for use in pLoF curation discussed below.

UK Biobank. UKB exome sequencing PLINK files were imported into Hail and all the same annotation described for AMP-T2D-GENES was added using appropriate files for genotype reference GRCh38 and VEP v95. In order to compare UKB variants to AMP-T2D-GENES variants we used Hail's liftover method to lift data from GRCh38 to GRCh37. Since the PLINK files do not contain genotype quality information that we can use for filtering low-quality genotypes, we downloaded the gVCFs for all variant carriers and determined which individuals genotypes were not high-quality (genotype quality ≥20, depth ≥10, and minor allele balance > 0.25 for heterozygous genotypes) and set each of these to missing in the VCF. After the initial analysis was completed, UKB reported that there was an error in the SPB gVCFs that led to a systematic under-marking of duplicate reads. Therefore, all genotypes in carriers of clinically significant variants were confirmed in the corrected SPB gVCFs (field: 23176).

ClinVar variant curation. We identified individuals carrying variants in the genes of interest that had at least one "pathogenic" or "likely pathogenic" submission in ClinVar by a clinical testing lab for the relevant trait. To streamline variant curation we first generated a list of high confidence clinical genetic testing laboratories. Using the April 2019 release of the ClinVar submission summary, a lab was considered high confidence if it had submitted >15,000 variants to ClinVar and had updated its submission after 2017 when the most recent ACMG variant interpretation guidelines were published[8]. This resulted in eight labs: Invitae; GeneDx; Ambry Genetics; EGL Genetic Diagnostics; Eurofins Clinical Diagnostics; PreventionGenetics; Laboratory of Molecular Medicine, Partners Healthcare Personalized Medicine; Genetic Services Laboratory, University of Chicago; and Counsyl. Variants that were reported by any lab on this list since January 1st, 2017 were then accepted as having the pathogenicity reported by the lab.

These labs were further verified through manual curation. First, five variants from each lab that were also present in our study were chosen to be manually curated, so that the manual curation could be compared to the lab's analysis. Through this, we found no differences in curation results. Then, five variants from each lab were chosen at random through ClinVar—one Pathogenic, one Likely Pathogenic, one VUS, one Likely Benign, and one Benign. As PreventionGenetics only submitted Benign and Likely Benign to ClinVar, their variants were limited to those categories. These variants were then also manually curated, and the results were compared. The only difference in curation of the non-study variants involved University of Chicago, due to internal data initially not available to our study curator; however, the same conclusion was reached upon inclusion of this internal data, which was included in their reporting in ClinVar. During the manual phenotype curation (described below), we discovered Counsyl reported conflicting phenotypes for the same variant, so we opted to manually curate variants assessed by Counsyl.

The variants not analyzed by high confidence labs were analyzed separately using manual curation with the curator blinded to carrier phenotypes. The ClinGen Variant Curation Interface (https://curation.clinicalgenome.org/) was used to analyze the variants and assign evidence following the ACMG guidelines[8] and recommendation for interpretation of LoF variants[67], with input from gene-specific rules under development by the Monogenic Diabetes Expert Panel VCEP (https://clinicalgenome.org/affiliation/50016/) for the MODY variants. Databases and other resources such as ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), Human Gene Mutation Database (HGMD) (https://digitalinsights.qiagen.com/products-

overview/clinical-insights-portfolio/human-gene-mutation-database/), gnomAD v2.1.1 (https://gnomad.broadinstitute.org/), PubMed (https://pubmed.ncbi.nlm.nih.gov/), Google Scholar (https://scholar.google.com/), Alamut v.2.11 (https://www.interactive-biosoftware.com/alamut-visual/), and the UCSC browser (https://genome.ucsc.edu/) were utilized to collect evidence for curation purposes. The general guidelines were adjusted slightly for certain criteria such as control population frequency as shown in Supplementary Table 6. Since most AMP-T2D-GENES participants are included in gnomAD, AMP-T2D-GENES allele frequency decisions were made by subtracting the number of AMP-T2D-GENES carriers from the number of total gnomAD carriers to determine an adjusted gnomAD allele frequency, which was compared to the cut-offs shown in Supplementary Table 6.

Three variants within *HNF1A* were excluded from further analysis because of poor genotyping quality at this site making it difficult to determine which individuals are actually carriers (GRCh37: 12-121432114-CG-C, 12-121432116-G-GC, 12-121432117-G-GC, GRCh38: chr12:120994311-CG-C, chr12:120994313-G-GC, chr12:120994314-G-GC). As all three are frameshifts, these variants were also excluded from the pLoF curation described below.

Variants in MODY genes were curated by a second set of reviewers at University of Maryland School of Medicine, the home institution of the ClinGen Monogenic Diabetes Expert Panel, to ensure accuracy. All variants were consistently classified as collectively pathogenic or likely pathogenic (Supplementary Data 2).

All variants curated for this project, along with their classification and supporting evidence, were submitted to ClinVar on January 30th, 2020.

High confidence loss of function variants. As described above, we used LOFTEE[65] to generate a list of high confidence pLoF variants, restricting to the set of genes we determined to have a LoF mechanism of pathogenicity. Each pLoF variant was assessed by manual review of reads by two independent reviewers. The reads were examined for poor quality, homopolymer artifacts, and multinucleotide variants (MNVs) causing a synonymous or missense variant instead of the reported stop codon. Where available, gnomAD data was examined to identify variants that were flagged as filtered by gnomAD's random forest variant quality control method. UCSC genome browser data was assessed to determine the conservation of the region, the location of the variant, and how many transcripts the variant was coding. If the variant was present in the last exon or last 50 base pairs of the penultimate exon, it was deemed not LoF due to a predicted lack of nonsense mediated decay. However, this was overruled if the variant was predicted to delete over 25% of the gene. The potential for a splice site rescue was assessed by examining +/− 21 bp around the variant. Any inframe splice site within 6 bp was considered an essential splice site rescue and possible inframe splice site rescues between 6 and 21 bp were considered a rescue if validated by the alternative splice site prediction tool Alamut v.2.11. We also used pext values obtained from the transcript expression-aware annotation[66] to indicate variants that fell in exons that have evidence of poor expression (specific cut-offs are detailed in Supplementary Data 6). Variants were classified into 5 categories, "LoF", "likely LoF", "uncertain", "likely not LoF", or "not LoF" using the guidelines described in Supplementary Data 6. Any variant that had a discordant assessment between the two reviewers ("LoF" or "likely LoF" by only one reviewer) was examined by a third reviewer to determine the final pLoF annotation.

Carrier vs non-carrier effect size analysis. We considered an individual to be a carrier of a clinically significant variant if they carry a ClinVar variant assessed as pathogenic or likely pathogenic or a pLoF variant passing manual curation ("LoF" or "likely LoF" as described above). For AMP-T2D-GENES, as previously described[17], we accounted for the diverse ancestry and different sequencing technologies by using a modified version of EPACTS v3.2.4 (http://genome.sph.umich.edu/wiki/EPACTS) that sets specified variants to missing based on QC of sample subgroups (as described in Flannick et al.[17], there are 25 subgroups that were determined by stratifying samples by cohort of origin, ancestry, and/or sequencing technology). As covariates in AMP-T2D-GENES analyses, we included sex, age, PCs 1–10, sample subgroup, and sequencing technology all as previously defined[17]. Analyses on UKB used covariates for sex, age, PCs 1–10 and the genotyping array.

For both AMP-T2D-GENES and UKB, we used VCFs produced after filtering variants as described above and performed the group b.burdenFirth for binary traits and q.burden test for continuous traits in EPACTS to compare carriers and non-carriers for the following condition/phenotype pairs: high LDL cholesterol with LDL cholesterol (mg/dL); low LDL cholesterol with LDL cholesterol (mg/dL); high HDL cholesterol with HDL cholesterol (mg/dL); high triglycerides with triglycerides (mg/dL); monogenic obesity with BMI (kg/m^2), MODY with diabetes status, and in diabetes cases only: HDL cholesterol, Triglycerides, and BMI.

In addition, we included T2D or T2D with prediabetes as covariates in all tests on lipid measurements and BMI. Triglycerides and BMI were log transformed. All of these analyses were also performed per gene to ensure that we captured possible gene level differences in phenotype values.

Estimation of penetrance. Unlike diabetes, phenotypes used to assess the possibility that individuals have each monogenic lipid condition or obesity, are continuous. The following clinical diagnosis cut-offs were used to dichotomize the phenotypes for estimating penetrance: High LDL cholesterol: LDL cholesterol ≥190 mg/dL[68], Low LDL cholesterol (familial hypobetalipoproteinemia): LDL cholesterol ≤80 mg/dL[69], High HDL cholesterol: HDL cholesterol ≥70 mg/dL[70], High triglycerides: triglycerides ≥200 mg/dL[68], and Monogenic obesity: BMI ≥30 kg/m^2.

Penetrance estimates were calculated as the proportion of individuals carrying a clinically significant variant that also exhibit the expected condition. To determine the significance for all penetrance estimates we used the group Firth burden test in the modified version of EPACTS and the same covariates as described in "Carrier vs non-carrier enrichment analysis".

Calculation of global extended polygenic score (gePS)

Body mass index and type 2 diabetes. Global extended polygenic scores for T2D and BMI were previously calculated on UKB participants using LDpred[5,43]. The variants and weights used in the calculation were downloaded (http://www.broadcvdi.org/informational/data). These weights were then applied to the UKB genotype data from the subset of individuals included in this study to calculate a gePS using Hail's equivalent to the—score method in PLINK version 1.9 (https://hail.is/docs/0.2/guides/genetics.html?highlight=prs). These values were then scaled and centered around zero with a standard deviation of one for downstream analysis. We confirmed that plots of T2D prevalence and BMI by respective polygenic scores converged at the same upper limits as previously published[5,43].

Lipid conditions. To estimate a gePS for each lipid phenotype, we filtered UK Biobank genotype data to only the individuals used in this study (unrelated, EUR ancestry, and exome sequenced) and excluded SNPs with an imputation INFO < 0.3 and allele frequency <1%. Summary statistics for lipid GWAS were downloaded from the European Network for Genetic and Genomic Epidemiology (ENGAGE) Consortium. This included LDL cholesterol, HDL cholesterol, and triglyceride GWAS summary stats from a meta-analysis of up to 62,166 individuals of European ancestry[71]. We filtered to variants observed in HapMap3 (—only-hm3) and both the summary statistics and genotype data, and then estimated SNP weights using the Bayesian computational method LDpred (version 1.0.6) which accounts for local LD patterns[72]. SNP weight estimates were obtained using the infinitesimal (inf) model (assumes all genetic variants impact phenotype) with heritability estimates (TG: 0.1525, LDL: 0.1347, HDL: 0.1572) as previously calculated using LD Score regression[42] and displayed on LD Hub[73]. We then used PLINK version 1.9 (—score) to calculate polygenic scores using the SNP weights[74]. As in the BMI and T2D gePRS, the distribution was scaled to have a mean of zero and one standard deviation around the mean. Since there is a single gePS for LDL cholesterol, the scaled gePS was multiplied by −1 for figures and analyses comparing low LDL cholesterol carrier phenotype values to phenotypes aggregated by gePS deciles or quantiles.

Statistical analysis. We used generalized linear models (GLM) to examine the gePS results in a few different ways. We compared the top 1% to the interquartile range (25–75%) of the gePS and to the clinically significant variant carriers (Supplementary Table 3). For both analyses we restricted the age in controls to >= 60. In addition, we determine the effect size of gePS on phenotypes in the subset of only clinically significant variant carriers and assessed the interaction of carrier status and gePS (Supplementary Table 5). In all GLMs age, sex and 10 PC's were included in the model as covariates. A linear regression was performed for all phenotypes except diabetes where a logistic regression was applied.

All plots were made using R version 3.5.2.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Sequence data and phenotypes from the AMP-T2D-GENES study are available via the database of Genotypes and Phenotypes (dbGAP) and/or the European Genome-phenome Archive, as indicated in Supplementary Data 1. Access to data from the UK Biobank can be obtained at https://www.ukbiobank.ac.uk/enable-your-research. All variants curated for this project, along with their classification and supporting evidence, were submitted to the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/) on January 30th, 2020. The following databases were accessed for this work: ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), Human Gene Mutation Database (https://digitalinsights.qiagen.com/products-overview/clinical-insights-portfolio/human-gene-mutation-database/), gnomAD v2.1.1 (https://gnomad.broadinstitute.org/), PubMed (https://pubmed.ncbi.nlm.nih.gov/), Google Scholar (https://scholar.google.com/), Alamut v.2.11 (https://www.interactive-biosoftware.com/alamut-visual/), and the UCSC browser (https://genome.ucsc.edu/).

Code availability

All software used in the analysis were open source and described in the "Methods" section of the paper. Existing software packages used were: Plink 1.9, EPACTS v3.2.4, Rv3.5.2, Hail v0.2.54, Alamut v2.11, LDpred v1.0.6, Ensembl's Variant Effect Predictor

(VEP) versions 85 and 95, Aberrant v1.0 R package, and LOFTEE. Code written for analyses performed in the paper are available in GitHub: https://github.com/broadinstitute/exome_penetrance.

Received: 1 December 2020; Accepted: 27 April 2021;
Published online: 09 June 2021

References

1. Directors, A. B. O. ACMG policy statement: updated recommendations regarding analysis and reporting of secondary findings in clinical genome-scale sequencing. *Genet. Med.* **17**, 68–69 (2015).
2. Green, R. C. et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **15**, 565–574 (2013).
3. Kalia, S. S. et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249–255 (2017).
4. Directors, A. B. O. The use of ACMG secondary findings recommendations for general population screening: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **21**, 1467–1468 (2019).
5. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
6. Oetjens, M. T., Kelly, M. A., Sturm, A. C., Martin, C. L. & Ledbetter, D. H. Quantifying the polygenic contribution to variable expressivity in eleven rare genetic disorders. *Nat. Commun.* **10**, 4897 (2019).
7. Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* **12**, 44 (2020).
8. Richards, S. 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).
9. Niehaus, A. et al. A survey assessing adoption of the ACMG-AMP guidelines for interpreting sequence variants and identification of areas for continued improvement. *Genet. Med.* **21**, 1699–1701 (2019).
10. Zlotogora, J. Penetrance and expressivity in the molecular age. *Genet. Med.* **5**, 347–352 (2003).
11. Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* **132**, 1077–1130 (2013).
12. Turner, H. & Jackson, L. Evidence for penetrance in patients without a family history of disease: a systematic review. *Eur. J. Hum. Genet.* **28**, 539–550 (2020).
13. Wright, C. F. et al. Assessing the pathogenicity, penetrance, and expressivity of putative disease-causing variants in a population setting. *Am. J. Hum. Genet.* **104**, 275–286 (2019).
14. Natarajan, P. et al. Aggregate penetrance of genomic variants for actionable disorders in European and African Americans. *Sci. Transl. Med.* **8**, 364ra151 (2016).
15. Abul-Husn, N. S., et al. Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science* **354**, aaf7000 (2016).
16. Flannick, J. et al. Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. *Nat. Genet.* **45**, 1380–1385 (2013).
17. Flannick, J. et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).
18. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
19. Harrison, S. M. et al. Using ClinVar as a resource to support variant interpretation. *Curr. Protoc. Hum. Genet.* **89**, 8.16.11–18.16.23 (2016).
20. Ellard, S. & Colclough, K. Mutations in the genes encoding the transcription factors hepatocyte nuclear factor 1 alpha (HNF1A) and 4 alpha (HNF4A) in maturity-onset diabetes of the young. *Hum. Mutat.* **27**, 854–869 (2006).
21. Osbak, K. K. et al. Update on mutations in glucokinase (GCK), which cause maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemic hypoglycemia. *Hum. Mutat.* **30**, 1512–1526 (2009).
22. Colclough, K., Bellanne-Chantelot, C., Saint-Martin, C., Flanagan, S. E. & Ellard, S. Mutations in the genes encoding the transcription factors hepatocyte nuclear factor 1 alpha and 4 alpha in maturity-onset diabetes of the young and hyperinsulinemic hypoglycemia. *Hum. Mutat.* **34**, 669–685 (2013).
23. Rehm, H. L. et al. ClinGen—the clinical genome resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
24. Yang, S. et al. Sources of discordance among germ-line variant classifications in ClinVar. *Genet. Med.* **19**, 1118–1126 (2017).
25. Harrison, S. M. et al. Scaling resolution of variant classification differences in ClinVar between 41 clinical laboratories through an outlier approach. *Hum. Mutat.* **39**, 1641–1649 (2018).
26. Campuzano, O. et al. Reanalysis and reclassification of rare genetic variants associated with inherited arrhythmogenic syndromes. *EBioMedicine* **54**, 102732 (2020).
27. Udler, M. S., McCarthy, M. I., Florez, J. C. & Mahajan, A. Genetic risk scores for diabetes diagnosis and precision medicine. *Endocr. Rev.* **40**, 1500–1520 (2019).
28. Hattersley, A. T. et al. ISPAD clinical practice consensus guidelines 2018: the diagnosis and management of monogenic diabetes in children and adolescents. *Pediatr. Diabetes* **19**, 47–63 (2018).
29. Home—Genetic Testing Registry (GTR)—NCBI.
30. Chakera, A. J. et al. Recognition and management of individuals with hyperglycemia because of a heterozygous glucokinase mutation. *Diabetes Care* **38**, 1383–1392 (2015).
31. Steele, A. M. et al. Use of HbA1c in the identification of patients with hyperglycaemia caused by a glucokinase mutation: observational case control studies. *PLoS One* **8**, e65326 (2013).
32. Patel, K. A. et al. Heterozygous RFX6 protein truncating variants are associated with MODY with reduced penetrance. *Nat. Commun.* **8**, 888 (2017).
33. Naylor. R., Johnson, A. K. & del Gaudio, D. Maturity-Onset Diabetes of the Young Overview. 2018 May 24. In: A (eds Adam, M. P. et al.) GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993–2021. Available from: https://www.ncbi.nlm.nih.gov/books/NBK500456/
34. Fajans, S. S., Bell, G. I. & Polonsky, K. S. Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young. *N. Engl. J. Med.* **345**, 971–980 (2001).
35. Tybjaerg-Hansen, A. et al. Phenotype of heterozygotes for low-density lipoprotein receptor mutations identified in different background populations. *Arterioscler. Thromb. Vasc. Biol.* **25**, 211–215 (2005).
36. Lange, L. A. et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am. J. Hum. Genet.* **94**, 233–245 (2014).
37. Aschard, H. et al. Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. *Hum. Genet.* **131**, 1591–1613 (2012).
38. Fahed, A. C. et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat. Commun.* **11**, 3635 (2020).
39. Stutzmann, F. et al. Prevalence of melanocortin-4 receptor deficiency in Europeans and their age-dependent penetrance in multigenerational pedigrees. *Diabetes* **57**, 2511–2518 (2008).
40. Chatterjee, N. et al. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45**, 400–405 (2013).
41. Zhang, Y., Qi, G., Park, J. H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* **50**, 1318–1326 (2018).
42. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
43. Khera, A. V. et al. Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell* **177**, 587–596.e589 (2019).
44. Narasimhan, V. M. et al. Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**, 474–477 (2016).
45. Matschinsky, F. M. & Wilson, D. F. The central role of glucokinase in glucose homeostasis: a perspective 50 years after demonstrating the presence of the enzyme in islets of langerhans. *Front. Physiol.* **10**, 148 (2019).
46. Tybjaerg-Hansen, A., Steffensen, R., Meinertz, H., Schnohr, P. & Nordestgaard, B. G. Association of mutations in the apolipoprotein B gene with hypercholesterolemia and the risk of ischemic heart disease. *N. Engl. J. Med.* **338**, 1577–1584 (1998).
47. Sorscher, S. Ascertainment bias and estimating penetrance. *JAMA Oncol.* **4**, 587 (2018).
48. Paquette, M. et al. Polygenic risk score predicts prevalence of cardiovascular disease in patients with familial hypercholesterolemia. *J. Clin. Lipidol.* **11**, 725–732.e725 (2017).
49. Trinder, M. et al. Risk of premature atherosclerotic disease in patients with monogenic versus polygenic familial hypercholesterolemia. *J. Am. Coll. Cardiol.* **74**, 512–522 (2019).
50. Mars, N. et al. The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nat. Commun.* **11**, 6383 (2020).
51. Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
52. Zuk, O. et al. Searching for missing heritability: designing rare variant association studies. *Proc. Natl Acad. Sci. U.S.A.* **111**, E455–E464 (2014).

53. Rivera-Muñoz, E. A. et al. ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation. *Hum. Mutat.* **39**, 1614–1622 (2018).
54. Cholesterol Treatment Trialists, C. et al. Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170,000 participants in 26 randomised trials. *Lancet* **376**, 1670–1681 (2010).
55. Zhao, Z. et al. Comparative efficacy and safety of lipid-lowering agents in patients with hypercholesterolemia: a frequentist network meta-analysis. *Med. (Baltim.)* **98**, e14400 (2019).
56. Patel, A. P. et al. Association of rare pathogenic DNA variants for familial hypercholesterolemia, hereditary breast and ovarian cancer syndrome, and lynch syndrome with disease risk in adults according to family history. *JAMA Netw. Open* **3**, e203959 (2020).
57. 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).
58. Eastwood, S. V. et al. Algorithms for the capture and adjudication of prevalent and incident diabetes in UK biobank. *PLoS One* **11**, e0162388 (2016).
59. Genomes Project, C., et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
60. Bellenguez, C. et al. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics* **28**, 134–135 (2012).
61. Sharifi, M., Futema, M., Nair, D. & Humphries, S. E. Genetic architecture of familial hypercholesterolaemia. *Curr. Cardiol. Rep.* **19**, 44 (2017).
62. Peterson, A. S., Fong, L. G. & Young, S. G. PCSK9 function and physiology. *J. Lipid Res.* **49**, 1152–1156 (2008).
63. Whitfield, A. J., Barrett, P. H. R., van Bockxmeer, F. M. & Burnett, J. R. Lipid disorders and mutations in the APOB gene. *Clin. Chem.* **50**, 1725–1732 (2004).
64. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
65. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
66. Cummings, B. B. et al. Transcript expression-aware annotation improves rare variant interpretation. *Nature* **581**, 452–458 (2020).
67. Abou Tayoun, A. N. et al. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum. Mutat.* **39**, 1517–1524 (2018).
68. National Cholesterol Education Program Expert Panel On Detection, Evaluation & Treatment of High Blood Cholesterol in Adults. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* **106**, 3143–3421 (2002).
69. Kwiterovich, P. O. Jr. Diagnosis and management of familial dyslipoproteinemias. *Curr. Cardiol. Rep.* **15**, 371 (2013).
70. Weissglas-Volkov, D. & Pajukanta, P. Genetic causes of high and low serum HDL-cholesterol. *J. Lipid Res.* **51**, 2032–2057 (2010).
71. Surakka, I. et al. The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.* **47**, 589–597 (2015).
72. Vilhjálmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
73. Zheng, J. et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
74. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

Acknowledgements

This work was supported by NIH/NIDDK U01 DK105554 to JCF. This research has been conducted using the UK Biobank Resource under application number 27892. MSU is supported by NIH/NIDDK K23 DK114551. AODL was supported by NIH/NICHD K12 HD052896. MB is supported by NIH/NIDDK DK062370. JCF is also supported by NIH/NIDDK K24 DK110550. JMM is supported by American Diabetes Association Innovative and Clinical Translational Award 1-19-ICTS-068. Please see Supplementary Information for additional Extended Acknowledgements.

Author contributions

Leadership: M.S.U, D.McA., J.C.F., J.F., M.I.M., M.B., N.P.B. Data Analysis: J.G., M.S.U, J.B.C., A.D., N.Z., J.M.M. Variant Curation: J.G., D.McA., S.B., A.O.D-L, M.S-B, R.S., A.S., J.W., E.E., T.I.P., H.Z., K.A.M. Project management/support roles: N.P.B., L.C. B.W., N.W., P.D., R.K., Z.Z. Data generation: A.C., R.A.D., G.A., N.B., J.B., E.B., D.W.B., J.C.C., N.C., E.C., J.C., C.-Y.C., Y.S.C., R.D., J.D., C.G., B.G., S.H., C.L.H., M.Y.H., H.M.K., B.-J.K., Y.J.K., H.A.K., J.S.K., J.K., S.-H.K., M.L., J.-Y.L., J.L., D.M.L., J.L., R.J.F.L., R.C.M., J.B.M., T.M., K.L.M., A.D.M., A.C.M., M.C.N., C.N.A.P., K.S.P., M.P., D.S., C.S., X.S., R.S., W.Y.S., K.S., T.M.S., E.S.T., C.H.T.T., Y.Y.T., F.T., B.T., J.T., R.M.v.D., R.S.V., J.G. W., T.-Y.W., L.L.B., L.G.,V.L., P.M.N., K.S.S. T.D.S., T.T., N.G., T.H.,M.E.J., A.L., O.P.D. R.W., C.A.A.-S., F.B.-O., F.C.-C., C.C.-C., E.C., M.E.G.-S., H.G.-O., C.G.-V., M.E.G., C.H., B.E.H., J.M.M.-H., S.I.-A., A.M.-H., E.M.-C., L.O.,C.R.-M., T.T.-L. E.B., M.G., N.L.H.-C., L.L., C.J.O'D., W.S.P, B.M.P., A.P.R., S.S.R., J.I.R., R.P.T.

Competing interests

M.I.M. has served on advisory panels for Pfizer, Novo Nordisk, and Zoe Global; has received honoraria from Merck, Pfizer, Novo Nordisk, and Eli Lilly; has stock options in Zoe Global; and has received research funding from Abbvie, Astra Zeneca, Boehringer Ingelheim, Eli Lilly, Janssen, Merck, Novo Nordisk, Pfizer, Roche, Sanofi Aventis, and Servier & Takeda. M.I.M. is an employee of Genentech and holds stock in Roche. Psaty serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. C.J.O'D. is an employee of the Novartis Institute for Biomedical Research. All other authors reported no relevant competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-23556-4.

Correspondence and requests for materials should be addressed to M.S.U.

Peer review information *Nature Communications* thanks David Ledbetter and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

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) 2021

Julia K. Goodrich[1], Moriel Singer-Berk[1], Rachel Son[1], Abigail Sveden[1], Jordan Wood[1], Eleina England[1], Joanne B. Cole[1], Ben Weisburd[1], Nick Watts[1], Lizz Caulkins[1], Peter Dornbos[1], Ryan Koesterer[1], Zachary Zappala[1], Haichen Zhang[2], Kristin A. Maloney[2], Andy Dahl[3], Carlos A. Aguilar-Salinas[4], Gil Atzmon[5,6,7], Francisco Barajas-Olmos[8], Nir Barzilai[5,7], John Blangero[9], Eric Boerwinkle[10,11], Lori L. Bonnycastle[12], Erwin Bottinger[13], Donald W. Bowden[14,15,16], Federico Centeno-Cruz[8],

John C. Chambers[17,18], Nathalie Chami[13,19], Edmund Chan[20], Juliana Chan[21,22,23,24], Ching-Yu Cheng[25,26,27], Yoon Shin Cho[28], Cecilia Contreras-Cubas[8], Emilio Córdova[8], Adolfo Correa[29], Ralph A. DeFronzo[30], Ravindranath Duggirala[9], Josée Dupuis[31], Ma Eugenia Garay-Sevilla[32], Humberto García-Ortiz[8], Christian Gieger[33,34,35], Benjamin Glaser[36], Clicerio González-Villalpando[37], Ma Elena Gonzalez[38], Niels Grarup[39], Leif Groop[40,41], Myron Gross[42], Christopher Haiman[43], Sohee Han[44], Craig L. Hanis[10], Torben Hansen[39], Nancy L. Heard-Costa[45,46], Brian E. Henderson[43,130], Juan Manuel Malacara Hernandez[32], Mi Yeong Hwang[44], Sergio Islas-Andrade[8], Marit E. Jørgensen[47,48,49], Hyun Min Kang[50], Bong-Jo Kim[44], Young Jin Kim[44], Heikki A. Koistinen[51,52,53], Jaspal Singh Kooner[54,55,56,57], Johanna Kuusisto[58], Soo-Heon Kwak[59], Markku Laakso[58], Leslie Lange[60], Jong-Young Lee[61], Juyoung Lee[44], Donna M. Lehman[30], Allan Linneberg[62,63,64], Jianjun Liu[20,65,66], Ruth J. F. Loos[13,19], Valeriya Lyssenko[38,67], Ronald C. W. Ma[21,22,23,24], Angélica Martínez-Hernández[8], James B. Meigs[1,68,69], Thomas Meitinger[70,71], Elvia Mendoza-Caamal[8], Karen L. Mohlke[72], Andrew D. Morris[73,74], Alanna C. Morrison[10], Maggie C. Y. Ng[14,15,16], Peter M. Nilsson[75], Christopher J. O'Donnell[76,77,78,79], Lorena Orozco[8], Colin N. A. Palmer[80], Kyong Soo Park[59,81,82], Wendy S. Post[83], Oluf Pedersen[39], Michael Preuss[13], Bruce M. Psaty[84,85], Alexander P. Reiner[86], Cristina Revilla-Monsalve[8], Stephen S. Rich[87], Jerome I. Rotter[88], Danish Saleheen[89,90,91], Claudia Schurmann[13,92,93], Xueling Sim[65], Rob Sladek[94,95,96], Kerrin S. Small[97], Wing Yee So[21,22,23], Timothy D. Spector[97], Konstantin Strauch[98,99], Tim M. Strom[70,100], E. Shyong Tai[20,27,65], Claudia H. T. Tam[21,22,23], Yik Ying Teo[65,101,102], Farook Thameem[103], Brian Tomlinson[104], Russell P. Tracy[105,106], Tiinamaija Tuomi[40,41,107,108,109], Jaakko Tuomilehto[110,111,112,113], Teresa Tusié-Luna[114,115], Rob M. van Dam[20,65,116], Ramachandran S. Vasan[45,117], James G. Wilson[118], Daniel R. Witte[119,120], Tien-Yin Wong[25,26,27], AMP-T2D-GENES Consortia, Noël P. Burtt[1], Noah Zaitlen[3], Mark I. McCarthy[121,122,129], Michael Boehnke[50], Toni I. Pollin[2], Jason Flannick[1,123,124], Josep M. Mercader[1,125,126], Anne O'Donnell-Luria[1,123,124], Samantha Baxter[1], Jose C. Florez[1,125,126], Daniel G. MacArthur[1,127,128] & Miriam S. Udler[1,125,126]✉

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]School of Medicine, University of Maryland Baltimore, Baltimore, MD, USA. [3]Department of Neurology, UCLA, Los Angeles, CA, USA. [4]Instituto Nacional de Ciencias Medicas y Nutricion, Mexico City, Mexico. [5]Department of Medicine, Albert Einstein College of Medicine, New York, NY, USA. [6]Faculty of Natural Science, University of Haifa, Haifa, Israel. [7]Department of Genetics, Albert Einstein College of Medicine, New York, NY, USA. [8]Instituto Nacional de Medicina Genómica, Mexico City, Mexico. [9]Department of Human Genetics and South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley, Brownsville and Edinburg, TX, USA. [10]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA. [11]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [12]Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [13]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [14]Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [15]Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [16]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. [17]Department of Epidemiology and Biostatistics, Imperial College London, London, UK. [18]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore. [19]The Mindich Child Health and Development Institute, Ichan School of Medicine at Mount Sinai, New York, NY, USA. [20]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore. [21]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China. [22]Chinese University of Hong Kong-Shanghai Jiao Tong University Joint Research Centre in Diabetes Genomics and Precision Medicine, The Chinese University of Hong Kong, Hong Kong, China. [23]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China. [24]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China. [25]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore. [26]Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore. [27]Duke-NUS Medical School, Singapore, Singapore. [28]Department of Biomedical Science, Hallym University, Chuncheon, South Korea. [29]Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA. [30]Department of Medicine, University of Texas Health San Antonio (aka University of Texas Health Science Center at San Antonio), San Antonio, TX, USA. [31]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [32]Department of Medical Science, División of Health Science, University of Guanjuato. Campus León. León, Guanjuato, Mexico. [33]Research Unit Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [34]Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [35]German Center for Diabetes Research (DZD), Neuherberg, Germany. [36]Endocrinology and Metabolism Service, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [37]Unidad de Investigacion en Diabetes y Riesgo Cardiovascular, Instituto Nacional de Salud Publica, Cuernavaca, Mexico. [38]Centro de Estudios en Diabetes, Mexico City, Mexico. [39]Novo

Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [40]Department of Clinical Sciences, Diabetes and Endocrinology, Lund University Diabetes Centre, Malmö, Sweden. [41]Institute for Molecular Genetics Finland, University of Helsinki, Helsinki, Finland. [42]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN, USA. [43]Department of Preventive Medicine, Keck School of Medicine of USC, Los Angeles, CA, USA. [44]Division of Genome Research, Center for Genome Science, National Institute of Health, Chungcheongbuk-do, South Korea. [45]Boston University and National Heart Lung and Blood Institute's Framingham Heart Study, Framingham, MA, USA. [46]Department of Neurology, Boston University School of Medicine, Boston, MA, USA. [47]Steno Diabetes Center Copenhagen, Gentofte, Denmark. [48]National Institute of Public Health, University of Southern Denmark, Copenhagen, Denmark. [49]Greenland Centre for Health Research, University of Greenland, Nuuk, Greenland. [50]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [51]Department of Public Health Solutions, Finnish Institute for Health and Welfare, Helsinki, Finland. [52]University of Helsinki and Department of Medicine, Helsinki University Central Hospital, Helsinki, Finland. [53]Minerva Foundation Institute for Medical Research, Helsinki, Finland. [54]Department of Cardiology, Ealing Hospital, London North West Healthcare NHS Trust, London, UK. [55]MRC-PHE Centre for Environment and Health, Imperial College London, London, UK. [56]Imperial College Healthcare NHS Trust, Imperial College London, London, UK. [57]National Heart and Lung Institute, Imperial College London, London, UK. [58]Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland. [59]Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea. [60]Department of Medicine, University of Colorado Denver, Anschutz Medical Campus, Aurora, CO, USA. [61]Oneomics Soonchunhyang Mirae Medical Center, Bucheon-si Gyeonggi-do, Republic of Korea. [62]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [63]Center for Clinical Research and Prevention, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark. [64]Department of Clinical Experimental Research, Rigshospitalet, Copenhagen, Denmark. [65]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore. [66]Genome Institute of Singapore, Agency for Science Technology and Research, Singapore, Singapore. [67]Department of Clinical Science, University of Bergen, Bergen, Norway. [68]Department of Medicine, Harvard Medical School, Boston, MA, USA. [69]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA. [70]Institute of Human Genetics, Technical University of Munich, Munich, Germany. [71]German Centre for Cardiovascular Research (DZHK), Partner Site Munich Heart Alliance, Munich, Germany. [72]Department of Genetics, University of North Carolina Chapel Hill, Chapel Hill, NC, USA. [73]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [74]Department of Biostatistics, University of Liverpool, Liverpool, UK. [75]Department of Clinical Sciences, Medicine, Lund University, Malmö, Sweden. [76]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [77]Section of Cardiology, Department of Medicine, VA Boston Healthcare, Boston, MA, USA. [78]Brigham and Women's Hospital, Boston, MA, USA. [79]Intramural Administration Management Branch, National Heart Lung and Blood Institute, NIH, Framingham, MA, USA. [80]Pat Macpherson Centre for Pharmacogenetics and Pharmacogenomics, University of Dundee, Dundee, UK. [81]Department of Internal Medicine, Seoul National University College of Medicine, Seoul, South Korea. [82]Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, South Korea. [83]Division of Cardiology, Department of Medicine, Johns Hopkins University, Baltimore, MD, USA. [84]Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology, and Health Services, University of Washington, Seattle, WA, USA. [85]Kaiser Permanente Washington Research Institute, Seattle, WA, USA. [86]Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [87]Center for Public Health Genomics, University of Virginia School of Medicine, Charlottesville, VA, USA. [88]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation (formerly Los Angeles Biomedical Research Institute) at Harbor-UCLA Medical Center, Torrance, CA, USA. [89]Division of Translational Medicine and Human Genetics, University of Pennsylvania, Philadelphia, PA, USA. [90]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA, USA. [91]Center for Non-Communicable Diseases, Karachi, Pakistan. [92]Digital Health Center, Hasso Plattner Institute, University of Potsdam, Prof.-Dr.-Helmert-Str. 2-3, Potsdam, Germany. [93]Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, NY, USA. [94]Department of Human Genetics, McGill University, Montreal, QC, Canada. [95]Division of Endocrinology and Metabolism, Department of Medicine, McGill University, Montreal, QC, Canada. [96]McGill University and Génome Québec Innovation Centre, Montreal, QC, Canada. [97]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [98]Institute of Genetic Epidemiology, Helmholtz Zentrum Munchen, German Research Center for Environmental Health, Neuherberg, Germany. [99]Institute for Medical Informatics Biometry and Epidemiology, Ludwig-Maximilians University, Munich, Germany. [100]Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [101]Life Sciences Institute, National University of Singapore, Singapore, Singapore. [102]Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore. [103]Department of Biochemistry, Faculty of Medicine, Health Science Center, Kuwait University, Safat, Kuwait. [104]Faculty of Medicine, Macau University of Science & Technology, Macau, China. [105]Department of Pathology and Laboratory Medicine, The Robert Larner M.D. College of Medicine, University of Vermont, Burlington, VT, USA. [106]Department of Biochemistry, The Robert Larner M.D. College of Medicine, University of Vermont, Burlington, VT, USA. [107]Folkhälsan Research Centre, Helsinki, Finland. [108]Department of Endocrinology, Abdominal Centre, Helsinki University Hospital, Helsinki, Finland. [109]Research Programs Unit, Clinical and Molecular Medicine, University of Helsinki, Helsinki, Finland. [110]Public Health Promotion Unit, Finnish Institute for Health and Welfare, Helsinki, Finland. [111]Department of Public Health, University of Helsinki, Helsinki, Finland. [112]Saudi Diabetes Research Group, King Abdulaziz University, Jeddah, Saudi Arabia. [113]Department of International Health, National School of Public Health, Instituto de Salud Carlos III, Madrid, Spain. [114]Unidad de Biología Molecular y Medicina Genómica, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico. [115]Departamento de Medicina Genómica y Toxiología Ambiental, Instituto de Investigaciones Biomédicas, UNAM, Mexico City, Mexico. [116]Department of Nutrition, Harvard School of Public Health, Boston, MA, USA. [117]Preventive Medicine & Epidemiology, and Cardiovascular Medicine, Medicine, Boston University School of Medicine, and Epidemiology, Boston University School of Public health, Boston, MA, USA. [118]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA. [119]Department of Public Health, Aarhus University, Aarhus, Denmark. [120]Danish Diabetes Academy, Odense, Denmark. [121]Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [122]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [123]Division of Genetics and Genomics, Boston Children's Hospital, Boston, Massachusetts, USA. [124]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [125]Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [126]Department of Medicine, Harvard Medical School, Boston, MA, USA. [127]Centre for Population Genomics, Garvan Institute of Medical Research, UNSW Sydney, Sydney, NSW, Australia. [128]Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, VIC, Australia. [129]Present address: Genentech, South San Francisco, CA, USA. [130]Deceased: Brian E. Henderson. A list of authors and their affiliations appears in the Supplementary Information. ✉email: mudler@mgh.harvard.edu