Gene-level analysis of rare variants in 379,066 whole exome sequences identifies an association of GIGYF1 loss of function with type 2 diabetes

Aimee M. Deaton1, Margaret M. Parker1, Lucas D. Ward1, Alexander O. Flynn-Carroll1, Lucas BonDurant1, Gregory Hinkle1, Parsa Akbari2, Luca A. Lotta2, Regeneron Genetics Center1, DiscovEHR Collaboration1, Aris Baras2 & Paul Nioi1

Sequencing of large cohorts offers an unprecedented opportunity to identify rare genetic variants and to find novel contributors to human disease. We used gene-based collapsing tests to identify genes associated with glucose, HbA1c and type 2 diabetes (T2D) diagnosis in 379,066 exome-sequenced participants in the UK Biobank. We identified associations for variants in GCK, HNF1A and PDX1, which are known to be involved in Mendelian forms of diabetes. Notably, we uncovered novel associations for GIGYF1, a gene not previously implicated by human genetics in diabetes. GIGYF1 predicted loss of function (pLOF) variants associated with increased levels of glucose (0.77 mmol/L increase, \( p = 4.42 \times 10^{-12} \)) and HbA1c (4.33 mmol/mol, \( p = 1.28 \times 10^{-14} \)) as well as T2D diagnosis (OR = 4.15, \( p = 6.14 \times 10^{-11} \)). Multiple rare variants contributed to these associations, including singleton variants. GIGYF1 pLOF also associated with decreased cholesterol levels as well as an increased risk of hypothyroidism. The association of GIGYF1 pLOF with T2D diagnosis replicated in an independent cohort from the Geisinger Health System. In addition, a common variant association for glucose and T2D was identified at the GIGYF1 locus. Our results highlight the role of GIGYF1 in regulating insulin signaling and protecting from diabetes.

Human genetics provides powerful methods for understanding the roles of genes and proteins in disease and can lead to new therapeutic hypotheses and drug targets. Genetic evidence based on sequence variants within coding regions of the genome is better at predicting the efficacy and safety of novel therapeutics than evidence from genome-wide association studies (GWAS), which tend to involve common noncoding variants. Among coding variants, predicted loss of function (pLOF) variants are particularly informative in association studies because they establish a direct causal link between reduction in gene function and biological outcomes. Additionally, rare missense variants predicted to be deleterious can provide valuable biological insights. However, interrogation of the effects of such variants is hampered by the rarity of these variants and the cohort sizes needed to identify associations. Exome or whole-genome sequencing of large biobanks coupled with gene-level aggregation of rare high impact variants can help to circumvent these challenges. Biobanks offer a considerable advantage over case–control cohorts as they contain richer phenotyping data which often includes biomarker measurements as well as disease diagnoses. This allows a more complete understanding of the biological consequences of damaging variants in particular genes.

Type 2 diabetes (T2D) is a disease that has been extensively studied in traditional array-based GWAS with hundreds of associations identified to date. Although these studies have given insight into some of the biological mechanisms contributing to T2D, most of the reported associations are with variants in non-coding regions, making identification of the causal gene challenging. More recently, exome sequencing has been applied to discover protein-coding variants that alter the risk of developing T2D. Sequencing of 20,791 T2D cases followed...
by the use of gene-based collapsing tests (to aggregate predicted damaging variants) identified associations of SLC30A8, MC4R and PAM with T2D diagnosis.

Using whole exome sequences from the UK Biobank (UKBB) we performed gene-level collapsing tests to examine the association of pLOF and damaging missense variants in ~17,000 genes with biomarkers of glycemic control, glucose, and glycated hemoglobin (HbA1c), as well as T2D diagnosis.

Results

Gene-level associations with glucose, HbA1c and T2D. We used 454,787 whole exome sequences from the UK Biobank (UKBB) to identify rare variants with a minor allele frequency (MAF) ≤ 1% likely to have functional impact; pLOF variants (i.e. frameshift, stop gain, splice donor or splice acceptor variants) called as high confidence by LOFTEE13 or missense variants predicted to be damaging (Combined Annotation Dependent Depletion [CADD] score ≥ 25). We identified 726,422 rare pLOF variants affecting 16,477 genes, 58.5% of which were singletons (carried by a single individual), and 2.14 million damaging missense variants in 17,312 genes, 49.6% of which were singletons (Supplementary Table 1).

Given the large proportion of variants present in just a single individual, we used gene-based collapsing tests to look for associations with biomarkers of glycemic control and T2D diagnosis. We used two variant aggregation strategies; (1) pLOF variants with MAF ≤ 1% and (2) damaging missense variants with MAF ≤ 1% and performed burden testing in the unrelated European ancestry population (n = 363,977) adjusting for age, sex and genetic ancestry via 12 principal components.

First, we tested genes for association with glucose and HbA1c levels. We required at least ten variant carriers per gene to have measurements based on an examination of genomic inflation at different carrier thresholds (Supplementary Figure 1). Using a p-value threshold adjusted for the number of variant sets and phenotypes tested (p ≤ 7.82 × 10−7), four variant sets in three genes significantly associated with glucose levels: GCK pLOF (p = 1.56 × 10−5, 1.24 mmol/L increase), GCK damaging missense (p = 6.15 × 10−11, 0.61 mmol/L increase), GIGYF1 pLOF (p = 4.42 × 10−13, 0.77 mmol/L increase) and G6PC2 damaging missense variants (p = 4.62 × 10−43, 0.33 mmol/L decrease) (Fig. 1, Table 1). The same variant sets also associated with HbA1c levels along with 27 additional sets including HNF1A pLOF (p = 2.14 × 10−7, 4.01 nmol/mol increase), TNRC6B pLOF (p = 2.36 × 10−7, 3.94 mmol/mol increase) and PDX1 damaging missense variants (p = 2.54 × 10−7, 0.41 nmol/mol increase) (Fig. 1, Table 1 and Supplementary Table 2). We then tested aggregated pLOF and damaging missense variants in all genes for association with T2D diagnosis (n = 24,695 cases). Six variant sets significantly associated with T2D (p ≤ 1.46 × 10−10); pLOF variants in GIGYF1, GCK, HNF1A and TNRC6B and damaging missense variants in GCK and PAM (Fig. 2, Table 2 and Supplementary Table 2). As the time of available follow-up differs between England, Scotland, and Wales, we controlled for country of recruitment in the regression. In addition, we confirmed that significant hits did not associate with country of recruitment (all p > 0.035) and that hits remained significant when only data from England were considered (Supplementary Table 3). Our results also remained significant when we adjusted for follow-up time in the burden test (Supplementary Table 4).

Sex-specific analyses of glucose, HbA1c and T2D were performed but did not identify associations for additional genes (Supplementary Tables 5–7). Meta-analysis of results for males and females gave similar effect estimates to the primary analysis, with one additional association detected, CREB3L3 pLOF and HbA1c levels. We did not observe significant heterogeneity in effects between sexes (Supplementary Tables 8, 9, 10).

We also tested genes for association with glucose, HbA1c and T2D in the African ancestry (n = 6221), South Asian ancestry (n = 7553) and East Asian ancestry (n = 1315) populations of UKBB. The only statistically significant association was between MYH6 damaging missense variants and HbA1c levels in the East Asian ancestry population (p = 1.11 × 10−8, 5.10 mmol/mol decrease) (Supplementary Table 11). Several of the variant sets significant in our primary analysis were nominally significant in the other populations. These included G6PC2 damaging missense variants which associated with decreased glucose levels in the African and South Asian ancestry populations (p = 0.01 and p = 8.31 × 10−4 respectively, Supplementary Tables 11 and 12). Carriers of GIGYF1 pLOF variants were too rare to test associations of these variants in the other populations.

Identification of genes with a biological role in diabetes. Variants in two genes, GCK and GIGYF1, significantly associated with glucose, HbA1c and T2D diagnosis, strongly suggesting a biological role in diabetes; GCK is involved in Mendelian forms of diabetes while GIGYF1 has not previously been implicated by genetics in the disease. Both GCK and GIGYF1 are located on chromosome 7 but are 56 Mb apart, strongly suggesting that these signals are independent; this independence was confirmed by conditional analysis (Supplementary Table 13). Two additional variant sets, HNF1A pLOF and TNRC6B pLOF, had genome-wide associations with both T2D diagnosis and HbA1c levels while G6PC2 damaging missense variants associated with decreased levels of both glucose and HbA1c but not T2D diagnosis (Table 3).

To see which other significant genes were likely to have a role in diabetes we looked at all variant sets with a significant glucose, HbA1c, or T2D association and examined whether they had associations with additional diabetes traits (p ≤ 0.0016, correcting for 32 sets tested). Damaging missense variants in PDX1 and PPFAS, which significantly associated with HbA1c levels in our primary analysis, associated with T2D diagnosis using this threshold (Table 3 and Supplementary Table 14).

Many HbA1c associations appeared to be secondary to effects on red blood cells. 22 out of 31 variant sets associated with HbA1c did not show effects on glucose levels or T2D diagnosis (Supplementary Table 14) and were not implicated in Mendelian forms of diabetes. Out of these 22 variant sets, 12 were in genes implicated in Mendelian disorders affecting red blood cells (for example EPB42 and TFR2, see Supplementary Table 15).
Figure 1. Gene-level associations with glucose and HbA1c levels. (A) pLOF associations with glucose levels. (B) Damaging missense variant (CADD score ≥ 25) associations with glucose levels. (C) pLOF associations with HbA1c. (D) Damaging missense variant associations with HbA1c levels. The red line indicates the threshold for significance, genes with significant associations are labeled.
and an additional five had highly significant associations with red blood cell traits in our data \((p \leq 7.82 \times 10^{-7}; \text{Supplementary Table 16})\).

We focused on the variant sets associated with multiple diabetes traits as these are strong candidates for regulating glucose homeostasis and/or causally contributing to T2D risk. The genes fall into three main groups; known MODY (maturity-onset diabetes of the young) genes \((\text{GCK}, \text{HNF1A} \text{ and } \text{PDX1})\)\(^{14}\), known genes reported in previous exome-wide analyses of glucose levels or T2D \((\text{G6PC2} \text{ and } \text{PAM})\)\(^{5,15}\), and novel genes not previously implicated by genetics in diabetes \((\text{GIGYF1}, \text{TNRC6B} \text{ and } \text{PFAS})\).

Because obesity is linked to the development of T2D, we adjusted for body mass index (BMI) in the regression and found that the association of these genes with diabetes-related traits remained significant (Supplementary Tables 17 and 18).

We used the generalized linear mixed model implemented by SAIGE-Gene which accounts for relatedness and adjusts for unbalanced case–control ratios\(^{16}\) to verify association of our variant sets of interest with glucose, HbA1c, and T2D diagnosis. SAIGE-Gene was run in the European ancestry population including related individuals \((n = 398,574)\). Using the p-value thresholds previously employed, all associations were statistically significant using this method apart from the associations of \text{TNRC6B} pLOF with HbA1c \((p = 6.85 \times 10^{-6})\) and T2D diagnosis \((p = 4.77 \times 10^{-5})\) which were less significant (Supplementary Table 19).

To maximize power to detect associations for rare variants, our original analysis of glucose and HbA1c included individuals with a diabetes diagnosis. Associations for all variant sets of interest were at least nominally significant when such individuals were excluded from the analysis (Supplementary Table 20). For \text{GIGYF1}
There was still a substantial effect on glucose ($p = 2.95 \times 10^{-8}$, effect = 0.53 SD) and HbA1c ($p = 8.29 \times 10^{-7}$, effect = 0.43 SD) levels in carriers without a formal diabetes diagnosis.

For the MODY genes, $GCK$, $HNF1A$, and $PDX1$, we also tested pathogenic or likely pathogenic variants annotated in ClinVar17 for association with glucose, HbA1c and T2D—either alone or combined with the variants used in our primary analysis. Pathogenic and likely pathogenic variants in $GCK$ and $HNF1A$ associated with T2D diagnosis and markers of glycemic control and added signal beyond what was seen in the primary analysis (Supplementary Tables 21 and 22).

**Validation of $GIGYF1$ pLOF associations using independent datasets.** We sought to use independent measurements of glucose and HbA1c to perform sensitivity analysis and verify the associations of interest seen in our primary analysis which used measurements taken as part of the UKBB assessment. To do this, we used independent measurements of glucose and HbA1c to perform sensitivity analysis and verify the associations of interest seen in our primary analysis which used measurements taken as part of the UKBB assessment. To do this, we

| Gene    | Variant set | Title | p-value     | OR  | 95% CI– | 95% CI+ | N cases | N carrier | N carrier cases | N expected |
|---------|-------------|-------|-------------|-----|---------|---------|---------|-----------|---------------|------------|
| GCK     | pLOF        | T2D   | 2.96 × 10^{-15} | 14.16 | 7.33    | 27.34   | 24,695  | 40        | 19           | 2.71       |
| GIGYF1  | pLOF        | T2D   | 6.14 × 10^{-11} | 4.15  | 2.71    | 6.37    | 24,695  | 131       | 29          | 8.89       |
| HNF1A   | pLOF        | T2D   | 1.23 × 10^{-9}  | 5.27  | 3.08    | 9       | 24,695  | 73        | 20          | 4.95       |
| TNRC6B  | pLOF        | T2D   | 2.00 × 10^{-7}  | 4.44  | 2.53    | 7.79    | 24,695  | 71        | 17          | 4.82       |
| PAM     | missense    | T2D   | 2.26 × 10^{-12} | 1.31  | 1.21    | 1.41    | 24,695  | 9357      | 801         | 634.85     |
| GCK     | missense    | T2D   | 1.70 × 10^{-4}  | 2.96  | 2.03    | 4.32    | 24,695  | 202       | 34          | 13.71      |

Table 2. Gene-level associations with T2D diagnosis. pLOF or damaging missense variants (CADD score ≥ 25) were aggregated per gene. OR odds ratio, CI confidence interval.

A red line indicates the threshold for significance, genes with significant associations are labeled.

**Figure 2.** Gene-level associations with T2D. (A) pLOF associations. (B) Damaging missense variant (CADD score ≥ 25) associations. The red line indicates the threshold for significance, genes with significant associations are labeled.
we extracted lab test values for glucose and HbA1c from primary care data, which is available for approximately half of the cohort. In gene-based burden tests all variant sets showed a direction of effect consistent with that seen in the primary analysis and 10 out of 12 of these were significant when correcting for the number of tests performed (p ≤ 0.004). This included the association of GIGYF1 pLOF with glucose (p = 2.10 × 10⁻⁸, effect = 0.65 SD) and HbA1c (p = 1.19 × 10⁻⁵, effect = 0.74 SD) levels (Supplementary Figure 2 and Supplementary Table 23).

We then assessed whether rare variants in GIGYF1 and the other novel genes associated with T2D replicated in a completely independent exome-sequencing cohort. Gene-based tests in European ancestry individuals from the Geisinger Health System (GHS; 25,846 T2D cases and 63,749 controls) confirmed the association of GIGYF1 pLOF with T2D (p = 0.01, OR = 1.8). We did not replicate the association of TNRC6B pLOF with T2D. We tested an expanded PFAS variant set (pLOF + deleterious missense) and did not detect an association with T2D (Supplementary Table 24). Notably variant set definitions varied somewhat from those used in our primary analysis (see “Methods”).

**Multiple variants contribute to associations with diabetes diagnosis and biomarkers.** To examine whether specific variants were driving the associations with diabetes traits we conducted “leave-one-out” burden tests. The association of PAM missense variants with T2D diagnosis was driven entirely by a previously reported variant Ser539Trp (rs78408340; p = 0.43 when Ser539Trp is excluded). For all other variant sets, multiple variants contributed to the associations observed (Supplementary Figure 3). Notably, when singleton variants were excluded, half of the associations no longer reached significance including those for GCK pLOF and glucose (p = 0.0015 without singletons versus p = 1.56 × 10⁻⁹) and GIGYF1 pLOF and T2D (p = 2.9 × 10⁻⁸ without singletons versus p = 6.14 × 10⁻¹¹) (Supplementary Table 25), demonstrating the power of including singletons in gene-based tests.

For the variants contributing to our novel discovered associations, GIGYF1 pLOF, TNRC6B pLOF and PFAS damaging missense variants, we examined the quality scores, sequencing depth, transcripts affected and presence of contributing variants in the gnomAD database. We found that for GIGYF1 and PFAS the variants contributing most to the associations had good quality scores and depth and were present in gnomAD. In contrast, TNRC6B is a highly constrained gene and the most common pLOF variant is not present in gnomAD (see “Supplementary Note” and Supplementary Figure 4). This observation along with the fact that the association of TNRC6B pLOF with T2D did not replicate in GHS leads us to view this association with suspicion.

**Replication of published gene-level associations with T2D and associations for T2D drug targets.** The association between predicted damaging variants in PAM and T2D diagnosis was previously reported in an exome-sequencing study performed by Flannick et al.². We examined whether the other two significant genes in the study, SLC30A8 and MC4R, associated with diabetes traits in our analysis. Both pLOF and damaging missense variants in SLC30A8 associated with reduced levels of HbA1c and glucose and suggestively associated with decreased incidence of T2D diagnosis (Supplementary Table 26). Combining SLC30A8 pLOF and missense variants resulted in more significant associations with glucose (p = 2.71 × 10⁻⁹), HbA1c (p = 8.64 × 10⁻¹⁰) and T2D diagnosis (p = 0.005) (Supplementary Table 26). There were no MC4R high confidence pLOF variants in our dataset and MC4R predicted damaging missense variants did not associate with diabetes-related traits in our study (all p > 0.19). We note that the MC4R Ile269Asn variant driving the association in Flannick and colleagues’ analysis is absent from our dataset, consistent with the fact that it is absent from European populations in gnomAD.

We also examined whether we detect associations for the 8 genes encoding T2D drug targets (GLP1R, IGF1R, PPARG, INSR, SLC5A2, DPP4, KCNJ11, ABCG8). Variant sets in three of these genes, DPP4, GLP1R and KCNJ11 significantly associated with either T2D diagnosis or HbA1c levels (p ≤ 0.003 correcting for 15 variant sets tested) and an additional 4 genes had a nominally significant association with T2D and/or HbA1c (Supplementary Figure 5 and Supplementary Table 27).

Table 3. Genes and variant sets associated with multiple diabetes-related traits. Variant sets significant for at least one trait in our primary analysis that are also associated with additional diabetes traits (p ≤ 0.0016, 32 sets tested) are shown. Effect is shown in SD of transformed values or as an odds ratio (OR).

| Gene     | Variant set | p-value glucose | Effect glucose | p-value HbA1c | Effect HbA1c | p-value T2D | OR T2D |
|----------|-------------|-----------------|----------------|-------------|-------------|-------------|--------|
| GCK      | pLOF        | 1.56 × 10⁻⁹     | 0.999          | 2.64 × 10⁻⁷  | 1.292       | 2.96 × 10⁻¹⁵ | 14.16  |
| HNF1A    | pLOF        | 0.01            | 0.317          | 2.14 × 10⁻⁷  | 0.592       | 1.23 × 10⁻⁴  | 5.27   |
| GIGYF1   | pLOF        | 4.42 × 10⁻¹²    | 0.616          | 1.28 × 10⁻¹⁵ | 0.639       | 6.14 × 10⁻¹¹ | 4.15   |
| GCK      | missense    | 6.15 × 10⁻¹¹    | 0.487          | 1.86 × 10⁻¹⁵ | 0.565       | 1.70 × 10⁻⁸  | 2.96   |
| PAM      | missense    | 0.92            | 0.001          | 0.009       | 0.026       | 2.26 × 10⁻¹² | 1.31   |
| TNRC6B   | pLOF        | 4.01 × 10⁻³     | −0.507         | 2.36 × 10⁻²  | 0.582       | 2.00 × 10⁻⁷  | 4.44   |
| PDX1     | missense    | 0.02            | 0.029          | 2.54 × 10⁻⁷  | 0.060       | 3.99 × 10⁻⁴  | 1.21   |
| PFAS     | missense    | 0.32            | 0.009          | 2.09 × 10⁻⁴  | −0.048      | 4.43 × 10⁻⁴  | 0.88   |
| G6PC2    | missense    | 4.62 × 10⁻⁹⁵    | −0.266         | 6.71 × 10⁻⁶  | −0.179      | 0.97        | 1.00   |
PhewAS of GIGYF1 pLOF reveals associations with cholesterol levels, hypothyroidism and complications of diabetes. The most significant novel associations were seen for GIGYF1 pLOF which associated with increased glucose and HbA1c levels as well as increased incidence of T2D diagnosis. To give additional insight into the biological roles of GIGYF1 we performed a phenome-wide association study (PhewAS) testing GIGYF1 pLOF for association with 142 quantitative traits and 262 ICD10-coded diagnoses (Fig. 3).

GIGYF1 pLOF strongly associated with decreased levels of total cholesterol ($p = 2.44 \times 10^{-12}$, effect $= -0.61$ SD) which was, in large part, driven by LDL cholesterol ($p = 2.40 \times 10^{-10}$, effect $= -0.56$ SD) although an effect on HDL cholesterol was also observed (Table 4). To understand the extent to which this is influenced by the use of cholesterol-lowering medication in diabetics, we adjusted for medication use in the regression and also performed a separate analysis excluding those on cholesterol-lowering medication. The association between GIGYF1 pLOF and LDL cholesterol levels was significant in both analyses (Supplementary Table 28). GIGYF1 pLOF also associated with decreased grip strength and decreased peak expiratory flow. Notably, GIGYF1 pLOF also associated with increased levels of the kidney injury biomarker cystatin c ($p = 6.65 \times 10^{-6}$, effect $= 0.36$ SD) and increased diagnosis of urinary system disorders ($p = 7.32 \times 10^{-5}$, OR $= 2.71$) (Tables 4 and 5).

After diabetes, the most significant disease association of GIGYF1 pLOF was with increased risk of hypothyroidism ($p = 1.25 \times 10^{-9}$, OR $= 4.53$). 21 out of the 131 GIGYF1 pLOF carriers had a diagnosis of unspecified hypothyroidism and seven of these also had a diagnosis of T2D. Given the autoimmune component in hypothyroidism and type 1 diabetes (T1D), we examined the association of GIGYF1 pLOF with T1D diagnoses but did not detect a significant association ($p = 0.1$). GIGYF1 pLOF also significantly associated with increased risk of syncope and collapse ($p = 1.92 \times 10^{-6}$, OR $= 3.75$) (Table 5).

Other phenome-wide significant associations with quantitative traits included waist circumference, total protein and mean corpuscular hemoglobin as well increased time to complete a cognitive test (Table 4). To ensure that the association of GIGYF1 pLOF with HbA1c was independent of effects on hemoglobin we adjusted for mean corpuscular hemoglobin level and verified that the association remained highly significant ($p = 4.10 \times 10^{-12}$). GIGYF1 pLOF also associated with increased diagnosis of emphysema and anemia (Table 5).
Common variants at **GIGYF1** associate with glucose, T2D and **GIGYF1** expression. We looked for more common variants that could further implicate the **GIGYF1** locus in diabetes. We tested array genotyped and imputed variants at the **GIGYF1** locus for association with glucose levels in 294,042 unrelated European ancestry individuals with measurements available. We found a cluster of variants in a linkage disequilibrium block covering **GIGYF1** and **EPO** significantly associating with glucose levels (Fig. 4). This signal is represented by rs221783, an intergenic variant whose minor T allele associated with decreased glucose (p = 1.8 × 10−11, effect = −0.03 SD) and HbA1c (p = 3.6 × 10−7, effect = −0.02 SD) levels as well as increased cholesterol (p = 7.0 × 10−12, effect = 0.03 SD). This variant also associated with a decreased risk of T2D (p = 0.005, OR = 0.96) and hypothyroidism (p = 6.95 × 10−7, OR = 0.92) (Table 6).

The independent glucose and T2D associations at the **GIGYF1** locus and their replication in other datasets further support the hypothesis that decreasing **GIGYF1** levels predisposes to diabetes while increasing **GIGYF1** levels may protect from diabetes.

### Identification of causal genes at GWAS loci.

Given the fact that the **GIGYF1** locus harbors both rare and common variants associated with T2D we examined whether our study points to the causal gene at additional GWAS loci. For 558 variants associated with T2D in a recent study by Vujkovic et al.9 we tested whether either of the two closest genes associated with T2D or HbA1c levels in our study. Just nine genes close to these

---

**Table 4.** PheWAS of **GIGYF1** pLOF – quantitative traits. Significant associations with quantitative traits are shown (p ≤ 1.22 × 10−4). Effect is shown in standard deviations (SD) of transformed values. RH right hand, LH left hand, CI confidence interval.

| Gene   | Variant set | Title                                | p-value       | Effect (SD) | 95% CI- | 95% CI+ | n carrier measured |
|--------|-------------|--------------------------------------|---------------|-------------|---------|---------|--------------------|
| GIGYF1 | pLOF        | HbA1c                                | 1.28 × 10−14  | 0.64        | 0.48    | 0.80    | 129                |
| GIGYF1 | pLOF        | Cholesterol                          | 2.44 × 10−13  | −0.61       | −0.78   | −0.44   | 128                |
| GIGYF1 | pLOF        | Glucose                              | 4.42 × 10−13  | 0.62        | 0.44    | 0.79    | 121                |
| GIGYF1 | pLOF        | LDL cholesterol                      | 2.40 × 10−13  | −0.56       | −0.73   | −0.38   | 128                |
| GIGYF1 | pLOF        | Apolipoprotein b                     | 2.52 × 10−13  | −0.56       | −0.73   | −0.39   | 127                |
| GIGYF1 | pLOF        | LH grip strength                     | 5.11 × 10−13  | −0.37       | −0.49   | −0.25   | 131                |
| GIGYF1 | pLOF        | RH grip strength                     | 1.01 × 10−6   | −0.34       | −0.46   | −0.23   | 123                |
| GIGYF1 | pLOF        | Peak expiratory flow                 | 5.73 × 10−6   | −0.41       | −0.56   | −0.26   | 114                |
| GIGYF1 | pLOF        | Cystatin c                           | 6.65 × 10−6   | 0.36        | 0.20    | 0.51    | 128                |
| GIGYF1 | pLOF        | Mean corpuscular hemoglobin          | 6.80 × 10−6   | −0.38       | −0.55   | −0.22   | 128                |
| GIGYF1 | pLOF        | HDL cholesterol                      | 1.53 × 10−5   | −0.35       | −0.52   | −0.19   | 121                |
| GIGYF1 | pLOF        | Time to complete round (cognitive test) | 1.67 × 10−5   | 0.35       | 0.19    | 0.51    | 129                |
| GIGYF1 | pLOF        | Waist circumference                  | 3.98 × 10−5   | 0.32       | 0.16    | 0.47    | 130                |
| GIGYF1 | pLOF        | Total protein                        | 6.45 × 10−5   | −0.36       | −0.53   | −0.18   | 121                |
| GIGYF1 | pLOF        | Apolipoprotein a                     | 6.88 × 10−5   | −0.33       | −0.49   | −0.17   | 121                |

**Table 5.** PheWAS of **GIGYF1** pLOF—ICD10-coded diagnoses. Significant associations with ICD10-coded diagnoses are shown (p ≤ 1.22 × 10−4). OR odds ratio, CI confidence interval.

| Gene   | Variant set | Title                                | p-value       | OR   | 95% CI- | 95% CI+ | N cases | N carrier cases | N expected |
|--------|-------------|--------------------------------------|---------------|------|---------|---------|---------|----------------|------------|
| GIGYF1 | pLOF        | E11 T2D                              | 6.14 × 10−11  | 4.15 | 2.71    | 6.37    | 24,695  | 29            | 8.89       |
| GIGYF1 | pLOF        | E03 other hypothyroidism             | 1.25 × 10−6   | 4.53 | 2.78    | 7.38    | 19,417  | 21            | 6.99       |
| GIGYF1 | pLOF        | R55 syncope and collapse             | 1.90 × 10−6   | 3.75 | 2.18    | 6.47    | 12,706  | 15            | 4.57       |
| GIGYF1 | pLOF        | D50 iron deficiency anaemia          | 8.52 × 10−6   | 3.56 | 2.04    | 6.23    | 12,886  | 14            | 4.64       |
| GIGYF1 | pLOF        | J43 emphysema                        | 1.99 × 10−5   | 6.13 | 2.67    | 14.10   | 3015    | 6             | 1.09       |
| GIGYF1 | pLOF        | N39 other disorders of urinary system | 7.32 × 10−5   | 2.71 | 1.66    | 4.45    | 24,581  | 19            | 8.85       |
558 variants significantly associated with T2D or HbA1c (p ≤ 2.41 × 10–5 adjusting for 2071 variant sets tested)—ANK1, GCK, HNF1A, TNRC6B, SLC30A8, NF1, IRS2, CFTR and HNF4A (Supplementary Figure 6 and Supplementary Table 32). Most of these genes are already known to be causal for T2D including GCK, HNF1A, SLC30A8, IRS2 and HNF4A. Given that there is a common variant association with T2D at TNRC6B but conflicting results for TNRC6B pLOF in UKBB and GHS, further study of this locus may be warranted.

**Discussion**

Our results highlight the power of whole exome sequencing to make novel discoveries relevant to human disease and to detect known associations of Mendelian disease genes. Gene-level aggregation and burden testing of rare pLOF and predicted damaging missense variants identified genes associating with T2D and biomarkers of glycemic control. These included several genes not previously implicated in diabetes, GIGYF1, TNRC6B and FFAS, as well as GCK, HNF1A and PDX1, known MODY genes[14,21–23]. We also identified PAM and G6PC2, genes highlighted by other rare-variant studies of T2D and glucose levels[5,15]. Gene-level tests were needed to detect the majority of these associations owing to the rarity of the variants. For example, out of 363,977 individuals, just 40 carried a pLOF variant in GCK and 131 carried a pLOF variant in GIGYF1. In general, singleton variants contributed a large part of the signal arguing strongly, as others have done[4], for including such variants in gene-based collapsing tests. One limitation of this study is its focus on individuals of European ancestry. Although we also
performed gene-level tests in the other sub-populations of UKBB we were underpowered to detect associations as they represent a small proportion of the cohort. The only significant association in non-European ancestry populations was between MYH6 damaging missense variants and HbA1c levels in the East Asian ancestry population. Given the small number of carriers and known function of MYH6 in cardiac function\(^2,20\), this may be a false positive and replication is required. An important area for further study will be assessing the impact of rare variants on T2D risk in larger multi-ethnic cohorts.

Our finding that rare variants in the MODY genes, GCK, HNF1A and PDX1 associate with common T2D are consistent with results reported by Bonneford et al.\(^26\). Bonneford et al. defined pathogenic variants in MODY genes according to the American Society of Human Genetics criteria\(^27\). In contrast, we used bioinformatic methods to select variants likely to have functional impact in all genes in the genome. This allowed us to identify associations for genes not previously implicated in diabetes such as GIGYF1. The fact that we detected associations for MODY genes strongly suggests that we are enriching for variants with functional impact. However, for GCK and HNF1A, adding known pathogenic variants did increase the significance of detected associations.

Test statistic inflation can be a challenge when testing rare variants as statistical assumptions break down when the number of carriers expected to have the disease of interest is low\(^2,28\). To avoid false positives in our analysis of diabetes, we initially examined associations with glucose and HbA1c because quantitative traits are less susceptible to inflation. All the variant sets that associated with T2D also affected HbA1c and/or glucose levels giving us confidence in these associations. A targeted analysis of the genes encoding T2D drug targets revealed HbA1c and/or T2D associations for variants in several of these genes. The lack of association for variants in some of these drug target genes may in part be due to a lack of statistical power for genes with small numbers of rare variant carriers. However, for some of these genes such as SLC5A2 (encoding SGLT2) we do not detect associations with diabetes traits despite having a reasonable number of variant carriers.

We uncovered novel associations with T2D and biomarkers of glycemic control for aggregated variants in GIGYF1, TNRC6B and PFAS and attempted replication of these associations in exome-sequenced individuals from GHS. The association of GIGYF1 pLOF with T2D replicated in GHS but we did not replicate associations for TNRC6B and PFAS. There are differences between these two cohorts; UKBB is a population-based cohort with T2D diagnoses obtained from inpatient records while GHS is a health system-based cohort and includes both inpatient and outpatient diagnoses. There is a larger effect size for GIGYF1 pLOF in UKBB compared to GHS which may be due to these differences in ascertainment. Differences in the definition of the variant sets tested or the frequency of the relevant variants may have contributed to the failure to replicate the TNRC6B and PFAS associations. Alternatively, this may suggest that the TNRC6B and PFAS associations are false positives.

We focused our analysis on understanding the consequences of GIGYF1 pLOF as it strongly associated with glucose, HbA1c and T2D and the T2D association replicated in GHS. GIGYF1 encodes a protein that was initially identified for its binding to the adapter protein GRB10 (GRB10 interacting YIF protein 1) which negatively regulates both the insulin and IGF-1 receptors\(^29,30\). Transfection of cells with GRB10-binding fragments of GIGYF1 led to greater activation of both the insulin and IGF-1 receptors\(^31\). This supports a hypothesis whereby GIGYF1 enhances insulin signaling by reducing the negative regulation of the insulin receptor by GRB10. When GIGYF1 is reduced, as is the case in individuals carrying pLOF variants, GRB10 presumably inhibits insulin signaling to a greater degree thereby reducing the action of insulin in its target tissues and leading to increased risk of T2D. However, the exact mechanistic details of these interactions remain to be determined. GRB10 variants have also been reported to associate with T2D and glycemic traits although interpretation of these results is complicated by imprinting\(^31,32\).

GIGYF1 is broadly expressed with high levels observed in endocrine tissues, pancreas and brain\(^33\). GIGYF1 and the related protein GIGYF2 have also been implicated in translational repression\(^34\) and translation-coupled mRNA decay\(^35\) suggesting biological roles beyond regulation of insulin and IGF-1 receptor signaling.

The WAS of GIGYF1 pLOF revealed a strong association with decreased cholesterol levels reflecting altered energy homeostasis in carriers. An inverse relationship between glucose and cholesterol levels has been observed for variants in other genes\(^36\). We also observed several associations that could reflect complications of diabetes in GIGYF1 pLOF carriers including increased cystatin c levels and increased diagnosis of urinary disorders, suggesting renal complications, as well as syncope and collapse which may be a side-effect of hyperglycemia and/or hypoglycemia in diabetics. Other associations may reflect poor health in carriers including decreased grip strength and decreased peak expiratory flow.

GIGYF1 pLOF associated with a 4.5-fold increased risk of hypothyroidism and GIGYF1 is highly expressed in thyroid\(^36,38\) consistent with a biological function in this tissue. IGF-1 and insulin have been implicated in the proliferation of thyroid cells which may, in part, explain the association with thyroid dysfunction\(^39-39\). An alternative possibility is that GIGYF1 contributes to thyroid function by affecting secretion of thyroid stimulating hormone in the anterior pituitary gland. Another explanation is that shared autoimmune mechanisms contribute to thyroid dysfunction and diabetes in pLOF carriers and that some of the carriers diagnosed with T2D have features of latent autoimmune diabetes in adults\(^40\). Damaging variants in GIGYF1 have recently been implicated in conferring risk for developmental delay and autism spectrum disorders\(^41\). Consistent with this, we see an association of GIGYF1 pLOF with increased time to complete a cognitive test. It may be that metabolic aberrations in carriers affect cognitive performance, that brain development is altered due to perturbation of IGF-1 signaling, or that other functions of GIGYF1 such as regulation of mRNA expression and decay are responsible for cognitive phenotypes.

In addition to replicating the association of GIGYF1 pLOF with T2D in an independent cohort we also used common genetic variants to further investigate the role of the GIGYF1 locus in diabetes. A non-coding variant at the GIGYF1 locus associated with glucose levels and T2D, and this replicated in independent datasets. This variant associated with increased GIGYF1 expression but a lower risk of T2D. This direction of effect is consistent
with what we see for the pLOF variants—reduced levels of GIGYF1 increases diabetes risk but increased levels of GIGYF1 are protective.

We observed an intersection of rare and common variant associations at GIGYF1 as well as at MODY genes such as GCK, HNF1A and HNF4A. However, in general, our gene-level analysis of rare variants did not identify many additional causal genes at GWAS loci; out of 558 variants associated with T2D just nine had rare variant associations at a nearby gene.

We assessed the impact of pLOF and predicted damaging missense variants on glycemic traits and uncovered a hitherto unappreciated role for GIGYF1 in regulating blood sugar and protecting from T2D. By highlighting the importance of GIGYF1 and GRB adapter proteins in modulating insulin signaling this finding may lead to new therapeutic approaches for the treatment of diabetes. Discoveries such as this are only possible by combining health-related data with the sequencing of rare variants on a biobank scale.

Methods

The UK Biobank resource and data access. The UK Biobank (UKBB) recruited ~ 500,000 participants in England, Wales, and Scotland between 2006 and 2010. Written informed consent was obtained from all participants. Phenotypic data available includes age, sex, biomarker data and self-reported diseases collected at the time of baseline assessment as well as disease diagnoses from inpatient hospital stays, the cancer registry and death records obtained through the NHS. Approximately half of the participants also have diagnoses from primary care available. Array genotypes are available for nearly all participants and exome sequencing data is available for 454,787 participants. The data used in this study were obtained from the UKBB through application 26041.

Population definition and PC calculation for subjects with exome data. Subject quality control was performed by Regeneron Genetics Center (RGC) and removed subjects with evidence of contamination, unresolved duplications, sex discrepancies and discordance between exome sequencing and genotyping data. Genetic relationships between participants were determined by RGC using the PRIMUS program. For the unrelated subset all first- and second-degrees relatives and some third-degree relatives were excluded.

Populations were defined through a combination of self-reported ethnicity and genetic principal components. We selected the unrelated individuals who identify as White (European), Black (African), Asian (South Asian) or Chinese (East Asian) (Field 21000) and ran an initial principal component analysis (PCA) on high quality common variants using eigenstrat. SNPs were filtered for missingness across individuals < 2%, MAF > 1%, regions of known long range LD, and pruned to independent markers with pairwise LD < 0.1. We then projected the principal components (PCs) onto related individuals and removed all individuals ± 3 standard deviations from the mean of PCs 1–6. A final PC estimation was performed in eigenstrat using unrelated subjects. We then projected related individuals onto the PCs.

Exome sequencing and variant calling. DNA was extracted from whole blood and was sequenced by the RGC as described elsewhere. Briefly, the xGen exome capture was used and reads were sequenced using the Illumina NovaSeq 6000 platform. Reads were aligned to the GRCh38 reference genome using BWA-mem. Duplicate reads were identified and excluded using the Picard MarkDuplicates tool (Broad Institute). Variant calling of SNVs and indels was done using the Wecall variant caller (Genomics Plc.) to produce a GVCF for each subject. GVCFs were combined using the GLnexus joint calling tool. Post-variant calling filtering was applied using the Goldilocks pipeline. Variants were annotated using the Ensembl Variant Effect Predictor which includes a LOFTEE plug-in to identify high confidence (HC) pLOF variants by the Neale lab, using PHESANT to transform values to normal distribution. For sex-stratified analyses, glucose and HbA1c values were inverse rank normalized separately in males and females.

For disease diagnoses, ICD10 codes were obtained from inpatient hospital diagnoses (Field 41270), causes of death (Field 40001 and 40002) and the cancer registry (Field 40,006) from UKBB. Diagnoses also included additional hospital episode statistics (HESIN) and death registry data made available by UKBB in July 2020. T2D was defined as ICD10 E11. For the purposes of excluding diagnosed diabetics from the glucose and HbA1c analysis we defined diabetes as ICD10 codes E10–E14 which includes both T1D and T2D diagnoses.

For genome-wide analyses, a selection of quantitative traits was obtained from other fields, encompassing anthropometric measurements, blood counts, as well as blood and urine biochemistry. Beyond these measurements, we selected additional quantitative traits found to be heritable (p ≤ 0.05 with a confidence level flagged as “medium” or “high”) by the Neale lab, using PHEASANT to transform values to quantitative traits when necessary as they describe. These included the results of cognitive tests. All quantitative traits were inverse rank normalized using the RNome R package. For burden testing, we required at least ten carriers to have measurements. We also tested associations with ICD10-coded diagnoses (using three character codes) that had more than 500 cases in the European ancestry participants with exome data and at least one expected case carrier based on variant frequency and disease prevalence.

Glucose and HbA1c values were also extracted from primary care data available for about half of the cohort using the following read codes. Glucose: read 2 codes 44U, 44g, 44g1, 44TJ, 44F, 44TK, 44F1, 44g0, 44f0 and read 3 codes XM0ly, X772z, XE2mq; HbA1c: read 2 codes 42WS, 44TB, 66Ae0, 44TC, 42W4 and read 3 codes...
The effect allele in these datasets was the alternate allele “C”. For consistency with T2D diagnosis using summary statistics from FinnGen release 3 for the phenotype “E4_DM2” (https://www.finnegen.fi/en/access_results). A subset of participants sequenced as part of the GHS-Regen-Eron Genetics Center DiscovEHR partnership were included in this study. T2D status was defined based on meeting at least one of the following criteria: (1) clinical encounters due to or problem-lists diagnosis code for type 2 diabetes (ICD-10 code E11), or (2) HbA1c greater than 6.5%, or (3) use of diabetic oral hypoglycemic medicine. Controls were participants who did not meet any of the criteria for case definition. Individuals were excluded from the analysis if they had clinical encounters due to or problem-lists diagnosis code for type 1 diabetes (ICD-10 code E10), or if they were treated with insulin but not with oral hypoglycemic medicines.

Exome-sequencing, variant calling, quality control and gene-based tests were performed as previously described. Variant sets tested were pLOF variants (GIGYF1 and TNRC6B) or pLOF plus missense variants for T2D drug targets were defined according to Flannick et al.

Manhattan plots were created using the R Package CMplot (https://github.com/YinLiLin/R-CMplot).

The expression of GIGYF1 in various tissues was assessed using the GTEX portal (accessed 08/04/2020) and Human Protein Atlas (http://www.proteinatlas.org). eQTL data for rs221783 was obtained from GTEx v8. For each tissue of interest, the best eQTL for GIGYF1 was identified (GTEx v8 “eGene”). R² for rs221783 and the best GIGYF1 eQTL was calculated as described above.

Replication analysis in GHS. The GHS MyCode Community Health Initiative study is a health system-based cohort and has been described previously. A subset of participants sequenced as part of the GHS-Regen-Eron Genetics Center DiscovEHR partnership were included in this study. T2D status was defined based on meeting at least one of the following criteria: (1) clinical encounters due to or problem-lists diagnosis code for type 2 diabetes (ICD-10 code E11), or (2) HbA1c greater than 6.5%, or (3) use of diabetic oral hypoglycemic medicine. Controls were participants who did not meet any of the criteria for case definition. Individuals were excluded from the analysis if they had clinical encounters due to or problem-lists diagnosis code for type 1 diabetes (ICD-10 code E10), or if they were treated with insulin but not with oral hypoglycemic medicines.

Exome-sequencing, variant calling, quality control and gene-based tests were performed as previously described. Variant sets tested were pLOF variants (GIGYF1 and TNRC6B) or pLOF plus missense variants.
predicted to be deleterious by 5/5 algorithms (PFAS) with MAF < 1%. The following variants were classified as pLOF variants: frameshift-causing indels, variants affecting splice acceptor and donor sites, variants leading to stop gain, stop loss and start loss. The five missense deleterious algorithms used were SIFT\(^5\), PolyPhen2 (HDIV), PolyPhen2 (HVAR)\(^6\), LRT\(^7\), and MutationTaster\(^8\). Association testing was performed in the European ancestry population using the Firth logistic regression test implemented in REGENIE\(^9\).

**Identification of potential causal genes at GWAS loci.** For 558 variants identified as associating with T2D\(^9\) we mapped the two closest protein coding genes using bedtools. This resulted in 1118 genes for which we had tested 2071 variant sets (pLOF and/or damaging missense) in our primary analysis. Genes with \(p < 2.41 \times 10^{-5}\) (correcting for 2071 variant sets tested) for HbA1c or T2D were considered significant.

**Ethics statement.** The UK Biobank study was approved by the National Health Service National Research Ethics Service and all participants provided written informed consent to participate in the study. The UK Biobank resource is an approved Research Tissue Bank and is registered with the Human Tissue Authority, which means that researchers who wish to use it do not need to seek separate ethics approval (unless re-contact of participants is required). Information about ethics oversight in the UK Biobank can be found at [https://www.ukbiobank.ac.uk/ethics/](https://www.ukbiobank.ac.uk/ethics/). This research has been conducted using the UK Biobank resource, application 26041. Research in GHS was approved by the GHS IRB, approval number 2006-0258. Written informed consent was obtained from all participants in GHS.

**Data availability**

For the primary analysis in UKBB, all phenotypic data and array genotypes used in this study are accessible through application to UKBB. Currently, exome sequencing data for ~200,000 participants is available\(^10\); the remainder of the exome data used is scheduled for release by UKBB in 2021. Full summary statistics for gene-level tests performed in UKBB will be made available upon publication. For the replication analysis in Geisinger Health System, the full summary statistics for all replication analyses performed are supplied in Supplemental Table 24. Summary statistics for Biobank Japan are available at [http://jenger.riken.jp/en/result](http://jenger.riken.jp/en/result) and summary statistics for FinnGen are available at [https://www.finngen.fi/en/access_results](https://www.finngen.fi/en/access_results).

Received: 22 April 2021; Accepted: 15 September 2021
Published online: 03 November 2021

**References**

1. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* 15, e1008489. [https://doi.org/10.1371/journal.pgen.1008489](https://doi.org/10.1371/journal.pgen.1008489) (2019).

2. Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. *Nat. Genet.* 47, 856–860. [https://doi.org/10.1038/ng.3314](https://doi.org/10.1038/ng.3314) (2015).

3. Nguyen, P. A., Born, D. A., Deaton, A. M., Nioi, P. & Ward, L. D. Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. *Nat. Commun.* 10, 1579. [https://doi.org/10.1038/s41467-019-09407-3](https://doi.org/10.1038/s41467-019-09407-3) (2019).

4. Cirulli, E. T. et al. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat Commun.* 11, 542. [https://doi.org/10.1038/s41467-020-14288-y](https://doi.org/10.1038/s41467-020-14288-y) (2020).

5. Flannick, J. et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* 570, 71–76. [https://doi.org/10.1038/s41586-019-1231-2](https://doi.org/10.1038/s41586-019-1231-2) (2019).

6. Moutsianas, L. et al. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet.* 11, e1005165. [https://doi.org/10.1371/journal.pgen.1005165](https://doi.org/10.1371/journal.pgen.1005165) (2015).

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

8. Diogo, D. et al. Phenome-wide association studies across large population cohorts support drug target validation. *Nat. Commun.* 9, 4285. [https://doi.org/10.1038/s41467-018-06540-3](https://doi.org/10.1038/s41467-018-06540-3) (2018).

9. Vukovic, M. et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* 52, 680–691. [https://doi.org/10.1038/s41467-020-0637-y](https://doi.org/10.1038/s41467-020-0637-y) (2020).

10. Xue, A. et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* 9, 2941. [https://doi.org/10.1038/s41467-018-04951-w](https://doi.org/10.1038/s41467-018-04951-w) (2018).

11. Huang, J., Ellinghaus, D., Franke, A., Howie, B. & Li, Y. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *Eur. J. Hum. Genet.* 20, 801–805. [https://doi.org/10.1038/ehgj.2012.3](https://doi.org/10.1038/ehgj.2012.3) (2012).

12. Scott, R. A. et al. An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes* 66, 2888–2902. [https://doi.org/10.2337/db16-1253](https://doi.org/10.2337/db16-1253) (2017).

13. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. [https://doi.org/10.1038/s41586-020-2308-7](https://doi.org/10.1038/s41586-020-2308-7) (2020).

14. 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. [https://doi.org/10.1056/NEJMr002168](https://doi.org/10.1056/NEJMr002168) (2001).

15. Wessel, J. et al. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat. Commun.* 6, 5897. [https://doi.org/10.1038/ncomms6897](https://doi.org/10.1038/ncomms6897) (2015).

16. Zhou, W. et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat. Genet.* 52, 634–639. [https://doi.org/10.1038/s41588-020-0621-6](https://doi.org/10.1038/s41588-020-0621-6) (2020).

17. Landrum, M. J. et al. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46, D1062–D1067. [https://doi.org/10.1093/nar/gkx1133](https://doi.org/10.1093/nar/gkx1133) (2018).

18. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* 550, 204–213. [https://doi.org/10.1038/nature24277](https://doi.org/10.1038/nature24277) (2017).

19. Kanai, M. et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* 50, 390–400. [https://doi.org/10.1038/s41588-018-0047-6](https://doi.org/10.1038/s41588-018-0047-6) (2018).
20. Saevarsdottir, S. et al. FLT3 stop mutation increases FLT3 ligand level and risk of autoimmune thyroid disease. *Nature* **584**, 619–623. https://doi.org/10.1038/s41586-020-2436-0 (2020).

21. Frognel, P. et al. Close linkage of glucokinaase locus on chromosome 7p to early-onset non-insulin-dependent diabetes mellitus. *Nature* **356**, 162–164. https://doi.org/10.1038/356162a0 (1992).

22. Ellard, S. Hepatocyte nuclear factor 1 alpha (HNF-1 alpha) mutations in maturity-onset diabetes of the young. *Hum. Mutat.* **16**, 377–385. https://doi.org/10.1002/1098-1004(200011)16:3<377::AID-HUMU16>3.0.CO;2-2 (2000).

23. Stoffers, D. A., Ferrer, J., Clarke, W. L. & Habener, J. F. Early-onset type-II diabetes mellitus (MODY4) linked to IPF1. *Nat. Genet.* **17**, 138–139. https://doi.org/10.1038/ng.1097-138 (1997).

24. Holm, H. et al. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat. Genet.* **43**, 316–320. https://doi.org/10.1038/ng.781 (2011).

25. Ching, Y. H. et al. Mutation in myosin heavy chain 6 causes atrial septal defect. *Nat. Genet.* **35**, 86–90. https://doi.org/10.1038/ng.1526 (2005).

26. Bonnefond, A. et al. Pathogenic variants in actionable MODY genes are associated with type 2 diabetes. *Nat. Metab.* **2**, 1126–1134. https://doi.org/10.1038/s42255-020-00294-3 (2020).

27. 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. https://doi.org/10.1038/gim.2015.30 (2015).

28. Churchhouse, C. Details and Considerations of the UK Biobank GWAS. *Neale lab*. http://www.nealelab.is/blog/2017/9/11/detali s-and-considerations-of-the-uk-biobank-gwas (2017).
61. Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods 7, 575–576. https://doi.org/10.1038/nmeth0810-575 (2010).
62. Mbathou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. Nat. Genet. 53, 1097–1103. https://doi.org/10.1038/s41588-021-00870-7 (2021).
63. Szuštakowski, J. D. et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Nat. Genet. 53, 942–948. https://doi.org/10.1038/s41588-021-00885-0 (2021).

Acknowledgements
This research has been conducted using the UK Biobank Resource (Project 26041). We would like to thank the participants and researchers of UK Biobank for creating an open-access resource. We thank the UK Biobank Exome Sequencing Consortium and UK Biobank for facilitating exome sequencing of participants. We also thank the participants of the GHS MyCode initiative as well as participants and investigators of the FinnGen study and Biobank Japan. We thank Mark McCarthy and Anna Glynn for comments on the manuscript. Data management and analytics were performed using the REVEAL/SciDB translational analytics platform from Paradigm4.

Author contributions
A.D., L.W., M.P., A.F.C. and P.N. performed computational analyses; P.A., L.L. and A.B. performed replication analysis in GHS; A.D. wrote the manuscript. All authors interpreted results and edited the manuscript.

Funding
This study was funded by Alnylam Pharmaceuticals and Regeneron Pharmaceuticals.

Competing interests
A.D., L.W., M.P., A.F.C., L.B., G.H. and P.N. are employees and stockholders of Alnylam Pharmaceuticals. P.A., L.L. and A.B. are employees and stockholders of Regeneron Pharmaceuticals.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-99091-5.

Correspondence and requests for materials should be addressed to A.M.D.

Reprints and permissions information is available at 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 licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence 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 licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021

Regeneron Genetics Center
RGC Management and Leadership Team

Goncalo Abecasis²Aris Baras²Michael Cantor²Giovanni Coppola²Aris Economides²Luca A. Lotta²
John D. Overton²Jeffrey G. Reid²Alan Shuldiner²Katia Karalis²Andrew Deubler²Katherine Siminovitch²

Sequencing and Lab Operations

Christina Beechert²Caitlin Forsythe²Erin D. Fuller²Zhenhua Gu²Michael Lattari²Alexander Lopez²
John D. Overton²Thomas D. Schleicher²Maria Sotiropoulos Padilla²Louis Widom²Sarah E. Wolf²
Manasi Pradhan²Kia Manoochehri²Ricardo H. Ulloa²

Genome Informatics

Xiaodong Bai²Suganthi Balasubramanian²Andrew Blumenfeld²Boris Boutkov²Gisu Eom²
Lucas Habegger²Alicia Hawes²Shareef Khalidi²Olga Krasheninina²Rouel Lanche²
Adam J. Mansfield²Evan K. Maxwell²Mona Nafde²Sean O’Keeffe²Max Orelus²Razvan Panea²
Tommy Polanco²Ayesha Rasool²Jeffrey G. Reid²William Salerno²Jeffrey C. Staples²
Clinical Informatics

Michael Cantor2Dadong Li2

Translational and Analytical Genetics

Adam Locke2Niek Verweij2Jonas Nielsen3Jonas Bovijn2Tanima De3Mary Haas2Parsa Akbari2
Olukayode Sosina2

Research Program Management

Marcus B. Jones2Jason Mighty2Michelle G. LeBlanc2Lyndon J. Mitnaul2

DiscovEHR Collaboration

Regeneron Personnel

Goncalo Abecasis2Aris Baras2Michael Cantor2Giovanni Coppola2Aris Economides2Luca A. Lotta2
John D. Overton2Jeffrey G. Reid2Alan Shuldiner2Andrew Deubler2Katia Karalis2
Christina Beechert2Caitlin Forsythe2Erin D. Fuller2Zhenhua Gu2Michael Lattari2
Alexander Lopez2Thomas D. Schleicher2Maria Sotiropoulos Padilla2Karina Toledo2
Louis Widom2Sarah E. Wolf2Manasi Pradhan2Kia Manoochehr2Ricardo H. Ulloa2
Xiaodong Bai2Suganthi Balasubramanian2Leland Barnard2Andrew Blumenfeld2Gisu Eom2
Lukas Habegger2Alicia Hawes2Shareef Khalid2Evan K. Maxwell2William Salerno2Jeffrey C. Staples2
Ashish Yadav2Dadong Li2Marcus B. Jones2Lyndon J. Mitnaul2Jason Mighty2Andrew Deubler2
Katia Karalis2Katherine Siminovitch2

Geisinger Personnel

Lance J. Adams3Jackie Blank3Dale Bodian3Derek Boris3Adam Buchanan3David J. Carey3
Ryan D. Colonie3E. Daniel Davis3Dustin N. Hartzel3Melissa Kelly3H. Lester Kirchner3Joseph B. Leader3
David H. Ledbetter3J. Neil Manus3Christa L. Martin3Raghu P. Metpally3Michelle Meyer3Tooraj Mirshahi3
Matthew Oetjens3Thomas Nate Person3Christopher Still3Natasha Strande3Amy Sturm3Jen Wagner3
Marc Williams3