medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Gene-level analysis of rare variants in 363,977 whole exome sequences identifies an association of *GIGYF1* loss of function with type 2 diabetes

Aimee M. Deaton[1], Margaret M. Parker[1], Lucas D. Ward[1], Alexander O. Flynn-Carroll[1], Lucas BonDurant[1], Gregory Hinkle[1], Parsa Akbari[2], Luca A. Lotta[2], Regeneron Genetics Center[2*], DiscovEHR Collaboration[3], Aris Baras[2] and Paul Nioi[1]

[1]Alnylam Pharmaceuticals, Cambridge, MA

[2]Regeneron Genetics Center, Regeneron Pharmaceuticals, Tarrytown, NY

[3]Geisinger Health System, Danville, PA

* For full author and contribution list see Supplementary Information.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

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 363,977 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.

Author Summary

Genetic studies focused on high impact variants in protein-coding regions of the genome can provide valuable insight into the biology of human disease. As these variants tend to be rare, studying them requires large cohort sizes and methods to aggregate variants that are likely to have a similar biological impact. We studied how rare genetic variants contribute to type 2 diabetes (T2D) using sequencing data from 363,977 participants in the UK Biobank, employing methods to aggregate variants at the level of individual genes. As well as identifying genes known to be involved in inherited forms of diabetes, we uncovered a novel association for *GIGYF1*. *GIGYF1* loss of function associated with increased risk of T2D and increased levels of the diabetes biomarkers glucose and HbA1c. This association was also seen in an independent dataset. *GIGYF1* encodes a protein that binds a negative regulator of the insulin receptor that has not been well-characterized in the literature. By highlighting the importance of GIGYF1 in modulating insulin signaling these results may lead to new therapeutic approaches for diabetes as well as a new appreciation for *GIGYF1* loss of function as a genetic risk factor for T2D.

Introduction

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 [1-3]. 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 [4, 5]. However, interrogation of the effects of such variants is hampered by the rarity of these variants and the cohort sizes needed to identify associations [6]. Exome or whole-genome sequencing of large biobanks coupled with gene-level aggregation of rare high impact variants can help to circumvent these challenges [4]. Biobanks offer a considerable advantage over case-control cohorts as they contain richer phenotyping data which often includes

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

biomarker measurements as well as disease diagnoses. This allows a more complete understanding of the biological consequences of damaging variants in particular genes [7, 8].

Diabetes is a disease that has been extensively studied in traditional array-based GWAS with hundreds of associations identified to date [9-12]. Although these studies have given insight into some of the biological mechanisms contributing to diabetes, 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 type 2 diabetes (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 [5].

Using 363,977 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 LOFTEE [13] 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 White 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 10 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 ($p \leq 7.82 \times 10^{-7}$), four genes significantly associated with glucose levels: *GCK* pLOF ($p = 1.56 \times 10^{-9}$, 1.24 mmol/L increase), *GCK* damaging missense ($p = 6.15 \times 10^{-11}$, 0.61 mmol/L increase), *GIGYF1* pLOF ($p = 4.42 \times 10^{-12}$, 0.77 mmol/L increase) and *G6PC2* damaging missense variants ($p = 4.62 \times 10^{-83}$, 0.33 mmol/L decrease) (Figure 1, Table 1). The same variant sets also associated with HbA1c levels along with 27 additional sets including *HNF1A* pLOF ($p = 2.14 \times 10^{-7}$, 4.01 mmol/mol increase), *TNRC6B* pLOF ($p = 2.36 \times 10^{-7}$, 3.94 mmol/mol increase) and *PDX1* damaging missense variants ($p = 2.54 \times 10^{-7}$, 0.41 mmol/mol increase) (Figure 1, Table 1).

We then tested aggregated pLOF and damaging missense variants for association with T2D diagnosis (n=24,695 cases). Using a p-value threshold adjusted for the number of variant sets tested ($p \leq 1.46 \times 10^{-6}$), 6 variant sets significantly associated with T2D; pLOF variants in *GIGYF1*, *GCK*, *HNF1A* and *TNRC6B*

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

and damaging missense variants in *GCK* and *PAM* (Figure 2, Table 2). As the time of available follow-up differs between England, Scotland, and Wales, we controlled for country of recruitment in the regression (see Methods). 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 2).

Table 1: Gene-level associations with glucose and HbA1c levels

Association of pLOF or damaging missense variants (CADD score ≥ 25) aggregated per gene with glucose and HbA1c levels. The effect is shown in standard deviations (SD) of transformed values as well as in International Federation of Clinical Chemistry (IFCC) units. CI; confidence interval.

gene	Variant set	title	pvalue	Effect (SD)	effect (units)	95% CI -	95% CI +	units	n carrier measured
GCK	pLOF	glucose	1.56×10^{-9}	1.00	1.24	0.84	1.65	mmol/L	35
GIGYF1	pLOF	glucose	4.42×10^{-12}	0.62	0.77	0.55	0.98	mmol/L	121
GCK	missense CADD>25	glucose	6.15×10^{-11}	0.49	0.61	0.42	0.79	mmol/L	173
G6PC2	missense CADD>25	glucose	4.62×10^{-83}	-0.27	-0.33	-0.36	-0.3	mmol/L	5128
GCK	pLOF	HbA1c	2.64×10^{-17}	1.29	8.75	6.73	10.78	mmol/mol	38
GIGYF1	pLOF	HbA1c	1.28×10^{-14}	0.64	4.33	3.23	5.43	mmol/mol	129
HNF1A	pLOF	HbA1c	2.14×10^{-7}	0.59	4.01	2.50	5.53	mmol/mol	68
TNRC6B	pLOF	HbA1c	2.36×10^{-7}	0.58	3.94	2.45	5.43	mmol/mol	70
RHAG	pLOF	HbA1c	3.31×10^{-34}	-0.86	-5.81	-6.75	-4.88	mmol/mol	179
EPB41	pLOF	HbA1c	3.14×10^{-17}	-0.53	-3.58	-4.41	-2.75	mmol/mol	226
PTPRH	pLOF	HbA1c	4.39×10^{-10}	0.11	0.74	0.51	0.97	mmol/mol	2924
APOB	pLOF	HbA1c	6.94×10^{-8}	0.23	1.57	1.00	2.15	mmol/mol	478
PLD1	pLOF	HbA1c	2.99×10^{-7}	0.23	1.56	0.96	2.16	mmol/mol	438
EPB42	pLOF	HbA1c	6.11×10^{-7}	-0.31	-2.08	-2.90	-1.26	mmol/mol	234
GCK	missense CADD ≥ 25	HbA1c	1.86×10^{-17}	0.56	3.83	2.94	4.71	mmol/mol	201
G6PC2	missense CADD ≥ 25	HbA1c	6.71×10^{-45}	-0.18	-1.21	-1.38	-1.04	mmol/mol	5574
PFAS	missense CADD ≥ 25	HbA1c	2.09×10^{-8}	-0.05	-0.32	-0.44	-0.21	mmol/mol	12621
PDX1	missense CADD ≥ 25	HbA1c	2.54×10^{-7}	0.06	0.41	0.25	0.56	mmol/mol	6694
PIEZO1	missense CADD ≥ 25	HbA1c	1.0×10^{-132}	-0.15	-1.00	-1.07	-0.92	mmol/mol	26726
AMPD3	missense CADD ≥ 25	HbA1c	7.28×10^{-34}	0.13	0.86	0.72	1.00	mmol/mol	8258
PFKM	missense CADD ≥ 25	HbA1c	2.16×10^{-28}	-0.28	-1.92	-2.26	-1.58	mmol/mol	1353
ANK1	missense CADD ≥ 25	HbA1c	3.12×10^{-19}	-0.13	-0.87	-1.06	-0.68	mmol/mol	4342
PFKL	missense CADD ≥ 25	HbA1c	2.69×10^{-14}	0.10	0.68	0.50	0.85	mmol/mol	5245
AXL	missense CADD ≥ 25	HbA1c	4.11×10^{-12}	-0.08	-0.54	-0.69	-0.39	mmol/mol	6827
LCAT	missense CADD ≥ 25	HbA1c	1.29×10^{-11}	-0.27	-1.82	-2.34	-1.29	mmol/mol	565

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

PLA2G12A	missense CADD ≥ 25	HbA1c	5.74×10^{-11}	-0.08	-0.51	-0.67	-0.36	mmol/mol	6761
PLD1	missense CADD ≥ 25	HbA1c	1.51×10^{-10}	0.08	0.57	0.40	0.75	mmol/mol	5180
APEH	missense CADD ≥ 25	HbA1c	1.86×10^{-10}	0.29	1.96	1.36	2.57	mmol/mol	429
TNFRSF13B	missense CADD ≥ 25	HbA1c	4.63×10^{-9}	-0.08	-0.51	-0.68	-0.34	mmol/mol	5470
BLVRB	missense CADD ≥ 25	HbA1c	2.53×10^{-8}	-0.07	-0.48	-0.65	-0.31	mmol/mol	5533
TMC8	missense CADD ≥ 25	HbA1c	3.01×10^{-8}	0.11	0.77	0.50	1.05	mmol/mol	2095
HK1	missense CADD ≥ 25	HbA1c	1.08×10^{-7}	-0.16	-1.08	-1.48	-0.68	mmol/mol	988
SLC4A1	missense CADD ≥ 25	HbA1c	1.76×10^{-7}	-0.15	-1.04	-1.43	-0.65	mmol/mol	1025
TFR2	missense CADD ≥ 25	HbA1c	4.24×10^{-7}	-0.12	-0.84	-1.16	-0.51	mmol/mol	1491
CARHSP1	missense CADD ≥ 25	HbA1c	6.78×10^{-7}	-0.13	-0.86	-1.20	-0.52	mmol/mol	1360

Table 2: Gene-level associations with T2D diagnosis

Association of pLOF or damaging missense variants (CADD score ≥ 25) aggregated per gene with T2D diagnosis. OR; odds ratio, CI; confidence interval.

gene	Variant set	title	pvalue	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	24695	40	19	2.71
GIGYF1	pLOF	T2D	6.14×10^{-11}	4.15	2.71	6.37	24695	131	29	8.89
HNF1A	pLOF	T2D	1.23×10^{-9}	5.27	3.08	9	24695	73	20	4.95
TNRC6B	pLOF	T2D	2.00×10^{-7}	4.44	2.53	7.79	24695	71	17	4.82
PAM	missense CADD ≥ 25	T2D	2.26×10^{-12}	1.31	1.21	1.41	24695	9357	801	634.85
GCK	missense CADD ≥ 25	T2D	1.70×10^{-8}	2.96	2.03	4.32	24695	202	34	13.71

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 56Mb apart, strongly suggesting that these signals are independent; this independence was confirmed by conditional analysis (Supplementary Table 3). Two additional variant sets, *HNF1A* pLOF and *TNRC6B* pLOF, had genome-wide associations with both T2D diagnosis and HbA1c levels while *G6PC2* damaging missense 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 using a more permissive p-value threshold correcting for the number of variant

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

sets tested ($p \leq 0.0016$, 32 sets tested). Damaging missense variants in *PDX1* and *PFAS*, which had significant associations with HbA1c levels in our primary analysis, associated with T2D diagnosis using this threshold (Table 3 and Supplementary Table 4).

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 4) 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 5) and an additional five had highly significant associations with red blood cell traits in our data ($p \leq 7.82 \times 10^{-7}$; Supplementary Table 6).

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

Because obesity is linked to the development of T2D, we adjusted for body mass index (BMI) in the burden tests and found that the association of variants in these genes with diabetes-related traits remained significant (Supplementary Tables 7 and 8).

Associations for rare variants can be susceptible to confounders such as population stratification and sample relatedness leading to false positives. Therefore, 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 White 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 *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 9).

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 10). For *GIGYF1* pLOF, 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.

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 ($p \leq 0.0016$, 32 sets tested). Effect is shown in SD of transformed values or as an odds ratio (OR).

gene	Variant set	Pvalue glucose	Effect glucose	Pvalue HbA1c	Effect Hba1c	Pvalue T2D	OR T2D
GCK	pLOF	1.56×10^{-9}	0.999	2.64×10^{-17}	1.292	2.96×10^{-15}	14.16
HNF1A	pLOF	0.01	0.317	2.14×10^{-7}	0.592	1.23×10^{-9}	5.27
GIGYF1	pLOF	4.42×10^{-12}	0.616	1.28×10^{-14}	0.639	6.14×10^{-11}	4.15
GCK	missense CADD ≥ 25	6.15×10^{-11}	0.487	1.86×10^{-17}	0.565	1.70×10^{-8}	2.96

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

PAM	missense CADD ≥ 25	0.92	0.001	0.009	0.026	2.26×10^{-12}	1.31
TNRC6B	pLOF	4.01×10^{-5}	0.507	2.36×10^{-7}	0.582	2.00×10^{-7}	4.44
PDX1	missense CADD ≥ 25	0.02	0.029	2.54×10^{-7}	0.060	3.99×10^{-5}	1.21
PFAS	missense CADD ≥ 25	0.32	0.009	2.09×10^{-8}	-0.048	4.43×10^{-4}	0.88
G6PC2	missense CADD ≥ 25	4.62×10^{-83}	-0.266	6.71×10^{-45}	-0.179	0.97	1.00

GIGYF1 pLOF associations replicate using independent datasets

We sought to use independent measurements of glucose and HbA1c to verify the associations of interest seen in our primary analysis which used measurements taken as part of the UKBB assessment. To do this we extracted lab test values for glucose and HbA1c from primary care data, which is available for approximately half of the cohort, taking the mean measurement per individual. 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 ($p \leq 0.004$). This included the association of *GIGYF1* pLOF with glucose ($p=2.10 \times 10^{-6}$, effect = 0.65 SD) and HbA1c ($p=1.19 \times 10^{-5}$, effect = 0.74 SD) levels (Supplementary Figure 2 and Supplementary Table 11).

We then assessed whether rare variants in *GIGYF1* and the other novel genes associated with T2D replicated in an 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 also tested an expanded *PFAS* variant set (pLOF + deleterious missense) and did not detect an association with T2D (Supplementary Table 12). 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 $p = 1.56 \times 10^{-9}$) and *GIGYF1* pLOF and T2D ($p = 2.9 \times 10^{-5}$ without singletons versus $p = 6.14 \times 10^{-11}$) (Supplementary Table 13), 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 gnomAD. We found that for *GIGYF1* and *PFAS* the variants contributing most to the associations had good quality scores and depth and were present in the non-Finnish European population in gnomAD. In contrast, *TNRC6B* is a highly constrained gene and the most common pLOF variant is not present in gnomAD. It is possible pLOF variants for constrained genes may not result in true loss of function (see Supplementary Note and Supplementary Figure 4). This

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

observation along with the fact that the association of *TNRC6B* pLOF with T2D did not replicate in Geisinger Health System leads us to view this association with suspicion.

Replication of published gene-level associations with T2D and associations for T2D drug target genes

The association between predicted damaging variants in *PAM* and T2D diagnosis was previously reported in an exome-sequencing study performed by Flannick and colleagues [5]. 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 14). Combining *SLC30A8* pLOF and missense variants resulted in more significant associations with glucose ($p = 2.71 \times 10^{-6}$), HbA1c ($p = 8.64 \times 10^{-10}$) and T2D diagnosis ($p = 0.005$) (Supplementary Table 14). 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, ABCC8*). 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 15).

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. *GIGYF1* encodes a protein named for its binding to GRB10 (GRB10 interacting GYF protein 1), an adapter protein that has been shown to bind both the insulin and IGF-1 receptors. The association between *GIGYF1* pLOF and increased diabetes risk indicates that *GIGYF1* has a role in regulating insulin signaling and in protecting from diabetes. 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. Based on the number of tests performed, the threshold for significance was $p \leq 1.22 \times 10^{-4}$ (Figure 4).

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 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 16). *GIGYF1* pLOF also associated with decreased grip strength and decreased peak expiratory flow which may reflect changes in body size, muscle mass or general health in carriers [17, 18]. Notably, *GIGYF1* pLOF also associated with increased levels of the kidney injury biomarker

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

cystatin c (p= 6.65 x 10^{-6}, effect = 0.36 SD) and increased diagnosis of urinary system disorders (p = 7.32 x 10^{-5}, OR = 2.71) which might suggest renal complications of diabetes in carriers (Table 4 and Table 5).

After diabetes, the most significant disease association of *GIGYF1* pLOF was with increased risk of hypothyroidism (p = 1.25 x 10^{-9}, OR = 4.53). 21 out of the 131 *GIGYF1* pLOF carriers had a diagnosis of unspecified hypothyroidism and 7 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 significantly associated with increased risk of syncope and collapse (p = 1.92 x 10^{-6}, OR = 3.75), possibly reflecting complications of diabetes or thyroid disorders (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 x 10^{-12}). *GIGYF1* pLOF also associated with increased diagnosis of emphysema and anemia (Table 5).

Table 4: PheWAS of *GIGYF1* pLOF – quantitative traits

Showing significant results for burden tests on quantitative traits (p ≤ 1.22 x 10^{-4}). Effect is shown in standard deviations (SD) of transformed values. RH; right hand, LH; left hand.

gene	variant set	title	pvalue	Effect (SD)	95% CI -	95% CI +	n carrier measured
GIGYF1	pLOF	HbA1c	1.28 x 10^{-14}	0.64	0.48	0.80	129
GIGYF1	pLOF	cholesterol	2.44 x 10^{-12}	-0.61	-0.78	-0.44	128
GIGYF1	pLOF	glucose	4.42 x 10^{-12}	0.62	0.44	0.79	121
GIGYF1	pLOF	LDL cholesterol	2.40 x 10^{-10}	-0.56	-0.73	-0.38	128
GIGYF1	pLOF	apolipoprotein b	2.52 x 10^{-10}	-0.56	-0.73	-0.39	127
GIGYF1	pLOF	LH grip strength	5.11 x 10^{-10}	-0.37	-0.49	-0.25	131
GIGYF1	pLOF	RH grip strength	1.01 x 10^{-8}	-0.34	-0.46	-0.23	131
GIGYF1	pLOF	peak expiratory flow	5.73 x 10^{-8}	-0.41	-0.56	-0.26	114
GIGYF1	pLOF	cystatin c	6.65 x 10^{-6}	0.36	0.20	0.51	128
GIGYF1	pLOF	mean corpuscular hemoglobin	6.80 x 10^{-6}	-0.38	-0.55	-0.22	128
GIGYF1	pLOF	HDL cholesterol	1.53 x 10^{-5}	-0.35	-0.52	-0.19	121
GIGYF1	pLOF	time to complete round (cognitive test)	1.67 x 10^{-5}	0.35	0.19	0.51	129
GIGYF1	pLOF	waist circumference	3.98 x 10^{-5}	0.32	0.16	0.47	130
GIGYF1	pLOF	total protein	6.45 x 10^{-5}	-0.36	-0.53	-0.18	121
GIGYF1	pLOF	apolipoprotein a	6.88 x 10^{-5}	-0.33	-0.49	-0.17	121

Table 5: PheWAS of *GIGYF1* pLOF – ICD10-coded diagnoses

Showing significant results for burden tests on ICD10 coded diagnoses with ≥ 500 cases and ≥ 1 expected case carrier (p ≤ 1.22 x 10^{-4}). OR; odds ratio.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

gene	Variant set	title	pvalue	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	24695	29	8.89
GIGYF1	pLOF	E03 other hypothyroidism	1.25×10^{-9}	4.53	2.78	7.38	19417	21	6.99
GIGYF1	pLOF	R55 syncope and collapse	1.90×10^{-6}	3.75	2.18	6.47	12706	15	4.57
GIGYF1	pLOF	D50 iron deficiency anemia	8.52×10^{-6}	3.56	2.04	6.23	12886	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	24581	19	8.85

Common variants at *GIGYF1* associate with glucose, T2D and *GIGYF1* expression

Replication is a challenge for rare variant association studies. Despite the rarity of *GIGYF1* pLOF variants, we replicated the T2D association in an independent cohort. In addition, 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 White individuals with measurements available. We found a cluster of variants in a linkage disequilibrium block covering *GIGYF1* and *EPO* significantly associating with glucose levels (Figure 4). This signal is represented by rs221783, an intergenic variant whose minor T allele associated with decreased glucose ($p = 1.8 \times 10^{-11}$, effect = -0.03 SD,) and HbA1c ($p = 3.6 \times 10^{-7}$, effect = -0.02 SD,) levels as well as increased cholesterol ($p = 7.0 \times 10^{-12}$, effect = 0.03 SD,). This variant also associated with a decreased risk of T2D (p = 0.005, OR = 0.96) and hypothyroidism ($p = 6.95 \times 10^{-7}$, OR=0.92) (Table 6). rs221783 is the best eQTL ($R^2 > 0.8$) for *GIGYF1* in several tissues including pancreas, adipose and thyroid [19] (Supplementary Table 17). In all tissues, the T allele associating with decreased glucose and decreased T2D risk associated with increased *GIGYF1* expression. Conditional analysis showed that the glucose and HbA1c associations of *GIGYF1* pLOF and rs221783 are independent of each other (Supplementary Table 18).

The association of rs221783 with glucose levels replicated in Biobank Japan ($p = 1.7 \times 10^{-4}$, effect = -0.05 SD for T allele) [20] whilst in FinnGen, rs221783 showed a nominal association with T2D diagnosis (p = 0.02, OR = 0.96 for T allele) (Supplementary Table 19). The association with thyroid disease has been replicated elsewhere [21].

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

Table 6: Common variant associations at the *GIGYF1* locus

Associations for the array-typed variant rs221783. For quantitative traits the effect is shown in standard deviations (beta) and for diagnoses as an odds ratio (OR). MAF; minor allele frequency.

phenotype	chrom	Pos (hg19/hg38)	Ref (effect allele)	Alt	rsid	MAF	pvalue	Effect (beta/OR)	95% CI -	95% CI +

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

glucose	7	100292914/ 100695291	T	C	rs221783	11%	1.82×10^{-11}	-0.03	-0.03	-0.02
HbA1c	7	100292914/ 100695291	T	C	rs221783	11%	3.58×10^{-7}	-0.02	-0.03	-0.01
Cholesterol	7	100292914/ 100695291	T	C	rs221783	11%	7.00×10^{-12}	0.03	0.02	0.03
LDL	7	100292914/ 100695291	T	C	rs221783	11%	6.25×10^{-10}	0.02	0.02	0.03
T2D	7	100292914/ 100695291	T	C	rs221783	11%	0.005	0.96	0.93	0.99
Hypothyroidism	7	100292914/ 100695291	T	C	rs221783	11%	6.95×10^{-7}	0.92	0.88	0.95

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 and colleagues [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 558 variants significantly associated with T2D or HbA1c ($p \leq 2.41 \times 10^{-5}$ adjusting for 2071 variant sets tested) - *ANK1, GCK, HNF1A, TNRC6B, SLC30A8, NF1, IRS2, CFTR* and *HNF4A* (Supplementary Figure 6 and Supplementary Table 20). 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 diabetes and biomarkers of glycemic control. These included several genes not previously implicated in diabetes, *GIGYF1*, *TNRC6B* and *PFAS*, as well as *GCK, HNF1A* and *PDX1*, known MODY genes [14, 22-24]. 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.

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 [4, 25]. 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 of the variant sets that associated with T2D also affected HbA1c and/or glucose levels giving us confidence in these associations. In addition, T2D associations for all genes, apart from *TNRC6B*, were significant ($p \leq 1.46 \times 10^{-6}$) using the linear mixed model implemented by SAIGE-Gene which can be more robust when dealing with low numbers of variant carriers [16]. We also verified the majority of our associations with glucose and HbA1c levels,

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

including those for *GIGYF1* pLOF, using independent measurements from primary care data. Additional confidence in our results comes from the fact that we identified genes known to be involved in Mendelian forms of diabetes and previously reported genes. In addition, 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 partly be due to a lack of statistical power. Several of these genes are constrained for pLOF variation and/or have small numbers of pLOF carriers in UKBB (for example, *PPARG* has just 16 pLOF carriers). However, for some of these genes such as *SLC5A2* (encoding SGLT2) we do not detect associations with diabetes traits despite good numbers 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 this cohort but we did not replicate associations for *TNRC6B* and *PFAS* variants. 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 especially for *PFAS* (see Methods) or the frequency of the relevant variants (for example, the frequency of *TNR6CB* pLOF is 0.01% in UKBB but 0.16% in GHS) 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 which negatively regulates both the insulin and IGF-1 receptors [26]. Transfection of cells with GRB10-binding fragments of GIGYF1 lead to greater activation of both the insulin and IGF-1 receptors [27]. 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 [28, 29]. *GIGYF1* is broadly expressed with high levels observed in endocrine tissues, pancreas and brain [19, 30]. GIGYF1 and the related protein GIGYF2 have also been implicated in translational repression [31] and translation-coupled mRNA decay [32] suggesting biological roles beyond regulation of insulin and IGF-1 receptor signaling.

PheWAS 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 [33]. 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 also associated with decreased mean corpuscular hemoglobin levels and increased

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

diagnosis of anemia as well as increased emphysema diagnosis. The biological basis for these associations is not clear. *GIGYF1* is highly expressed in lung [19, 30] although the emphysema association is driven by small numbers of individuals, so replication is required.

GIGYF1 pLOF associated with a 4.5-fold increased risk of hypothyroidism and *GIGYF1* is highly expressed in thyroid [19, 30] 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 [34-36]. 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 [37]. Damaging variants in *GIGYF1* have recently been implicated in conferring risk for developmental delay and autism spectrum disorders [38]. 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. Non-coding variants at the *GIGYF1* locus associated with glucose levels and T2D, and we replicated these findings in independent datasets. These variants 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 [9] just nine had rare variant associations at a nearby gene.

We assessed the impact of pLOF and predicted damaging missense variants in approximately 17,000 genes 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 [39]. 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

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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 [40]. 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 (Field 21000) and ran an initial principal component analysis (PCA) on high quality common variants using eigenstrat [41]. SNPs were filtered for missingness across individuals < 2%, MAF > 1%, regions of known long range LD [42], 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 [41] 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 [43]. 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 [44]. 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 to using the GLnexus joint calling tool [45]. Post-variant calling filtering was applied using the Goldilocks pipeline [43]. Variants were annotated using the Ensembl Variant Effect Predictor v95 [46] which includes a LOFTEE plug-in to identify high confidence (HC) pLOF variants [13]. Combined Annotation Dependent Depletion (CADD) scores were generated using the Whole Genome Sequence Annotator (WGSA) AMI version 0.8.

Phenotype definitions

Blood biochemistry values were obtained for glucose (Field 30740) and HbA1c (Field 30750) from UKBB and inverse rank normalized using the RNOmni R package [47], resulting in an approximately normal distribution.

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 40006) 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 phenome-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 (h^2 significance flagged as at least “nominal” with a confidence level flagged as “medium” or “high”) by the

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Neale lab [25], using PHESANT 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 RNOmni R package. [47]. For burden testing, we required at least 10 carriers to have measurements. We also tested associations with ICD10-coded diagnoses (using 3 character codes) that had more than 500 cases in the White subset of 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 42W5., 44TB., 66Ae0, 44TC., 42W4. and read 3 codes XaPbt, X772q, XaWP9, XaBLm, XaERp.

Values were converted to IFCC units where necessary. Aberrantly high (≥ 45 mmol/L for glucose, ≥ 300 mmol/mol for HbA1c) and extremely low values (≤ 0.6 mmol/L for glucose, ≤ 10 mmol/mol for HbA1c) were excluded. The mean measurement per individual was then taken and inverse rank normalized prior to association testing. The mean age at measurement was also extracted and used as a covariate in the regression.

Individuals taking cholesterol-lowering medication were identified using self-reported medications recorded at their UKBB interview (Field 20003) and whether cholesterol-lowering medications were recorded using the touchscreen questionnaire (Fields 6177 and 6153).

Gene-based association testing

For gene-based tests, autosomal rare pLOF variants were identified as follows; LOFTEE high confidence LOFs, MAF ≤ 1%, missingness across individuals ≤ 2%, HWE p-value ≥ 10^{-10}. Predicted damaging missense variants were defined as missense variants with a CADD PHRED-scaled score ≥ 25, MAF ≤ 1%, missingness across individuals ≤ 2%, HWE p-value ≥ 10^{-10}**.** Only genes with more than one pLOF variant or damaging missense variant were tested.

Burden testing was performed unrelated White subset using glm in R, using a gaussian model for quantitative traits and a binomial model for case-control analyses. Genotype was coded as 0 (no variant) or 1 (any number of variants). We adjusted for age, sex and the first 12 PCs of genetic ancestry in the regression. Additionally, when testing for association with disease diagnoses, we included country of recruitment as a covariate as the time of available follow-up differs between England, Scotland and Wales. Recruitment country was defined using the location of the relevant UKBB recruitment center (Field 54). Associations were later confirmed using just participants recruited in England. For case-control analyses we only ran tests where there was at least one expected case carrier based on variant frequency and disease prevalence. For quantitative traits we required at least 10 carriers to have measurements.

For glucose and HbA1c, to convert effect sizes from normalized values back to measured units, the estimates from the regression were multiplied by the standard deviation of these traits in the entire cohort.

SAIGE-Gene was run using the SAIGE R package (v0.36.5) [48] using settings recommend by the developers and related individuals were included.

T2D drug targets were defined according to Flannick et al. [5].

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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

Array association testing

Genotypes were obtained through array typing and imputation as described previously [49]. Population definition and PC estimation for individuals with array data was performed as previously described [50]. We tested all variants with imputation quality score (info) ≥ 0.8 and minor allele frequency (MAF) ≥ 0.1% in a 200Mb region around *GIGYF1* for association with glucose, HbA1c, T2D and hypothyroidism. Association analyses were performed using an additive model in PLINK adjusting for age at recruitment to UKBB, sex and the first 12 PCs of genetic ancestry. We also adjusted for country of recruitment where appropriate. The most significant variant with info > 0.95 was selected as the lead variant at the locus.

We replicated the association of rs221783 with glucose using available summary statistics for Biobank Japan for the trait "blood sugar" (http://jenger.riken.jp/en/result) [20]. We replicated the association of this variant with T2D diagnosis using summary statistics from FinnGen release 3 for the phenotype "E4_DM2" (https://www.finngen.fi/en/access_results). The effect allele in these datasets was the alternate allele "C". For consistency with the UKBB associations we have shown the effect for the "T" allele.

Meta-analysis of the UKBB and replication dataset association results was performed with the METAL software package using the classical method [51].

Region plots were created using LocusZoom [52]. LD calculations were performed in the White population for array variants in a 500kb sliding window as follows; we extracted genotypes with info > 0.9, rounded them to whole numbers, mean-imputed missing genotypes and used the R "cor" function to compute R which was then squared to get an R^2 value.

Gene expression and eQTL analysis

The expression of *GIGYF1* in various tissues was assessed using the GTEx portal (accessed 08/04/2020) [19] and Human Protein Atlas (http://www.proteinatlas.org) [30]. eQTL data for rs221783 was obtained from GTEx v8. For each tissue of interest, the best eQTL for *GIGYF1* was identified (GTEx v8 "eGene"). R^2 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 [53]. A subset of participants sequenced as part of the GHS-Regeneron 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 [54]. 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

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

sites, variants leading to stop gain, stop loss and start loss. The five missense deleterious algorithms used were SIFT [55], PolyPhen2 (HDIV), PolyPhen2 (HVAR) [56], LRT [57], and MutationTaster [58]. Association testing was performed in the European ancestry population using the Firth logistic regression test implemented in REGENIE [59] as previously described [54].

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 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 with participants is required). Research in GHS was approved by the GHS IRB, approval number 2006-0258. Written informed consent was obtained from all participants in UKBB and GHS.

Data availability

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 [38]; the remainder of the exome data used is scheduled for public release in 2021. Summary statistics for gene-level tests will be made available upon publication.

Author Contributions

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

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.

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 Gloyn for comments on the manuscript. Data management and analytics were performed using the REVEAL/SciDB translational analytics platform from Paradigm4.

References

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

1. King EA, Davis JW, Degner JF. 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 Genetics. 2019;15(12):e1008489. doi: 10.1371/journal.pgen.1008489.

2. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, et al. The support of human genetic evidence for approved drug indications. Nature Genetics. 2015;47(8):856-60. doi: 10.1038/ng.3314.

3. Nguyen PA, Born DA, Deaton AM, Nioi P, Ward LD. Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. Nature Communications. 2019;10(1):1579. doi: 10.1038/s41467-019-09407-3.

4. Cirulli ET, White S, Read RW, Elhanan G, Metcalf WJ, Tanudjaja F, et al. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. Nat Commun. 2020;11(1):542. Epub 2020/01/30. doi: 10.1038/s41467-020-14288-y. PubMed PMID: 31992710; PubMed Central PMCID: PMCPMC6987107.

5. Flannick J, Mercader JM, Fuchsberger C, Udler MS, Mahajan A, Wessel J, et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. Nature. 2019;570(7759):71-6. Epub 2019/05/24. doi: 10.1038/s41586-019-1231-2. PubMed PMID: 31118516; PubMed Central PMCID: PMCPMC6699738.

6. Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, et al. The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. PLOS Genetics. 2015;11(4):e1005165. doi: 10.1371/journal.pgen.1005165.

7. Bush WS, Oetjens MT, Crawford DC. Unravelling the human genome-phenome relationship using phenome-wide association studies. Nat Rev Genet. 2016;17(3):129-45. Epub 2016/02/16. doi: 10.1038/nrg.2015.36. PubMed PMID: 26875678.

8. Diogo D, Tian C, Franklin CS, Alanne-Kinnunen M, March M, Spencer CCA, et al. Phenome-wide association studies across large population cohorts support drug target validation. Nat Commun. 2018;9(1):4285. Epub 2018/10/18. doi: 10.1038/s41467-018-06540-3. PubMed PMID: 30327483; PubMed Central PMCID: PMCPMC6191429.

9. Vujkovic M, Keaton JM, Lynch JA, Miller DR, Zhou J, Tcheandjieu C, 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. 2020;52(7):680-91. Epub 2020/06/17. doi: 10.1038/s41588-020-0637-y. PubMed PMID: 32541925; PubMed Central PMCID: PMCPMC7343592.

10. Xue A, Wu Y, Zhu Z, Zhang F, Kemper KE, Zheng Z, et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. Nature Communications. 2018;9(1):2941. doi: 10.1038/s41467-018-04951-w.

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. 2012;20(7):801-5. Epub 2012/02/02. doi: 10.1038/ejhg.2012.3. PubMed PMID: 22293688; PubMed Central PMCID: PMCPMC3376268.

12. Scott RA, Scott LJ, Magi R, Marullo L, Gaulton KJ, Kaakinen M, et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes. 2017;66(11):2888-902. Epub 2017/06/02. doi: 10.2337/db16-1253. PubMed PMID: 28566273; PubMed Central PMCID: PMCPMC5652602.

13. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43. Epub 2020/05/29. doi: 10.1038/s41586-020-2308-7. PubMed PMID: 32461654; PubMed Central PMCID: PMCPMC7334197.

14. Fajans SS, Bell GI, Polonsky KS. Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young. N Engl J Med. 2001;345(13):971-80. Epub 2001/09/29. doi: 10.1056/NEJMra002168. PubMed PMID: 11575290.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

15. Wessel J, Chu AY, Willems SM, Wang S, Yaghootkar H, Brody JA, et al. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. Nature Communications. 2015;6(1):5897. doi: 10.1038/ncomms6897.

16. Zhou W, Zhao Z, Nielsen JB, Fritsche LG, LeFaive J, Gagliano Taliun SA, et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. Nature Genetics. 2020;52(6):634-9. doi: 10.1038/s41588-020-0621-6.

17. Tikkanen E, Gustafsson S, Amar D, Shcherbina A, Waggott D, Ashley EA, et al. Biological Insights Into Muscular Strength: Genetic Findings in the UK Biobank. Scientific Reports. 2018;8(1):6451. doi: 10.1038/s41598-018-24735-y.

18. Willems SM, Wright DJ, Day FR, Trajanoska K, Joshi PK, Morris JA, et al. Large-scale GWAS identifies multiple loci for hand grip strength providing biological insights into muscular fitness. Nature Communications. 2017;8(1):16015. doi: 10.1038/ncomms16015.

19. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550(7675):204-13. doi: 10.1038/nature24277.

20. Kanai M, Akiyama M, Takahashi A, Matoba N, Momozawa Y, Ikeda M, et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat Genet. 2018;50(3):390-400. Epub 2018/02/07. doi: 10.1038/s41588-018-0047-6. PubMed PMID: 29403010.

21. Saevarsdottir S, Olafsdottir TA, Ivarsdottir EV, Halldorsson GH, Gunnarsdottir K, Sigurdsson A, et al. FLT3 stop mutation increases FLT3 ligand level and risk of autoimmune thyroid disease. Nature. 2020;584(7822):619-23. doi: 10.1038/s41586-020-2436-0.

22. Froguel P, Vaxillaire M, Sun F, Velho G, Zouali H, Butel MO, et al. Close linkage of glucokinase locus on chromosome 7p to early-onset non-insulin-dependent diabetes mellitus. Nature. 1992;356(6365):162-4. Epub 1992/03/12. doi: 10.1038/356162a0. PubMed PMID: 1545870.

23. Ellard S. Hepatocyte nuclear factor 1 alpha (HNF-1 alpha) mutations in maturity-onset diabetes of the young. Hum Mutat. 2000;16(5):377-85. Epub 2000/11/03. doi: 10.1002/1098-1004(200011)16:5<377::AID-HUMU1>3.0.CO;2-2. PubMed PMID: 11058894.

24. Stoffers DA, Ferrer J, Clarke WL, Habener JF. Early-onset type-II diabetes mellitus (MODY4) linked to IPF1. Nat Genet. 1997;17(2):138-9. Epub 1997/11/05. doi: 10.1038/ng1097-138. PubMed PMID: 9326926.

25. Churchhouse C. Details and Considerations of the UK Biobank GWAS. Neale lab. 2017;http://www.nealelab.is/blog/2017/9/11/details-and-considerations-of-theuk-biobank-gwas.

26. Holt LJ, Siddle K. Grb10 and Grb14: enigmatic regulators of insulin action--and more? Biochem J. 2005;388(Pt 2):393-406. Epub 2005/05/20. doi: 10.1042/BJ20050216. PubMed PMID: 15901248; PubMed Central PMCID: PMCPMC1138946.

27. Giovannone B, Lee E, Laviola L, Giorgino F, Cleveland KA, Smith RJ. Two novel proteins that are linked to insulin-like growth factor (IGF-I) receptors by the Grb10 adapter and modulate IGF-I signaling. J Biol Chem. 2003;278(34):31564-73. Epub 2003/05/29. doi: 10.1074/jbc.M211572200. PubMed PMID: 12771153.

28. Prokopenko I, Poon W, Magi R, Prasad BR, Salehi SA, Almgren P, et al. A central role for GRB10 in regulation of islet function in man. PLoS Genet. 2014;10(4):e1004235. Epub 2014/04/05. doi: 10.1371/journal.pgen.1004235. PubMed PMID: 24699409; PubMed Central PMCID: PMCPMC3974640.

29. Rampersaud E, Damcott CM, Fu M, Shen H, McArdle P, Shi X, et al. Identification of novel candidate genes for type 2 diabetes from a genome-wide association scan in the Old Order Amish: evidence for replication from diabetes-related quantitative traits and from independent populations. Diabetes. 2007;56(12):3053-62. Epub 2007/09/12. doi: 10.2337/db07-0457. PubMed PMID: 17846126.

30. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419. Epub 2015/01/24. doi: 10.1126/science.1260419. PubMed PMID: 25613900.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

31. Peter D, Weber R, Sandmeir F, Wohlbold L, Helms S, Bawankar P, et al. GIGYF1/2 proteins use auxiliary sequences to selectively bind to 4EHP and repress target mRNA expression. Genes Dev. 2017;31(11):1147-61. Epub 2017/07/13. doi: 10.1101/gad.299420.117. PubMed PMID: 28698298; PubMed Central PMCID: PMCPMC5538437.

32. Weber R, Chung MY, Keskeny C, Zinnall U, Landthaler M, Valkov E, et al. 4EHP and GIGYF1/2 Mediate Translation-Coupled Messenger RNA Decay. Cell Rep. 2020;33(2):108262. Epub 2020/10/15. doi: 10.1016/j.celrep.2020.108262. PubMed PMID: 33053355.

33. Klimentidis YC, Arora A, Newell M, Zhou J, Ordovas JM, Renquist BJ, et al. Type-2 diabetes with low LDL-C: genetic insights into a unique phenotype. bioRxiv. 2019;10.1101/837013 %J bioRxiv:837013. doi: 10.1101/837013 %J bioRxiv.

34. Clement S, Refetoff S, Robaye B, Dumont JE, Schurmans S. Low TSH requirement and goiter in transgenic mice overexpressing IGF-I and IGF-Ir receptor in the thyroid gland. Endocrinology. 2001;142(12):5131-9. Epub 2001/11/20. doi: 10.1210/endo.142.12.8534. PubMed PMID: 11713206.

35. Kimura T, Van Keymeulen A, Golstein J, Fusco A, Dumont JE, Roger PP. Regulation of thyroid cell proliferation by TSH and other factors: a critical evaluation of in vitro models. Endocr Rev. 2001;22(5):631-56. Epub 2001/10/06. doi: 10.1210/edrv.22.5.0444. PubMed PMID: 11588145.

36. Zaballos MA, Santisteban P. FOXO1 controls thyroid cell proliferation in response to TSH and IGF-I and is involved in thyroid tumorigenesis. Mol Endocrinol. 2013;27(1):50-62. Epub 2012/11/20. doi: 10.1210/me.2012-1032. PubMed PMID: 23160481; PubMed Central PMCID: PMCPMC5416949.

37. Mishra R, Hodge KM, Cousminer DL, Leslie RD, Grant SFA. A Global Perspective of Latent Autoimmune Diabetes in Adults. Trends Endocrinol Metab. 2018;29(9):638-50. Epub 2018/07/26. doi: 10.1016/j.tem.2018.07.001. PubMed PMID: 30041834.

38. Satterstrom FK, Kosmicki JA, Wang J, Breen MS, De Rubeis S, An JY, et al. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell. 2020;180(3):568-84 e23. Epub 2020/01/26. doi: 10.1016/j.cell.2019.12.036. PubMed PMID: 31981491; PubMed Central PMCID: PMCPMC7250485.

39. Allen N, Sudlow C, Downey D, Peakman T, Danesh J, Elliott P, et al. UK Biobank: current status and what it means for epidemiology. Health Policy Technol. 2012;1(3):123-6.

40. Staples J, Qiao D, Cho MH, Silverman EK, University of Washington Center for Mendelian G, Nickerson DA, et al. PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. Am J Hum Genet. 2014;95(5):553-64. Epub 2014/12/03. doi: 10.1016/j.ajhg.2014.10.005. PubMed PMID: 25439724; PubMed Central PMCID: PMCPMC4225580.

41. Wang L, Zhang W, Li Q. AssocTests: An R Package for Genetic Association Studies. Journal of Statistical Software; Vol 1, Issue 5 (2020). 2020.

42. Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, et al. Long-range LD can confound genome scans in admixed populations. Am J Hum Genet. 2008;83(1):132-5; author reply 5-9. Epub 2008/07/09. doi: 10.1016/j.ajhg.2008.06.005. PubMed PMID: 18606306; PubMed Central PMCID: PMCPMC2443852.

43. Van Hout CV, Tachmazidou I, Backman JD, Hoffman JD, Liu D, Pandey AK, et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature. 2020;586(7831):749-56. Epub 2020/10/23. doi: 10.1038/s41586-020-2853-0. PubMed PMID: 33087929.

44. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60. Epub 2009/05/20. doi: 10.1093/bioinformatics/btp324. PubMed PMID: 19451168; PubMed Central PMCID: PMCPMC2705234.

45. Lin MF, Rodeh O, Penn J, Bai X, Reid JG, Krasheninina O, et al. GLnexus: joint variant calling for large cohort sequencing. bioRxiv. 2018:343970. doi: 10.1101/343970.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

46. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):122. Epub 2016/06/09. doi: 10.1186/s13059-016-0974-4. PubMed PMID: 27268795; PubMed Central PMCID: PMCPMC4893825.

47. McCaw ZR, Lane JM, Saxena R, Redline S, Lin X. Operating Characteristics of the Rank-Based Inverse Normal Transformation for Quantitative Trait Analysis in Genome-Wide Association Studies. bioRxiv. 2019:635706. doi: 10.1101/635706.

48. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018;50(9):1335-41. Epub 2018/08/15. doi: 10.1038/s41588-018-0184-y. PubMed PMID: 30104761; PubMed Central PMCID: PMCPMC6119127.

49. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. Genome-wide genetic data on ~500,000 UK Biobank participants. bioRxiv. 2017:166298. doi: 10.1101/166298.

50. Ward LD, Tu H-C, Quenneville C, Flynn-Carroll AO, Parker MM, Deaton AM, et al. Genome-wide association study of circulating liver enzymes reveals an expanded role for manganese transporter SLC30A10 in liver health. 2020:2020.05.19.104570. doi: 10.1101/2020.05.19.104570 %J bioRxiv.

51. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190-1. Epub 2010/07/10. doi: 10.1093/bioinformatics/btq340. PubMed PMID: 20616382; PubMed Central PMCID: PMCPMC2922887.

52. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26(18):2336-7. Epub 2010/07/17. doi: 10.1093/bioinformatics/btq419. PubMed PMID: 20634204; PubMed Central PMCID: PMCPMC2935401.

53. Dewey FE, Murray MF, Overton JD, Habegger L, Leader JB, Fetterolf SN, et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science. 2016;354(6319). Epub 2016/12/23. doi: 10.1126/science.aaf6814. PubMed PMID: 28008009.

54. Kosmicki JA, Horowitz JE, Banerjee N, Lanche R, Marcketta A, Maxwell E, et al. A catalog of associations between rare coding variants and COVID-19 outcomes. medRxiv. 2021. Epub 2021/03/04. doi: 10.1101/2020.10.28.20221804. PubMed PMID: 33655273; PubMed Central PMCID: PMCPMC7924298.

55. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016;11(1):1-9. Epub 2015/12/04. doi: 10.1038/nprot.2015.123. PubMed PMID: 26633127.

56. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7 20. Epub 2013/01/15. doi: 10.1002/0471142905.hg0720s76. PubMed PMID: 23315928; PubMed Central PMCID: PMCPMC4480630.

57. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553-61. Epub 2009/07/16. doi: 10.1101/gr.092619.109. PubMed PMID: 19602639; PubMed Central PMCID: PMCPMC2752137.

58. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7(8):575-6. Epub 2010/08/03. doi: 10.1038/nmeth0810-575. PubMed PMID: 20676075.

59. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. Computationally efficient whole genome regression for quantitative and binary traits. 2020:2020.06.19.162354. doi: 10.1101/2020.06.19.162354 %J bioRxiv.

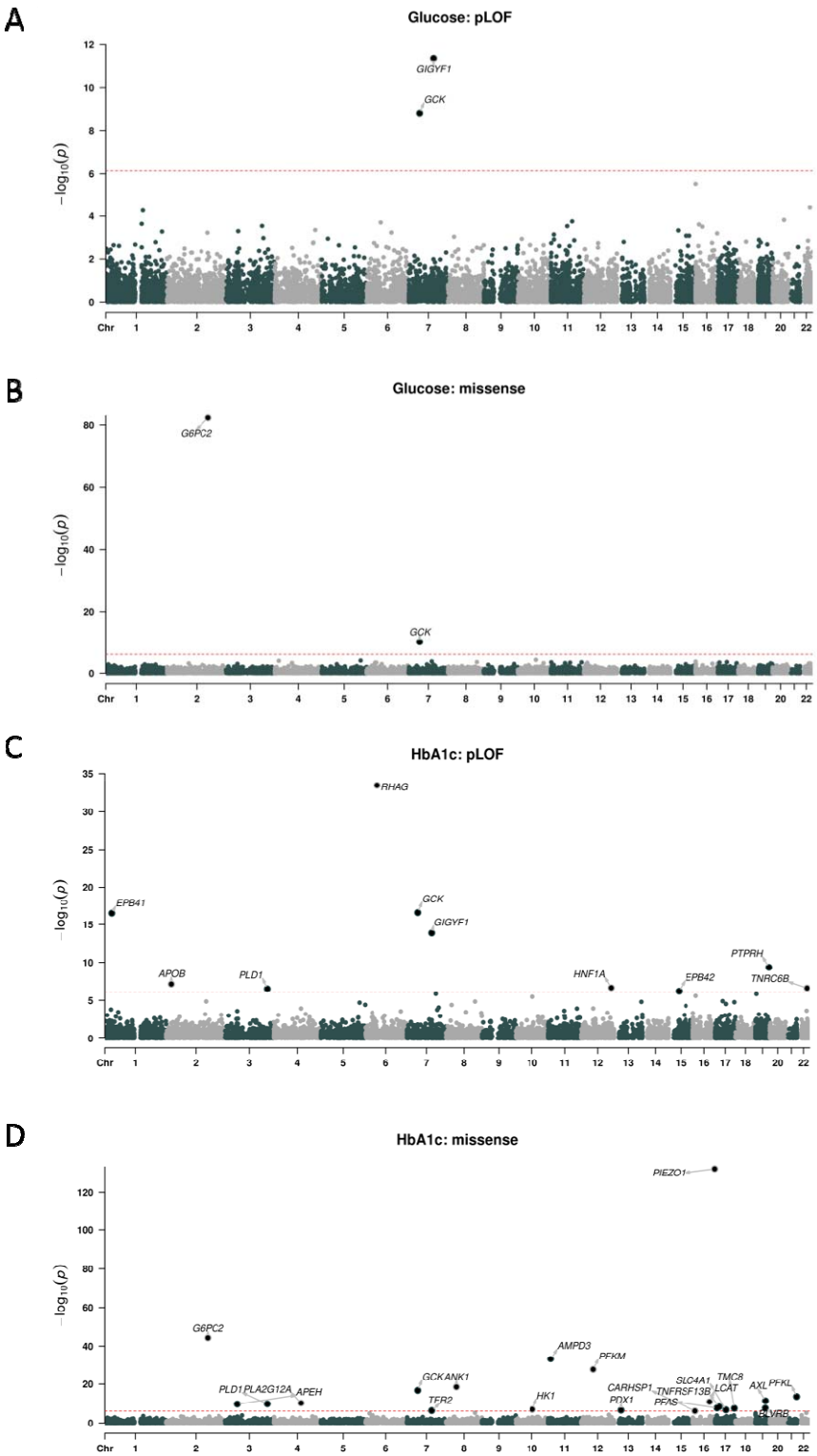
medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figures

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

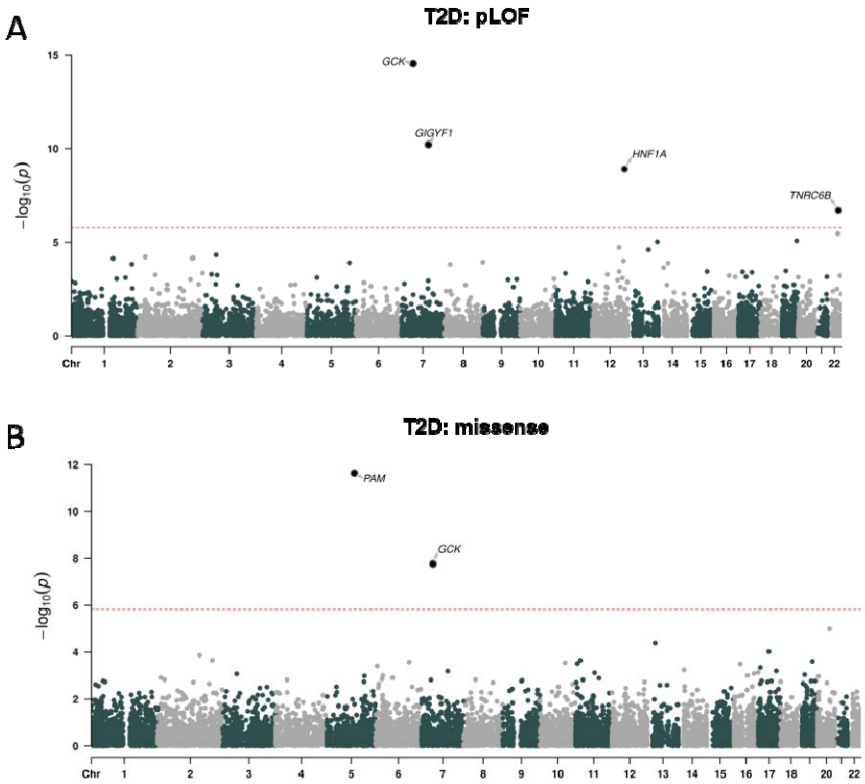
medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .



medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 2: Gene-level associations with T2D

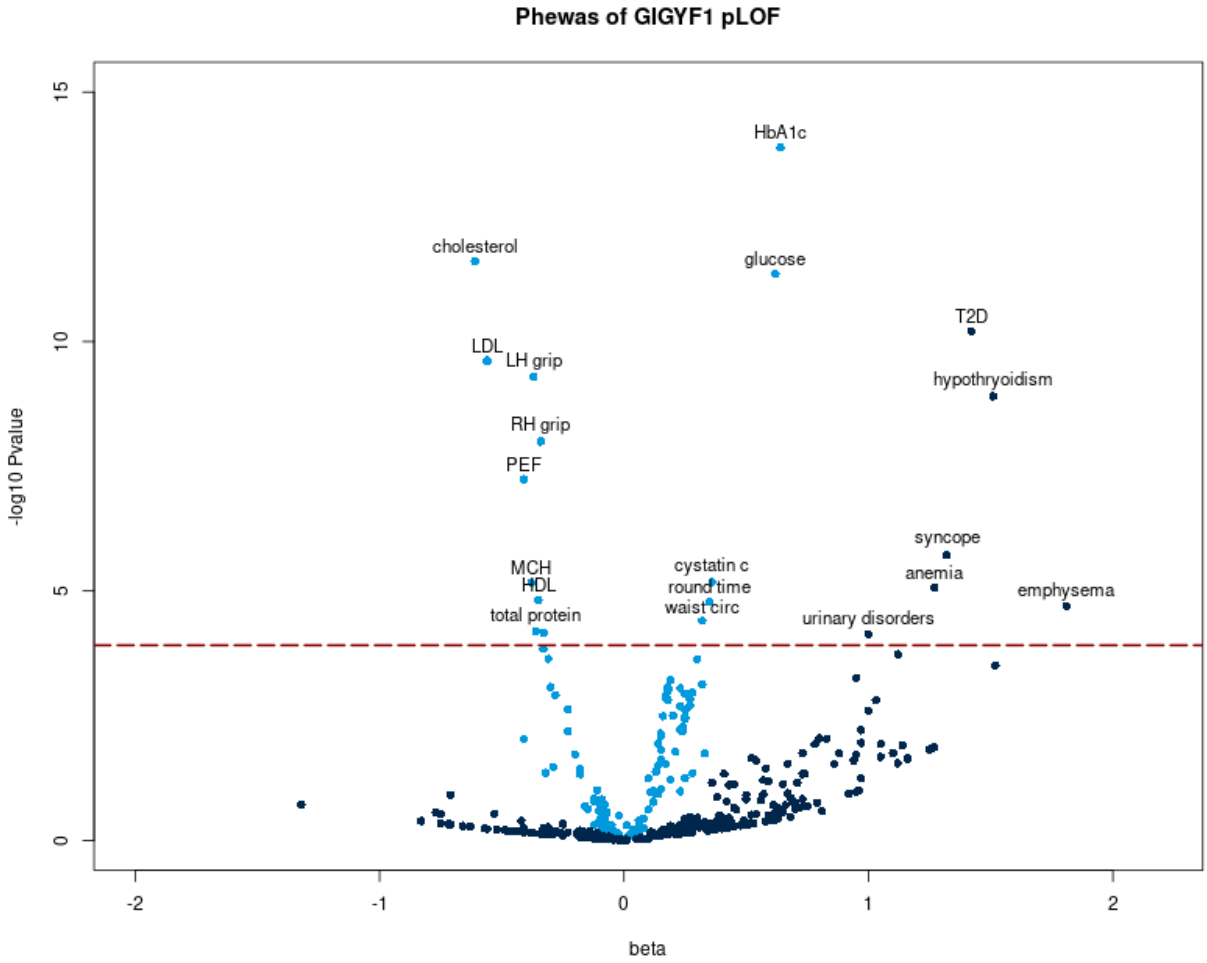
A) pLOF associations with T2D diagnosis. B) Damaging missense variant (CADD score ≥ 25) associations with T2D diagnosis. The red line indicates the threshold for significance, genes with significant associations are labeled.



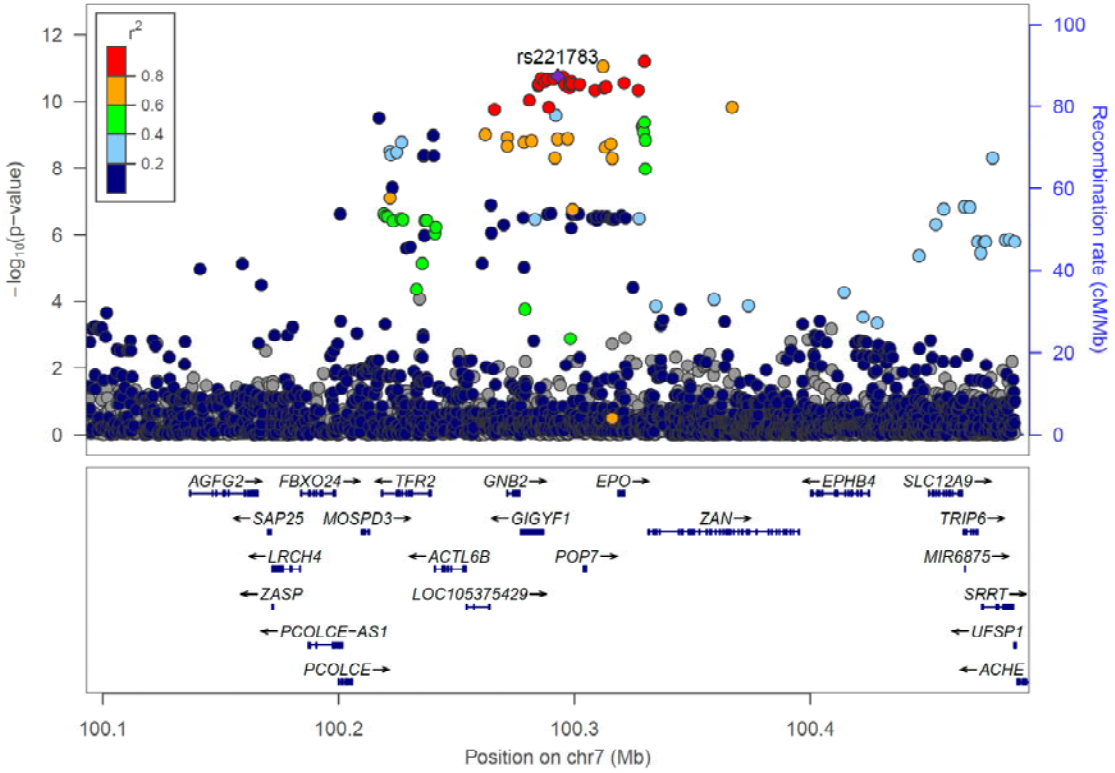
medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 3: PheWAS of *GIGYF1* pLOF

The x-axis is the beta (effect size in standard deviations) for the association and the y-axis is -log10(p-value). Quantitative traits are colored light blue and ICD10 diagnoses colored dark blue. Phenome-wide significant associations are labeled. The dashed line indicates the p-value threshold for phenome-wide significance. Protein; total protein, RH grip; right hand grip strength, round time: time to complete round (cognitive test), LH grip; left hand grip strength, PEF; peak expiratory flow.



medRxiv preprint doi: https://doi.org/10.1101/2021.01.19.21250105; this version posted April 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 4: Locus plot of glucose associations at the *GIGYF1* locus

Association results for array genotyped and imputed variants are shown. The purple diamond represents the lead variant rs221783. Other variants are colored according to correlation (R^2) with this marker (legend at top-left). The region displayed is chr7: 100092914-100492914. Genomic coordinates are for hg19.