Analysis of rare coding variants in 200,000 exome-sequenced subjects reveals novel genetic risk factors for type 2 diabetes

Running title: Rare variants in type 2 diabetes

David Curtis

UCL Genetics Institute, UCL, Darwin Building, Gower Street, London WC1E 6BT.

Centre for Psychiatry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ.

Correspondence:

David Curtis *d.curtis@ucl.ac.uk*

Abstract 250 words

Main text 3981 words

36 references

4 tables

1 figure

Abstract

Aims

The study aimed to elucidate the effects of rare genetic variants on the risk of type 2 diabetes (T2D).

Materials and methods

Weighted burden analysis of rare variants was applied to a sample of 200,000 exome-sequenced participants in the UK Biobank project, of whom over 13,000 were identified as having T2D. Variant weights were allocated based on allele frequency and predicted effect, as informed by a previous analysis of hyperlipidaemia.

Results

There was an exome-wide significant increased burden of rare, functional variants in three genes, *GCK*, *HNF4A* and *GIGYF1*. *GIGYF1* has not previously been identified as a diabetes risk gene and its product appears to be involved in the modification of insulin signalling. A number of other genes did not attain exome-wide significance but were highly ranked and potentially of interest, including *ALAD*, *PPARG*, *GYG1* and *GHRL*. Loss of function (LOF) variants were associated with T2D in *GCK* and *GIGYF1* whereas nonsynonymous variants annotated as probably damaging were associated in *GCK* and *HNF4A*. Overall, fewer than 1% of T2D cases carried one of these variants. In *HNF1A* and *HNF1B* there was an excess of LOF variants among cases but the small numbers of these fell short of statistical significance.

Conclusions

Rare genetic variants make an identifiable contribution to T2D in a small number of cases but these may provide valuable insights into disease mechanisms. As larger samples become available it is likely that additional genetic factors will be identified.

This research has been conducted using the UK Biobank Resource.

Keywords`

GCK; *HNF4A*; *GIGYF1*; *ALAD*; *PPARG*.

Introduction

Genome-wide association studies of type 2 diabetes (T2D) have implicated a large number of common genetic variants [1]. In the UK Biobank, a genetic risk score derived from common variants was associated with T2D and incorporating it alongside conventional risk factors in order to predict T2D increased the area under the curve (AUC) from 0.851 to 0.855 [2]. Common variants have individually modest effects on risk but a small number of genes have been identified in which rare coding variants with large effects can result in a phenotype of non-insulin dependent diabetes. Maturity onset diabetes of the young (MODY) is caused by mutations in a number of genes including *HNF1A*, *HNF4A*, *GCK*, *HNF1B*, *KCNJ11*, *ABCC8* [3]. Loss of function (LOF) and nonsynonymous variants in *PPARG* can cause a familial lipodystrophy with insulin resistant diabetes [4]. Recessively acting mutations in the *INSR* gene can cause insulin resistance with hyperglycaemia [5]. These rare coding variants have typically been identified by targeted studies of familial cases and individuals with severe phenotypes but the availability of large samples of exome sequenced subjects now allows the

possibility to explore the effects of variations in these genes in the population more broadly and potentially to discover novel genes. A study of 16,000 exome and genome sequenced cases and controls failed to detect a substantial contribution from rare variants to risk of type 2 diabetes [6]. However a larger study using 20,791 cases and 24,440 controls identified three exome-wide significant genes, *MC4R*, *SLC30A8* and *PAM*, with *SLC30A8* variants tending to be protective, as well as a specific variant in *MC4R*, rs79783591 (Ile269Asn), which was exome-wide significant in single variant analyses [7]. A study in which actionable MODY genes were sequenced revealed an increased burden of variants predicted to be pathogenic or likely pathogenic in T2D cases versus controls, especially driven by variants in *GCK* [8]. Rare variant association studies and functional studies have implicated two specific nonsynonymous variants in *PAM*, rs78408340 and rs35658696 [6, 9, 10].

Exome sequence data is now available for 200,000 of the 500,000 UK Biobank subjects [11]. We have recently analysed this in order to illuminate the effect of rare, coding variants on susceptibility to hyperlipidaemia and we now apply the same approach to study T2D [12].

Materials and Methods

The UK Biobank dataset was downloaded along with the variant call files for 200,632 subjects who had undergone exome-sequencing and genotype-calling by the UK Biobank Exome Sequencing Consortium using the GRCh38 assembly with coverage 20X at 95.6% of sites on average [11]. UK Biobank had obtained ethics approval from the North West Multi-centre Research Ethics Committee which covers the UK (approval number: 11/NW/0382) and had obtained informed consent from all participants. The UK Biobank approved an application for use of the data (ID 51119) and ethics approval for the analyses was obtained from the UCL Research Ethics Committee (11527/001). All variants were annotated using the standard software packages VEP, PolyPhen and SIFT [13–15]. To obtain population principal components reflecting ancestry, all variants with minor allele frequency (MAF) >= 0.1 were extracted using version 2.0 of *plink* (https://www.cog-genomics.org/plink/2.0/) [16, 17]. Then the command *--pca 20 approx* was run. This treats the genotype at each variant as a vector and then outputs the first 20 principal components.

The UK Biobank sample contains 503,317 subjects of whom 94.6% are of white ethnicity. As we have discussed in the report of a previous analysis of BMI, it has become standard practice for investigators to simply discard data from participants with other ancestries and we regard this as regrettable [18]. In that analysis we demonstrated that if population principal components are included as covariates then it is possible to include all participants, regardless of ancestry, in the type of weighted burden analysis described below without any inflation of the test statistic.

To define cases, a similar approach was used as was previously implemented for the investigation of hyperlipidaemia [12, 19]. The T2D phenotype was determined from three sources in the dataset: self-reported diabetes or type 2 diabetes (but not type 1 or gestational diabetes); reporting taking any of a list of named medications commonly used to treat T2D in the UK (https://www.diabetes.co.uk/Diabetes-drugs.html); having an ICD10 code for non-insulin-dependent diabetes mellitus in hospital records or as a cause of death. Subjects in any of these categories were deemed to be cases while all other subjects were taken to be controls.

The SCOREASSOC program was used to carry out a weighted burden analysis to test whether, in each gene, sequence variants which were rarer and/or predicted to have more severe functional effects occurred more commonly in cases than controls. Attention was restricted to rare variants with MAF <= 0.01 in both cases and controls. As previously described, variants were weighted by overall MAF so that variants with MAF=0.01 were given a weight of 1 while very rare variants with MAF close to zero were given a weight of 10 [18]. Variants were also weighted according to their functional

annotation using the GENEVARASSOC program, which was used to generate input files for weighted burden analysis by SCOREASSOC [20, 21]. The weights were informed from the analysis of the effects of different categories of variant in *LDLR* on hyperlipidaemia risk [12]. Variants predicted to cause complete LOF of the gene were assigned a weight of 100. Nonsynonymous variants were assigned a weight of 5 but if PolyPhen annotated them as possibly or probably damaging then 5 or 10 was added to this and if SIFT annotated them as deleterious then 20 was added. In order to allow exploration of the effects of different types of variant on disease risk the variants were also grouped into broader categories to be used in multivariate analyses as described below. The full set of weights and categories is displayed in Table 1. As described previously, the weight due to MAF and the weight due to functional annotation were multiplied together to provide an overall weight for each variant. Variants were excluded if there were more than 10% of genotypes missing in the controls or cases or if the heterozygote count was smaller than both homozygote counts in the controls or cases. If a subject was not genotyped for a variant then they were assigned the subject-wise average score for that variant. For each subject a gene-wise weighted burden score was derived as the sum of the variant-wise weights, each multiplied by the number of alleles of the variant which the given subject possessed. For variants on the X chromosome, hemizygous males were treated as homozygotes.

For each gene, logistic regression analysis was carried out including the first 20 population principal components and sex as covariates. The likelihood was first calculated for the case and control phenotypes predicted just by these covariates and then the likelihood was calculated again additionally incorporating the gene-wise weighted burden score for each subject. Then a likelihood ratio test was performed taking twice the natural log of the likelihood ratio as a chi-squared statistic with one degree of freedom. The statistical significance was summarised as a signed log p value (SLP), which is the log base 10 of the p value given a positive sign if the score is higher in cases and negative if it is higher in controls.

Gene set analyses were carried out as before using the 1454 "all GO gene sets, gene symbols" pathways as listed in the file *c5.all.v5.0.symbols.gmt* downloaded from the Molecular Signatures Database at http://www.broadinstitute.org/gsea/msigdb/collections.jsp [22]. For each set of genes, the natural logs of the gene-wise p values were summed according to Fisher's method to produce a chi-squared statistic with degrees of freedom equal to twice the number of genes in the set. The p value associated with this chi-squared statistic was expressed as a minus log10 p (MLP) as a test of association of the set with the T2D phenotype. In order to test for an overlap with common variant effects, an additional set was constructed consisting of the 33 putative functional genes whose expression was implicated as being associated with T2D risk in the previous GWAS meta-analysis [1].

For selected genes, additional analyses were carried out to clarify the contribution of different categories of variant. As described previously, logistic regression analyses were performed using as predictor variables the counts of the separate categories of variant as listed in Table 1, again including principal components and sex as covariates, to estimate the effect size for each category [12]. The odds ratios associated with each category were estimated along with their standard errors and the Wald statistic was used to obtain a p value, except for categories in which variants occurred fewer than 50 times in which case Fisher's exact test was applied to the variant counts. The associated p value was converted to an SLP, again with the sign being positive if the mean count was higher in cases than controls.

Data manipulation and statistical analyses were performed using GENEVARASSOC, SCOREASSOC and R [23].

Results

There were 13,938 cases of T2D and 186,694 controls. There were 20,384 genes for which there were qualifying variants. Given that there were 20,384 informative genes, the critical threshold for the absolute value of the SLP to declare a result as formally statistically significant is -log10(0.05/20384) = 5.61 and this was achieved by three genes, *GCK* (SLP = 22.24), *HNF4A* (SLP = 6.82) and *GIGYF1* (SLP = 6.22). The quantile-quantile (QQ) plot for the SLPs obtained for all genes except *GCK* is shown in Figure 1. This shows that the test appears to be well-behaved and conforms well with the expected distribution. Omitting the genes with the 100 highest and 100 lowest SLPs, which might be capturing a real biological effect, the gradient for positive SLPs is 1.06 with intercept at -0.007 and the gradient for negative SLPs is 1.02 with intercept at -0.009, indicating only modest inflation of the test statistic in spite of the fact that participants of all ancestries are included.

Table 2 shows all the genes achieving SLP with absolute value greater than 3, equivalent to an uncorrected p value of 0.001. Given that 20,384 genes were tested, one would expect that by chance about 20 would reach this level of significance whereas in fact there are 32. If we take the average gradient of the QQ plot, 1.04, as an indication of an inflation factor then we would use a corrected threshold of 3*1.04 = 3.12 as an appropriate threshold to indicate p = 0.001 and 29 genes exceed this. Thus it is possible that some of these highly ranked genes do demonstrate a biological signal which fails to reach statistical significance after correction for multiple testing. As discussed below, a number of these seem of potential interest, comprising *ALAD* (SLP = 3.63), *PPARG* (SLP = 3.45), *GYG1* (SLP = 3.22) and *GHRL* (SLP = -3.15). The results for all genes are provided in Supplementary Table S1.

In order to see if any additional genes were highlighted by analysing gene sets, gene set analysis was performed as described above after first removing all genes with absolute SLP value greater than 3. Given that 1,454 sets were tested a critical MLP to achieve to declare results significant after correction for multiple testing would be log10(1454*20) = 4.46 and this was not achieved by any set. There were two sets with MLP > 3, EXTERNAL SIDE OF PLASMA MEMBRANE (MLP = 3.29) and MONOSACCHARIDE TRANSMEMBRANE TRANSPORTER ACTIVITY (MLP = 3.06). The latter is of some interest because it consists of 10 genes, of which three were individually significant at p<0.05, these being *SLC2A2* (SLP = 2.54), *SLC2A3* (SLP = -2.37) and *SLC2A4* (SLP = 1.70). The set of 33 putative functional T2D genes implicated in the previous GWAS meta-analysis showed no evidence for association (MLP=0.86). The results for all sets are provided in Supplementary Table S2.

For the genes of possible interest listed above, a logistic regression analysis of different categories of variant was carried out to elucidate their relative contributions. The results for the three exome-wide significant genes are shown in Table 3, which shows differences between the genes relating to the implicated pattern of variants. The results for *GCK* demonstrate that splice site variants and gene disruptive variants, comprising frameshift and stop variants, are associated with large effects on risk. These occur a total of 17 times among the 13,938 cases. However of note is that nonsynonymous variants annotated as probably damaging by PolyPhen are also associated with increased risk, with OR = 2.97 (1.59 - 5.54), and these occur 33 times among cases. The situation for *HNF4A* is quite different. There are no splice site variants and only 6 gene disruptive variants and these all occur in controls. Only probably damaging nonsynonymous variants show an effect, with OR = 2.97 (1.61 - 5.50), and these occur 34 times among cases. Finally, for *GIGYF1* probably damaging nonsynonymous variants have no discernible effect and it is only the splice site (OR = 7.70 (2.62 - 22.67)) and disruptive (OR = 5.65 (3.07 - 10.40)) variants which increase risk and these occur 24 times in cases.

Table 3 also shows the results of variant category analysis for the genes which, while not exome-wide significant, had SLP with magnitude >3 and which appeared to be potentially of interest, *ALAD*,

PPARG, *GYG1* and *GHRL*. The signal for *ALAD* seems to be driven by the fact that although splice site and disruptive variants occur only a total of 5 times in cases this still makes them about 7 times as frequent as in controls. The signal for *PPARG* arises from the fact that disruptive variants occur 4 times in cases and 7 times in controls, OR = 8.23 (2.29 - 29.63). The findings for *GYG1* seem somewhat more robust, being based on an excess of both disruptive (OR = 1.98 (1.40 - 2.81)) and splice site (OR = 3.08 (0.96 - 9.81)) variants in cases, with a total of 37 occurrences. For *GHRL*, which has SLP = -3,15, implying that variants in it might be protective, no individual category has a statistically significant effect and there is more a general tendency for there to be fewer variants among cases which is spread over a number of categories, including 3 prime UTR, protein altering, indel, disruptive and deleterious.

The analyses described above failed to highlight a number of genes which have previously been implicated in T2D, comprising *HNF1A* (SLP = 1.66), *HNF1B* (SLP = -0.28), *ABCC8* (SLP = 1.94), *INSR* (SLP = -0.25), *MC4R* (SLP = 1.41), *SLC30A8* (SLP = -2.64) and *PAM* (SLP = 0.19). The association with T2D for each category of variant within these genes is shown in Table 4. From this it can be seen that LOF variants in *HNF1A* and *HNF1B* are commoner in cases than controls but that their absolute numbers are too small to produce statistically significant effects. By contrast, LOF variants have higher overall frequency in *ABCC8* and *INSR* and they have approximately equal frequencies in cases and controls. The results for *MC4R* suggest that variants which are disruptive or annotated as deleterious or damaging may be associated with a moderate increase in risk, with OR 1.2 – 1.5, but no individual category is statistically significant. Unfortunately genotypes for the single *MC4R* variant previously reported to be associated with T2D, rs79783591 producing an Ile269Asn substitution, were not provided in the supplied dataset even though it is clearly exonic and genotypes for nearby variants are available. The frequency of nonsynonymous variants annotated as deleterious in *SLC30A8* is higher in controls than cases, consistent with previous findings that missense variants in this gene typically result in reduced T2D risk, but again this is not statistically significant. In *PAM* there is no real suggestion of any signal among the included variants. However the frequency criteria used excluded the two variants which had previously been individually implicated and when these were analysed separately they both showed evidence for a small effect on risk, with rs35658696 (Asp563Gly) having SLP = 5.68, OR = 1.15 (1.08 – 1.21) and rs78408340 having SLP = 6.92, OR = 1.20 (1.13 – 1.28).

Discussion

Biobank data differs from that which can be obtained using specifically designed association studies in a number of ways. One must make use of measures which are available rather than being able to specify predetermined inclusion and exclusion criteria in order to decide which subjects to define as cases and controls. It also needs to be borne in mind that in the UK Biobank not all fields are complete for all participants and in order to maximise sample size one may wish to choose fields for which relatively complete information is available rather than those which may more closely define the phenotype of interest. Here, a pragmatic method for distinguishing cases and controls has been used which is intended to broadly reflect a real world diagnosis of type 2 diabetes. This is based on a combination of self-report, recorded diagnoses and prescribed medication. Of course it is to be expected that some participants categorised as cases might not meet operationally defined criteria for diabetes while some controls might well actually be found to have diabetes if they were specifically assessed for this. Nevertheless, this approach does produce a large sample which yields some interesting results.

Burden analyses makes the general assumption that rare variants which make changes to the function of a gene on average impair functioning rather than enhance it. This is clearly expected for LOF variants and for nonsynonymous variants it seems reasonable to expect that a random change is more likely to have a negative than a positive effect. However it must be recognised that gain of function variants can occur and that including all variants together in a burden analysis will tend to water down the effects of variants with opposing effects, reducing power. For variants that are common this issue can be addressed with methods such as the sequence kernel association test (SKAT) but such approaches are not applicable to extremely rare variants [24]. For a given category of nonsynonymous variant, for example those reported as deleterious by SIFT, it is to be expected that there will be range of effect sizes, perhaps in both directions. However because the variants concerned are extremely rare it is generally not possible to gain any accurate estimate of their individual effects and the ORs reported represent an average for the whole category. As more sequence data becomes available it may be possible to develop improved methods for predicting likely effects of coding variants on gene function. In turn, this will support the development of improved weighting schemes, for example by assigning negative weights to variants expected to produce gain of function.

In spite of the inherent limitations of available data and the analytical approaches used, some confidence in their validity may be gained from the fact that two established genes, *GCK* and *HNF4A*, are implicated with exome wide significance. Variants in *GCK* are recognised as the cause of up to half of cases of MODY, itself accounting for around 1-2% of cases of all diabetes diagnoses [25]. Likewise, *HNF4A* variants cause 5-10% of cases of MODY [26].

The third gene which is exome wide significant, *GIGYF1*, has not previously been implicated in the aetiology of diabetes. Its product binds to growth factor receptor-bound protein 10 (GRB10) and has a role in modulating the insulin-like growth factor (IGF1) receptor signalling pathway [27, 28]. A variant in *GRB10* has been reported to be associated with decreased early-phase insulin secretion and the muscle-specific ablation of *Grb10* in mice causes increased glucose uptake into muscles with increased insulin signalling [29, 30].

Among genes significant at $p < 0.001$, a number are of interest in the light of previous research. The expression of *ALAD* (SLP = 3.63) is reduced in obese subjects while the expression of *Alad* is reduced in rats with high-fat diet-induced weight gain and inhibition of ALAD with aminotriazole led to reduced glucose uptake in cultured human adipocytes [31]. The common P12A variant of *PPARG* (SLP = 3.45) reduces risk of T2D whereas rare LOF variants and nonsynonymous variants which cause reduced activity (occurring in approximately 1 in 1,000 individuals) have previously been reported to substantially increase risk [32]. Damaging variants in *GYG1* (SLP = 3.22) cause deficiency of glycogenin 1, resulting in glycogen storage myopathies, but have not been reported to be associated with diabetes [33]. *GHRL* (SLP = -3.15) encodes the ghrelin-obestatin preproprotein which is cleaved to yield two peptides, ghrelin and obestatin, which are involved in appetite and energy metabolism and there have been some studies which have claimed that the common Leu72Met (rs696217) variant is associated with reduced risk of T2D although the effect does not seem to be consistent and the gene was not highlighted in a large GWAS meta-analysis [1, 34].

It is noteworthy that in the gene set analyses the second highest ranked set is MONOSACCHARIDE TRANSMEMBRANE TRANSPORTER ACTIVITY. This includes *SLC2A2* (SLP = 2.54), *SLC2A3* (SLP = -2.37) and *SLC2A4* (SLP = 1.70). *SLC2A2*, previously known as *GLUT2*, codes for a glucose transporter expressed by beta cells which senses glucose levels and recessively acting variants in it can cause neonatal diabetes [35]. A common intronic variant of *SLC2A2*, rs8192675, is associated with the glycaemic response to metformin [36]. *SLC2A4* codes for a glucose transporter whose levels in cell membranes increase in response to insulin but although candidate gene studies claim that common variants in it are associated with T2D these results are not supported by properly powered GWAS

metanalysis [1, 37]. The results presented here suggest that rare coding variants in these glucose transporter genes might have some effect on risk of T2D.

Some genes which have been previously implicated in T2D failed to produce significant results. LOF variants in *HNF1A* and *HNF1B* were commoner in cases but are too rare to produce definitive results in this sample. By contrast, LOF variants in ABCC8 are less rare but do not differ in frequency between cases and controls. These results are consistent with reports that it is only activating variants in *ABCC8* which result in diabetes whereas LOF variants can cause hyperinsulinaemia and this reflects the fact that weighted burden analysis is not expected to detect gain of function effects [38]. The fact that the frequency of LOF variants in *INSR* is similar in cases and controls, indicates that, although recessively acting variants can cause infantile hyperinsulinaemia followed by insulin dependent diabetes, the loss of function of a single copy of this gene has little discernible effect on risk of T2D [5].

These analyses provide an overview of the way very rare genetic variants contribute to risk of type 2 diabetes in a large sample broadly representative of the UK population. The weighted burden analysis successfully identifies two known diabetes genes, *GCK* and *HNF4A*, and implicates a novel gene, *GIGYF1*. A few other biologically plausible genes do not reach formal levels of statistical significance after correction for multiple testing but might be worthy of further investigation, including *ALAD*, *PPARG*, *GYG1*, *GHRL*, *SCL2A2* and *SLC2A4*. Any possible role for these genes will become clearer as additional data becomes available, for example from the remaining 300,000 UK Biobank subjects for whom exome sequence data has not yet been released. Typically, hundreds of variants are identified per gene, mostly occurring in only a handful of subjects each. The findings for some previously implicated genes, *HNF1A*, *HNF1B*, *MC4R* and *SLC30A8*, are not formally statistically significant but are consistent with an effect which could be detected once the sample size is increased.

Together, variants which can be identified as having large effects on risk occur in fewer than 1% of the cases of T2D in this sample. There is no doubt that identifying specific genetic causes may be useful to guide treatment for some patients [4]. However it needs to be acknowledged that for the vast majority of patients with T2D exome sequencing will be not be helpful in terms of identifying specific subtypes of disease which might benefit from specific treatments. Thus, the potential to apply a personalised medicine approach to T2D based on genetic testing seems to be somewhat limited.

The main potential utility of genetic investigations such as this might be to better characterise the mechanisms which can lead to disease, identify novel drug targets and develop improved therapeutic approaches which would benefit T2D patients in general, rather than only the small number carrying the relevant genetic variant. If some of the findings reported here can be replicated then the genes identified could become the objects of more intensive investigation.

Conflicts of interest

The author declares he has no conflict of interest.

Author contributions

DC conceived the project, carried out the analyses and wrote the manuscript.

Data availability

The raw data is available on application to UK Biobank. Detailed results with variant counts cannot be made available because they might be used for subject identification. Scripts and relevant derived

variables will be deposited in UK Biobank. Software and scripts used to carry out the analyses are also available at https://github.com/davenomiddlenamecurtis.

Acknowledgments

This research has been conducted using the UK Biobank Resource. The author wishes to acknowledge the staff supporting the High Performance Computing Cluster, Computer Science Department, University College London. This work was carried out in part using resources provided by BBSRC equipment grant BB/R01356X/1. The author wishes to thank the participants who volunteered for the UK Biobank project.

References

1 Xue A, Wu Y, Zhu Z, Zhang F, Kemper KE, Zheng Z, Yengo L, Lloyd-Jones LR, Sidorenko J, Wu Y, Agbessi M, Ahsan H, Alves I, Andiappan A, Awadalla P, Battle A, Beutner F, Bonder MJJ, Boomsma D, Christiansen M, Claringbould A, Deelen P, Esko T, Favé MJ, Franke L, Frayling T, Gharib S, Gibson G, Hemani G, Jansen R, Kähönen M, Kalnapenkis A, Kasela S, Kettunen J, Kim Y, Kirsten H, Kovacs P, Krohn K, Kronberg-Guzman J, Kukushkina V, Kutalik Z, Lee B, Lehtimäki T, Loeffler M, Marigorta UM, Metspalu A, Milani L, Müller-Nurasyid M, Nauck M, Nivard M, Penninx B, Perola M, Pervjakova N, Pierce B, Powell J, Prokisch H, Psaty B, Raitakari O, Ring S, Ripatti S, Rotzschke O, Ruëger S, Saha A, Scholz M, Schramm K, Seppälä I, Stumvoll M, Sullivan P, Teumer A, Thiery J, Tong L, Tönjes A, van Dongen J, van Meurs J, Verlouw J, Völker U, Võsa U, Yaghootkar H, Zeng B, McRae AF, Visscher PM, Zeng J, Yang J. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat Commun* 2018;**9**:1–14.

2 Chen X, Liu C, Si S, Li Y, Li W, Yuan T, Xue F. Genomic risk score provides predictive performance for type 2 diabetes in the UK biobank. *Acta Diabetol* 2021;**1**:3.

3 Murphy R, Ellard S, Hattersley AT. Clinical implications of a molecular genetic classification of monogenic β-cell diabetes. Nat. Clin. Pract. Endocrinol. Metab. 2008;**4**:200–13.

4 Agostini M, Schoenmakers E, Beig J, Fairall L, Szatmari I, Rajanayagam O, Muskett FW, Adams C, Marais AD, O'Rahilly S, Semple RK, Nagy L, Majithia AR, Schwabe JWR, Blom DJ, Murphy R, Chatterjee K, Savage DB. A pharmacogenetic approach to the treatment of patients with PPARG mutations. *Diabetes* 2018;**67**:1086–92.

5 Semple RK, Savage DB, Cochran EK, Gorden P, O'Rahilly S. Genetic syndromes of severe insulin resistance. Endocr. Rev. 2011;**32**:498–514.

6 Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, Ma C, Fontanillas P, Moutsianas L, McCarthy DJ, Rivas MA, Perry JRB, Sim X, Blackwell TW, Robertson NR, Rayner NW, Cingolani P, Locke AE, Tajes JF, Highland HM, Dupuis J, Chines PS, Lindgren CM, Hartl C, Jackson AU, Chen H, Huyghe JR, Van De Bunt M, Pearson RD, Kumar A, Müller-Nurasyid M, Grarup N, Stringham HM, Gamazon ER, Lee J, Chen Y, Scott RA, Below JE, Chen P, Huang J, Go MJ, Stitzel ML, Pasko D, Parker SCJ, Varga T V., Green T, Beer NL, Day-Williams AG, Ferreira T, Fingerlin T, Horikoshi M, Hu C, Huh I, Ikram MK, Kim BJ, Kim Y, Kim YJ, Kwon MS, Lee J, Lee S, Lin KH, Maxwell TJ, Nagai Y, Wang X, Welch RP, Yoon J, Zhang W, Barzilai N, Voight BF, Han BG, Jenkinson CP, Kuulasmaa T, Kuusisto J, Manning A, Ng MCY, Palmer ND, Balkau B, Stančáková A, Abboud HE, Boeing H, Giedraitis V, Prabhakaran D, Gottesman O, Scott J, Carey J, Kwan P, Grant G, Smith JD, Neale BM, Purcell S, Butterworth AS, Howson JMM, Lee HM, Lu Y, Kwak SH, Zhao W, Danesh J, Lam VKL, Park KS, Saleheen D, So WY, Tam CHT, Afzal U, Aguilar D, Arya R, Aung T, Chan E, Navarro C, Cheng CY, Palli D, Correa A, Curran JE, Rybin D, Farook VS, Fowler SP, Freedman BI, Griswold M, Hale DE, Hicks PJ, Khor CC, Kumar S, Lehne B, Thuillier D, Lim WY, Liu J, Van Der Schouw YT, Loh M, Musani SK, Puppala

S, Scott WR, Yengo L, Tan ST, Taylor HA, Thameem F, Wilson G, Wong TY, Njolstad PR, Levy JC, Mangino M, Bonnycastle LL, Schwarzmayr T, Fadista J, Surdulescu GL, Herder C, Groves CJ, Wieland T, Bork-Jensen J, Brandslund I, Christensen C, Koistinen HA, Doney ASF, Kinnunen L, Esko T, Farmer AJ, Hakaste L, Hodgkiss D, Kravic J, Lyssenko V, Hollensted M, Jorgensen ME, Jorgensen T, Ladenvall C, Justesen JM, Käräjämäki A, Kriebel J, Rathmann W, Lannfelt L, Lauritzen T, Narisu N, Linneberg A, Melander O, Milani L, Neville M, Orho-Melander M, Qi L, Qi Q, Roden M, Rolandsson O, Swift A, Rosengren AH, Stirrups K, Wood AR, Mihailov E, Blancher C, Carneiro MO, Maguire J, Poplin R, Shakir K, Fennell T, DePristo M, De Angelis MH, Deloukas P, Gjesing AP, Jun G, Nilsson P, Murphy J, Onofrio R, Thorand B, Hansen T, Meisinger C, Hu FB, Isomaa B, Karpe F, Liang L, Peters A, Huth C, O'Rahilly SP, Palmer CNA, Pedersen O, Rauramaa R, Tuomilehto J, Salomaa V, Watanabe RM, Syvänen AC, Bergman RN, Bharadwaj D, Bottinger EP, Cho YS, Chandak GR, Chan JCN, Chia KS, Daly MJ, Ebrahim SB, Langenberg C, Elliott P, Jablonski KA, Lehman DM, Jia W, Ma RCW, Pollin TI, Sandhu M, Tandon N, Froguel P, Barroso I, Teo YY, Zeggini E, Loos RJF, Small KS, Ried JS, DeFronzo RA, Grallert H, Glaser B, Metspalu A, Wareham NJ, Walker M, Banks E, Gieger C, Ingelsson E, Im HK, Illig T, Franks PW, Buck G, Trakalo J, Buck D, Prokopenko I, Mägi R, Lind L, Farjoun Y, Owen KR, Gloyn AL, Strauch K, Tuomi T, Kooner JS, Lee JY, Park T, Donnelly P, Morris AD, Hattersley AT, Bowden DW, Collins FS, Atzmon G, Chambers JC, Spector TD, Laakso M, Strom TM, Bell GI, Blangero J, Duggirala R, Tai ES, McVean G, Hanis CL, Wilson JG, Seielstad M, Frayling TM, Meigs JB, Cox NJ, Sladek R, Lander ES, Gabriel S, Burtt NP, Mohlke KL, Meitinger T, Groop L, Abecasis G, Florez JC, Scott LJ, Morris AP, Kang HM, Boehnke M, Altshuler D, McCarthy MI. The genetic architecture of type 2 diabetes. *Nature* 2016;**536**:41–7.

7 Flannick J, Mercader JM, Fuchsberger C, Udler MS, Mahajan A, Wessel J, Teslovich TM, Caulkins L, Koesterer R, Barajas-Olmos F, Blackwell TW, Boerwinkle E, Brody JA, Centeno-Cruz F, Chen L, Chen S, Contreras-Cubas C, Córdova E, Correa A, Cortes M, DeFronzo RA, Dolan L, Drews KL, Elliott A, Floyd JS, Gabriel S, Garay-Sevilla ME, García-Ortiz H, Gross M, Han S, Heard-Costa NL, Jackson AU, Jørgensen ME, Kang HM, Kelsey M, Kim BJ, Koistinen HA, Kuusisto J, Leader JB, Linneberg A, Liu CT, Liu J, Lyssenko V, Manning AK, Marcketta A, Malacara-Hernandez JM, Martínez-Hernández A, Matsuo K, Mayer-Davis E, Mendoza-Caamal E, Mohlke KL, Morrison AC, Ndungu A, Ng MCY, O'Dushlaine C, Payne AJ, Pihoker C, Post WS, Preuss M, Psaty BM, Vasan RS, Rayner NW, Reiner AP, Revilla-Monsalve C, Robertson NR, Santoro N, Schurmann C, So WY, Soberón X, Stringham HM, Strom TM, Tam CHT, Thameem F, Tomlinson B, Torres JM, Tracy RP, van Dam RM, Vujkovic M, Wang S, Welch RP, Witte DR, Wong TY, Atzmon G, Barzilai N, Blangero J, Bonnycastle LL, Bowden DW, Chambers JC, Chan E, Cheng CY, Cho YS, Collins FS, de Vries PS, Duggirala R, Glaser B, Gonzalez C, Gonzalez ME, Groop L, Kooner JS, Kwak SH, Laakso M, Lehman DM, Nilsson P, Spector TD, Tai ES, Tuomi T, Tuomilehto J, Wilson JG, Aguilar-Salinas CA, Bottinger E, Burke B, Carey DJ, Chan JCN, Dupuis J, Frossard P, Heckbert SR, Hwang MY, Kim YJ, Kirchner HL, Lee JY, Lee J, Loos RJF, Ma RCW, Morris AD, O'Donnell CJ, Palmer CNA, Pankow J, Park KS, Rasheed A, Saleheen D, Sim X, Small KS, Teo YY, Haiman C, Hanis CL, Henderson BE, Orozco L, Tusié-Luna T, Dewey FE, Baras A, Gieger C, Meitinger T, Strauch K, Lange L, Grarup N, Hansen T, Pedersen O, Zeitler P, Dabelea D, Abecasis G, Bell GI, Cox NJ, Seielstad M, Sladek R, Meigs JB, Rich SS, Rotter JI, Altshuler D, Burtt NP, Scott LJ, Morris AP, Florez JC, McCarthy MI, Boehnke M. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* 2019;**570**:71–6.

8 Bonnefond A, Boissel M, Bolze A, Durand E, Toussaint B, Vaillant E, Gaget S, Graeve F De, Dechaume A, Allegaert F, Guilcher D Le, Yengo L, Dhennin V, Borys JM, Lu JT, Cirulli ET, Elhanan G, Roussel R, Balkau B, Marre M, Franc S, Charpentier G, Vaxillaire M, Canouil M, Washington NL, Grzymski JJ, Froguel P. Pathogenic variants in actionable MODY genes are associated with type 2 diabetes. *Nat Metab* 2020;**2**:1126–34.

9 Steinthorsdottir V, Thorleifsson G, Sulem P, Helgason H, Grarup N, Sigurdsson A, Helgadottir

HT, Johannsdottir H, Magnusson OT, Gudjonsson SA, Justesen JM, Harder MN, Jørgensen ME, Christensen C, Brandslund I, Sandbæk A, Lauritzen T, Vestergaard H, Linneberg A, Jørgensen T, Hansen T, Daneshpour MS, Fallah MS, Hreidarsson AB, Sigurdsson G, Azizi F, Benediktsson R, Masson G, Helgason A, Kong A, Gudbjartsson DF, Pedersen O, Thorsteinsdottir U, Stefansson K. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet* 2014;**46**:294–8.

10 Thomsen SK, Raimondo A, Hastoy B, Sengupta S, Dai XQ, Bautista A, Censin J, Payne AJ, Umapathysivam MM, Spigelman AF, Barrett A, Groves CJ, Beer NL, Manning Fox JE, McCarthy MI, Clark A, Mahajan A, Rorsman P, MacDonald PE, Gloyn AL. Type 2 diabetes risk alleles in PAM impact insulin release from human pancreatic β-cells. *Nat Genet* 2018;**50**:1122–31.

11 Szustakowski JD, Balasubramanian S, Sasson A, Khalid S, Bronson PG, Kvikstad E, Wong E, Liu D, Davis JW, Haefliger C, Loomis AK, Mikkilineni R, Noh HJ, Wadhawan S, Bai X, Hawes A, Krasheninina O, Ulloa R, Lopez A, Smith EN, Waring J, Whelan CD, Tsai EA, Overton J, Salerno W, Jacob H, Szalma S, Runz H, Hinkle G, Nioi P, Petrovski S, Miller MR, Baras A, Mitnaul L, Reid JG. Advancing Human Genetics Research and Drug Discovery through Exome Sequencing of the UK Biobank. *medRxiv* 2020;:2020.11.02.20222232.

12 Curtis D. Analysis of 200,000 exome-sequenced UK Biobank subjects illustrates the contribution of rare genetic variants to hyperlipidaemia. medRxiv. 2021. doi:10.1101/2021.02.10.21251503

13 McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. *Genome Biol* 2016;**17**:122.

14 Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* 2013;**7 Unit7.20**. doi:10.1002/0471142905.hg0720s76

15 Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;**4**:1073–81.

16 Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;**4**:7.

17 Galinsky KJ, Bhatia G, Loh PR, Georgiev S, Mukherjee S, Patterson NJ, Price AL. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am J Hum Genet* 2016;**98**:456–72.

18 Curtis D. Multiple Linear Regression Allows Weighted Burden Analysis of Rare Coding Variants in an Ethnically Heterogeneous Population. *Hum Hered* 2021;:1–10.

19 Curtis D. A weighted burden test using logistic regression for integrated analysis of sequence variants, copy number variants and polygenic risk score. *Eur J Hum Genet* 2019;**27**:114–24.

20 Curtis D. A rapid method for combined analysis of common and rare variants at the level of a region, gene, or pathway. *Adv Appl Bioinform Chem* 2012;**5**:1–9.

21 Curtis D. Pathway analysis of whole exome sequence data provides further support for the involvement of histone modification in the aetiology of schizophrenia. *Psychiatr Genet* 2016;**26**:223–7.

22 Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;**102**:15545–50.

23 R Core Team. *R: A language and environment for statistical computing.* Vienna, Austria.: : R Foundation for Statistical Computing 2014. http://www.r-project.org

24 Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012;**91**:224–37.

25 Bishay RH, Greenfield JR. A review of maturity onset diabetes of the young (MODY) and challenges in the management of glucokinase-MODY. *Med J Aust* 2016;**205**:480–5.

26 Naylor R, Johnson AK, Gaudio D del. Maturity-Onset Diabetes of the Young Overview. In: Adam M, Ardinger H, Pagon R, Wallace S, Bean L, Stephens K, Amemiya A, eds. *GeneReviews*. Seattle (WA): : University of Washington, Seattle 2018. https://www.ncbi.nlm.nih.gov/books/NBK500456/ (accessed 6 Jan2021).

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**:31564–73.

28 Zhou T, Ma Y, Tang J, Guo F, Dong M, Wei Q. Modulation of IGF1R Signaling Pathway by GIGYF1 in High Glucose-Induced SHSY-5Y Cells. *DNA Cell Biol* 2018;**37**:1044–54.

29 Lyssenko V, Groop L, Prasad RB. Genetics of type 2 diabetes: It matters from which parent we inherit the risk. Rev. Diabet. Stud. 2015;**12**:233–42.

30 Holt LJ, Brandon AE, Small L, Suryana E, Preston E, Wilks D, Mokbel N, Coles CA, White JD, Turner N, Daly RJ, Cooney GJ. Ablation of Grb10 Specifically in Muscle Impacts Muscle Size and Glucose Metabolism in Mice. *Endocrinology* 2018;**159**:1339–51.

31 Moreno-Navarrete JM, Rodríguez A, Ortega F, Becerril S, Girones J, Sabater-Masdeu M, Latorre J, Ricart W, Frühbeck G, Fernández-Real JM. Heme Biosynthetic Pathway is Functionally Linked to Adipogenesis via Mitochondrial Respiratory Activity. *Obesity* 2017;**25**:1723–33.

32 Majithia AR, Flannick J, Shahinian P, Guo M, Bray MA, Fontanillas P, Gabriel SB, Manning AK, Hartl C, Agarwala V, Green T, Banks E, DePristo M, Poplin R, Shakir K, Fennell T, Njølstad PR, Altshuler D, Burtt NP, Fuchsberger C, Kang HM, Sim X, Ma C, Locke AE, Blackwell T, Jackson A, Teslovich TM, Stringham H, Chines PS, Kwan P, Huyghe JR, Tan A, Jun G, Stitzel M, Bergman RN, Bonnycastle L, Tuomilehto J, Collins FS, Scott LJ, Mohlke KL, Abecasis G, Boehnke M, Strom T, Gieger C, Müller-Nurasyid M, Grallert H, Kriebel J, Ried J, De Angelis MH, Huth C, Meisinger C, Peters A, Rathmann W, Strauch K, Meitinger T, Kravic J, Algren P, Ladenvall C, Toumi T, Isomaa B, Groop L, Gaulton K, Moutsianas L, Rivas M, Pearson R, Mahajan A, Prokopenko I, Kumar A, Perry J, Howie B, Van De Bunt M, Small K, Lindgren CM, Lunter G, Robertson N, Rayner W, Morris A, Buck D, Hattersley A, Spector T, McVean G, Frayling T, Donnelly P, McCarthy M, Gupta N, Taylor H, Fox E, Newton-Cheh C, Wilson JG, O'Donnell CJ, Kathiresan S, Hirschhorn J, Seidman JG, Seidman C, Williams AL, Jacobs SBR, Moreno-Macías H, Huerta-Chagoya A, Churchhouse C, Luna CM, García-Ortíz H, Gómez-Vázquez MJ, Estrada K, Mercader JM, Ripke S, Neale B, Stram DO, Fernández-López JC, Romero-Hidalgo S, Aguilar-Delfín I, Martínez-Hernández A, Centeno-Cruz F, Mendoza-Caamal E, Monsalve CR, Islas-Andrade S, Córdova E, Rodríguez-Arellano E, Soberón X, González-Villalpando ME, Monroe K, Wilkens L, Kolonel LN, Le Marchand L, Riba L, Ordóñez-Sánchez ML, Guillén RR, Cruz-Bautista I, Rodríguez-Torres M, Muñoz-Hernández LL, Sáenz T, Gómez D, Alvirde U, Onofrio RC, Brodeur WM, Gage D, Murphy J, Franklin J, Mahan S, Ardlie K, Crenshaw AT, Winckler W, MacArthur DG, Florez JC, Haiman CA, Henderson BE, Aguilar-Salinas CA, González-Villalpando C, Orozco L, Tusié-Luna T, Almeida M, Asimit JL, Atzmon G, Barber M, Beer NL, Bell GI, Below J, Blackwell T, Blangero J, Bowden DW, Chambers J, Chen H, Chen P, Choi S, Cingolani P, Cornes BK, Cox N, Day-Williams AG, Duggirala R, Dupuis J, Dyer T, Feng S, Fernandez-Tajes J, Ferreira T, Fingerlin TE, Frayling TM, Gamazon ER, Ghosh S, Gloyn A, Grossman RL, Grundstad

J, Hanis C, Heath A, Highland H, Hirokoshi M, Huh IS, Ikram K, Jablonski KA, Kim YJ, Kato N, Kim J, King CR, Kooner J, Kwon MS, Im HK, Laakso M, Lam KKY, Lee J, Lee S, Lee S, Lehman DM, Li H, Liu X, Livne OE, Maller JB, Maxwell TJ, Mazoure A, McCarthy MI, Meigs JB, Min B, Musani S, Nagai Y, Ng MCY, Nicolae D, Oh S, Palmer N, Park T, Pollin TI, Reich D, Rivas MA, Seielstad M, Cho YS, Tai ES, Sladek R, Smith P, Tachmazidou I, Torres J, Trubetskoy V, Willems SM, Wiltshire S, Won S, Wood AR, Xu W, Teo YY, Yoon J, Lee JY, Zawistowski M, Zeggini E, Zhang W, Zöllner S, Rosen ED, Scolnick E. Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *Proc Natl Acad Sci U S A* 2014;**111**:13127–32.

33 Ben Yaou R, Hubert A, Nelson I, Dahlqvist JR, Gaist D, Streichenberger N, Beuvin M, Krahn M, Petiot P, Parisot F, Michel F, Malfatti E, Romero N, Carlier RY, Eymard B, Labrune P, Duno M, Krag T, Cerino M, Bartoli M, Bonne G, Vissing J, Laforet P, Petit FM. Clinical heterogeneity and phenotype/genotype findings in 5 families with GYG1 deficiency . *Neurol Genet* 2017;**3**:e208.

34 Rivera-León EA, Llamas-Covarrubias MA, Sánchez-Enríquez S, Martínez-López E, González-Hita M, Llamas-Covarrubias IM. Leu72Met polymorphism of GHRL gene decreases susceptibility to type 2 diabetes mellitus in a Mexican population. *BMC Endocr Disord* 2020;**20**. doi:10.1186/s12902-020-00596-3

35 Sansbury FH, Flanagan SE, Houghton JAL, Shuixian Shen FL, Al-Senani AMS, Habeb AM, Abdullah M, Kariminejad A, Ellard S, Hattersley AT. SLC2A2 mutations can cause neonatal diabetes, suggesting GLUT2 may have a role in human insulin secretion. *Diabetologia* 2012;**55**:2381–5.

36 Zhou K, Yee SW, Seiser EL, Van Leeuwen N, Tavendale R, Bennett AJ, Groves CJ, Coleman RL, Van Der Heijden AA, Beulens JW, De Keyser CE, Zaharenko L, Rotroff DM, Out M, Jablonski KA, Chen L, Javorský M, Zidzik J, Levin AM, Keoki Williams L, Dujic T, Semiz S, Kubo M, Chien HC, Maeda S, Witte JS, Wu L, Tká I, Kooy A, Van Schaik RHN, Stehouwer CDA, Logie L, Sutherland C, Klovins J, Pirags V, Hofman A, Stricker BH, Motsinger-Reif AA, Wagner MJ, Innocenti F, Hart LM, Holman RR, McCarthy MI, Hedderson MM, Palmer CNA, Florez JC, Giacomini KM, Pearson ER. Variation in the glucose transporter gene SLC2A2 is associated with glycemic response to metformin. *Nat Genet* 2016;**48**:1055–9.

37 Hu S, Ma S, Li X, Tian Z, Liang H, Yan J, Chen M, Tan H. Relationships of SLC2A4, RBP4, PCK1, and PI3K gene polymorphisms with gestational diabetes mellitus in a Chinese population. *Biomed Res Int* 2019;**2019**. doi:10.1155/2019/7398063

38 De Franco E, Saint-Martin C, Brusgaard K, Knight Johnson AE, Aguilar-Bryan L, Bowman P, Arnoux JB, Larsen AR, May S, Greeley SAW, Calzada-León R, Harman B, Houghton JAL, Nishimura-Meguro E, Laver TW, Ellard S, del Gaudio D, Christesen HT, Bellanné-Chantelot C, Flanagan SE. Update of variants identified in the pancreatic β-cell KATP channel genes KCNJ11 and ABCC8 in individuals with congenital hyperinsulinism and diabetes. *Hum Mutat* 2020;**41**:884–905.

Table 1

The table shows the weight which was assigned to each type of variant as annotated by VEP, Polyphen and SIFT as well as the broad categories which were used for multivariate analyses of variant effects [13–15]. The category referred to as "Unused" indicates variant annotations which were not assigned to any variants and/or were not included in the multivariate analyses.

VEP / SIFT / Polyphen annotation	Weight	Category
intergenic_variant	0	Unused
feature_truncation	0	Intronic, etc.
regulatory_region_variant	0	Intronic, etc.
feature_elongation	0	Intronic, etc.
regulatory_region_amplification	1	Intronic, etc.
regulatory_region_ablation	1	Intronic, etc.
TF_binding_site_variant	1	Intronic, etc.
TFBS_amplification	1	Intronic, etc.
TFBS_ablation	1	Intronic, etc.
downstream_gene_variant	0	Intronic, etc.
upstream_gene_variant	0	Intronic, etc.
non_coding_transcript_variant	0	Intronic, etc.
NMD_transcript_variant	0	Intronic, etc.
intron_variant	0	Intronic, etc.
non_coding_transcript_exon_variant	0	Intronic, etc.
3_prime_UTR_variant	1	3 prime UTR
5_prime_UTR_variant	1	5 prime UTR
mature_miRNA_variant	5	Unused
coding_sequence_variant	0	Unused
synonymous_variant	0	Synonymous
stop_retained_variant	5	Unused
incomplete_terminal_codon_variant	5	Unused
splice_region_variant	1	Splice region
protein_altering_variant	5	Protein altering
missense_variant	5	Protein altering
inframe_deletion	10	InDel, etc
inframe_insertion	10	InDel, etc
transcript_amplification	10	InDel, etc
start_lost	10	Unused
stop_lost	10	Unused
frameshift_variant	100	Disruptive
stop_gained	100	Disruptive
splice_donor_variant	100	Splice site variant
splice_acceptor_variant	100	Splice site variant
transcript_ablation	100	Disruptive
SIFT deleterious	20	Deleterious
PolyPhen possibly damaging	5	Possibly damaging
PolyPhen probably damaging	10	Probably damaging

Table 2

Genes with absolute value of SLP exceeding 3 or more (equivalent to p<0.001) for test of association of weighted burden score with T2D.

Symbol	SLP	Name
GCK	22.25	Glucokinase
HNF4A	6.82	Hepatocyte Nuclear Factor 4 Alpha
GIGYF1	6.22	GRB10 Interacting GYF Protein 1
ZNF620	3.78	Zinc Finger Protein 620
RAI2	3.74	Retinoic Acid Induced 2
TM4SF20	3.65	Transmembrane 4 L Six Family Member 20
ALAD	3.63	Aminolevulinate Dehydratase
PPARG	3.45	Peroxisome Proliferator Activated Receptor Gamma
LOC105370752	3.42	Uncharacterized LOC105370752
KLHL11	3.35	Kelch Like Family Member 11
HMGXB4	3.35	HMG-Box Containing 4
MIR6825	3.31	MicroRNA 6825
TAZ	3.30	Tafazzin
WDR33	3.25	WD Repeat Domain 33
HECTD1	3.24	HECT Domain E3 Ubiquitin Protein Ligase 1
ZNF571-AS1	3.23	ZNF571 Antisense RNA 1
GYG1	3.22	Glycogenin 1
APTX	3.20	Aprataxin
KCNK15	3.19	Potassium Two Pore Domain Channel Subfamily K Member 15
XPO1	3.19	Exportin 1
PKD1	3.10	Polycystin 1, Transient Receptor Potential Channel Interacting
ZNF763	-3.01	Zinc Finger Protein 763
COA5	-3.05	Cytochrome C Oxidase Assembly Factor 5
GHRL	-3.15	Ghrelin And Obestatin Prepropeptide
DEUP1	-3.20	Deuterosome Assembly Protein 1
C7orf50	-3.22	Chromosome 7 Open Reading Frame 50
MFSD12	-3.34	Major Facilitator Superfamily Domain Containing 12
C19orf73	-3.34	Chromosome 19 Open Reading Frame 73
ATXN1L	-3.35	Ataxin 1 Like
EML4	-3.58	EMAP Like 4
DLEC1	-3.72	DLEC1 Cilia And Flagella Associated Protein
RPS5	-3.76	Ribosomal Protein S5

Table 3

Results from logistic regression analysis showing the effects on risk of T2D of different categories of variant within exome-wide significant genes and genes of interest with absolute value of gene-wise SLP > 3. Odds ratios for each category are estimated including principal components and sex as covariates.

Table 3A

Results for *GCK*.

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR	SLP
Intronic, etc	9531	0.051051	1034	0.074186	1.04 (0.97 - 1.12)	0.55
5 prime UTR	940	0.005035	137	0.009829	1.08 (0.89 - 1.31)	0.39
Synonymous	4550	0.024371	333	0.023892	0.91 (0.81 - 1.02)	-0.98
Splice region	1810	0.009695	132	0.009471	1.04 (0.87 - 1.25)	-0.20
3 prime UTR	1533	0.008211	100	0.007175	0.92 (0.75 - 1.13)	-0.37
Protein altering	744	0.003985	103	0.007390	1.05 (0.77 - 1.41)	0.12
InDel, etc	4	0.000021	2	0.000143	7.27 (1.26 - 41.81)	1.22
Disruptive	3	0.000016	11	0.000789	61.80 (16.62 - 229.79)	10.27
Splice site variant	2	0.000011	6	0.000430	36.48 (6.79 - 196.08)	5.56
Deleterious	300	0.001607	47	0.003372	1.23 (0.69 - 2.16)	0.32
Possibly damaging	89	0.000477	10	0.000717	1.14 (0.52 - 2.51)	0.13
Probably damaging	129	0.000691	33	0.002368	2.97 (1.59 - 5.54)	3.32

Table 3B

Results for *HNF4A.*

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR	SLP
Intronic, etc	11703	0.062685	1181	0.084732	1.03 (0.97 - 1.09)	0.42
5 prime UTR	275	0.001473	17	0.001220	0.84 (0.51 - 1.39)	-0.31
Synonymous	5140	0.027532	429	0.030779	1.01 (0.91 - 1.12)	0.07
Splice region	1159	0.006208	115	0.008251	0.89 (0.72 - 1.10)	0.55
3 prime UTR	460	0.002464	53	0.003803	0.99 (0.74 - 1.34)	0.02
Protein altering	1326	0.007103	163	0.011695	0.92 (0.72 - 1.18)	0.29
InDel, etc	3	0.000016	0	0.000000		0.00
Disruptive	6	0.000032	0	0.000000		0.00
Splice site variant	0	0.000000	0	0.000000		0.00
Deleterious	541	0.002898	81	0.005811	1.46 (0.85 - 2.50)	0.80
Possibly damaging	290	0.001553	35	0.002511	1.30 (0.72 - 2.36)	0.42
Probably damaging	117	0.000627	34	0.002439	2.97 (1.61 - 5.50)	3.41

Table 3C

Results for *GIGYF1.*

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR	SLP
Intronic, etc	23552	0.126153	1950	0.139905	0.98 (0.94 - 1.03)	0.44
5 prime UTR	163	0.000873	20	0.001435	1.02 (0.63 - 1.66)	0.03
Synonymous	4794	0.025678	419	0.030062	0.99 (0.89 - 1.10)	0.08
Splice region	3394	0.018179	352	0.025255	1.04 (0.94 - 1.15)	0.33
3 prime UTR	2102	0.011259	149	0.010690	0.88 (0.74 - 1.04)	-0.88
Protein altering	5653	0.030279	463	0.033219	1.05 (0.89 - 1.24)	0.27
InDel, etc	489	0.002619	38	0.002726	1.11 (0.79 - 1.55)	0.26
Disruptive	35	0.000187	16	0.001148	5.65 (3.07 - 10.40)	7.85
Splice site variant	8	0.000043	6	0.000430	7.70 (2.62 - 22.67)	3.68
Deleterious	1365	0.007311	114	0.008179	0.95 (0.75 - 1.20)	0.17
Possibly damaging	2347	0.012571	175	0.012556	1.03 (0.82 - 1.29)	-0.10
Probably damaging	1237	0.006626	111	0.007964	1.14 (0.87 - 1.47)	0.48

Table 3D

Results for *ALAD.*

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR	SLP
Intronic, etc	11859	0.063521	974	0.069881	1.00 (0.93 - 1.07)	0.02
5 prime UTR	160	0.000857	22	0.001578	1.30 (0.81 - 2.08)	0.58
Synonymous	3296	0.017655	251	0.018008	0.91 (0.79 - 1.04)	0.83
Splice region	188	0.001007	26	0.001865	1.37 (0.89 - 2.10)	0.85
3 prime UTR	149	0.000798	12	0.000861	1.04 (0.57 - 1.90)	0.05
Protein altering	939	0.005030	87	0.006242	0.94 (0.67 - 1.32)	0.15
InDel, etc	0	0.000000	0	0.000000		0.00
Disruptive	6	0.000032	3	0.000215	6.67 (1.55 - 28.73)	1.69
Splice site variant	3	0.000016	2	0.000143	8.31 (1.26 - 55.06)	1.38
Deleterious	407	0.002180	49	0.003516	1.60 (0.80 - 3.18)	0.76
Possibly damaging	191	0.001023	25	0.001794	1.20 (0.58 - 2.50)	0.21
Probably damaging	177	0.000948	18	0.001291	0.82 (0.37 - 1.82)	0.21

Table 3E

Results for *PPARG.*

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR	SLP
Intronic, etc	5096	0.027296	506	0.036304	0.97 (0.88 - 1.07)	0.27
5 prime UTR	119	0.000637	26	0.001865	1.17 (0.75 - 1.83)	0.33
Synonymous	3003	0.016085	231	0.016573	0.96 (0.84 - 1.11)	0.23
Splice region	95	0.000509	7	0.000502	0.94 (0.43 - 2.08)	-0.05
3 prime UTR	174	0.000932	14	0.001004	1.07 (0.61 - 1.88)	0.10
Protein altering	454	0.002432	49	0.003516	1.20 (0.79 - 1.81)	0.42
InDel, etc	1	0.000005	0	0.000000		0.00
Disruptive	7	0.000037	4	0.000287	8.23 (2.29 - 29.63)	2.29
Splice site variant	0	0.000000	1	0.000072		1.16
Deleterious	138	0.000739	15	0.001076	1.24 (0.58 - 2.66)	0.24
Possibly damaging	46	0.000246	6	0.000430	1.25 (0.47 - 3.31)	0.19
Probably damaging	131	0.000702	13	0.000933	0.92 (0.41 - 2.06)	0.08

Table 3F

Results for *GYG1.*

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR	SLP
Intronic, etc	8535	0.045717	787	0.056464	0.98 (0.91 - 1.07)	0.17
5 prime UTR	3285	0.017596	250	0.017937	1.04 (0.91 - 1.18)	0.24
Synonymous	1348	0.007220	184	0.013201	0.94 (0.78 - 1.14)	0.28
Splice region	270	0.001446	37	0.002655	1.58 (1.11 - 2.27)	1.99
3 prime UTR	1457	0.007804	118	0.008466	0.91 (0.75 - 1.10)	0.51
Protein altering	3509	0.018795	313	0.022457	1.10 (0.91 - 1.32)	0.51
InDel, etc	5	0.000027	3	0.000215	3.70 (0.83 - 16.43)	1.84
Disruptive	203	0.001087	33	0.002368	1.98 (1.40 - 2.81)	4.03
Splice site variant	15	0.000080	4	0.000287	3.08 (0.96 - 9.81)	1.41
Deleterious	1934	0.010359	158	0.011336	0.90 (0.63 - 1.30)	0.24
Possibly damaging	157	0.000841	28	0.002009	0.95 (0.60 - 1.50)	0.09
Probably damaging	1611	0.008629	114	0.008179	1.00 (0.69 - 1.47)	-0.01

Table 3G

Results for *GHRL.*

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR	SLP
Intronic, etc	6647	0.035604	621	0.044554	0.99 (0.91 - 1.08)	0.07
5 prime UTR	6997	0.037478	621	0.044554	1.02 (0.94 - 1.11)	0.19
Synonymous	258	0.001382	24	0.001722	1.06 (0.68 - 1.63)	0.10
Splice region	3237	0.017339	230	0.016502	0.98 (0.86 - 1.13)	-0.10
3 prime UTR	268	0.001436	13	0.000933	0.65 (0.37 - 1.15)	-0.89
Protein altering	2500	0.013391	140	0.010044	0.82 (0.45 - 1.49)	-0.30
InDel, etc	32	0.000171	1	0.000072	0.44 (0.06 - 3.35)	-0.14
Disruptive	50	0.000268	1	0.000072	0.23 (0.03 - 1.77)	-0.82
Splice site variant	11	0.000059	0	0.000000		0.00
Deleterious	103	0.000552	6	0.000430	0.60 (0.22 - 1.66)	-0.50
Possibly damaging	47	0.000252	6	0.000430	2.14 (0.64 - 7.14)	0.69
Probably damaging	2279	0.012207	122	0.008753	0.94 (0.50 - 1.76)	-0.07

Table 4

Results from logistic regression analysis showing effect on risk of T2D of different categories of variant within genes previously implicated in diabetes pathogenesis.

Table 4A

Results for *HNF1A.*

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR	SLP
Intronic, etc	8981	0.048105	798	0.057254	0.93 (0.86 - 1.00)	1.24
5 prime UTR	413	0.002212	54	0.003874	1.26 (0.94 - 1.69)	0.93
Synonymous	1561	0.008361	157	0.011264	1.05 (0.89 - 1.24)	0.27
Splice region	624	0.003342	87	0.006242	1.03 (0.81 - 1.31)	0.08
3 prime UTR	330	0.001768	30	0.002152	0.78 (0.53 - 1.16)	0.66
Protein altering	2806	0.015030	268	0.019228	1.07 (0.90 - 1.28)	0.38
InDel, etc	8	0.000043	1	0.000072	1.96 (0.23 - 16.51)	0.32
Disruptive	29	0.000155	7	0.000502	3.27 (1.39 - 7.70)	1.96
Splice site variant	2	0.000011	1	0.000072	7.91 (0.67 - 93.47)	0.71
Deleterious	1274	0.006824	107	0.007677	0.87 (0.64 - 1.20)	0.41
Possibly damaging	430	0.002303	39	0.002798	1.07 (0.72 - 1.58)	0.13
Probably damaging	449	0.002405	43	0.003085	1.27 (0.83 - 1.94)	0.58

Table 4B

Results for *HNF1B.*

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR	SLP
Intronic, etc	9825	0.052626	1188	0.085235	1.03 (0.97 - 1.09)	0.43
5 prime UTR	2883	0.015442	259	0.018582	1.08 (0.94 - 1.23)	0.57
Synonymous	848	0.004542	78	0.005596	1.02 (0.80 - 1.29)	0.05
Splice region	140	0.000750	14	0.001004	1.44 (0.82 - 2.54)	0.72
3 prime UTR	1116	0.005978	145	0.010403	0.89 (0.74 - 1.08)	0.65
Protein altering	1734	0.009288	184	0.013201	1.19 (0.96 - 1.47)	1.00
InDel, etc	7	0.000037	1	0.000072	2.04 (0.24 - 17.58)	0.36
Disruptive	2	0.000011	1	0.000072	6.79 (0.57 - 81.16)	0.71
Splice site variant	0	0.000000	0	0.000000		0.00
Deleterious	1046	0.005603	60	0.004305	0.58 (0.33 - 1.00)	-1.34
Possibly damaging	591	0.003166	32	0.002296	1.00 (0.55 - 1.82)	0.00
Probably damaging	445	0.002384	33	0.002368	1.39 (0.76 - 2.53)	-0.56

Table 4C

Results for *ABCC8.*

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR	SLP
Intronic, etc	40619	0.217570	3992	0.286411	1.02 (0.99 - 1.05)	0.84
5 prime UTR	3981	0.021324	443	0.031784	1.06 (0.96 - 1.17)	0.65
Synonymous	6894	0.036927	686	0.049218	0.98 (0.90 - 1.06)	0.26
Splice region	1489	0.007976	115	0.008251	0.92 (0.76 - 1.12)	0.41
3 prime UTR	117	0.000627	6	0.000430	0.62 (0.27 - 1.45)	-0.58
Protein altering	5651	0.030269	521	0.037380	1.07 (0.93 - 1.23)	0.49
InDel, etc	13	0.000070	1	0.000072	0.90 (0.11 - 7.17)	0.00
Disruptive	103	0.000552	9	0.000646	1.19 (0.59 - 2.40)	0.21
Splice site variant	58	0.000311	4	0.000287	0.83 (0.29 - 2.37)	-0.14
Deleterious	1861	0.009968	189	0.013560	1.17 (0.96 - 1.43)	0.97
Possibly damaging	469	0.002512	38	0.002726	0.70 (0.48 - 1.02)	1.26
Probably damaging	2290	0.012266	215	0.015425	1.10 (0.90 - 1.33)	0.46

Table 4D

Results for *INSR.*

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR	SLP
Intronic, etc	36762	0.196910	3725	0.267255	0.99 (0.96 - 1.02)	0.31
5 prime UTR	165	0.000884	15	0.001076	0.86 (0.50 - 1.49)	0.23
Synonymous	15478	0.082906	1385	0.099369	1.02 (0.97 - 1.07)	0.29
Splice region	4126	0.022100	387	0.027766	0.96 (0.82 - 1.13)	0.22
3 prime UTR	400	0.002143	48	0.003444	1.24 (0.91 - 1.70)	0.78
Protein altering	8249	0.044185	603	0.043263	0.89 (0.79 - 1.00)	-1.33
InDel, etc	18	0.000096	2	0.000143	1.41 (0.49 - 4.07)	0.19
Disruptive	45	0.000241	5	0.000359	1.62 (0.63 - 4.19)	0.51
Splice site variant	16	0.000086	1	0.000072	0.64 (0.08 - 5.12)	0.00
Deleterious	3668	0.019647	292	0.020950	1.21 (0.87 - 1.67)	0.61
Possibly damaging	3181	0.017039	249	0.017865	0.90 (0.65 - 1.25)	0.27
Probably damaging	357	0.001912	26	0.001865	0.98 (0.59 - 1.62)	-0.03

Table 4E

Results for *MC4R.*

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR	SLP
Intronic, etc	0	0.000000	0	0.000000		0.00
5 prime UTR	265	0.001419	25	0.001794	1.28 (0.84 - 1.96)	0.63
Synonymous	799	0.004280	93	0.006672	0.97 (0.77 - 1.21)	0.12
Splice region	0	0.000000	0	0.000000		0.00
3 prime UTR	45	0.000241	4	0.000287	1.26 (0.44 - 3.61)	0.11
Protein altering	1248	0.006685	117	0.008394	0.91 (0.69 - 1.22)	0.27
InDel, etc	3	0.000016	1	0.000072	4.85 (0.48 - 49.39)	0.60
Disruptive	72	0.000386	8	0.000574	1.47 (0.69 - 3.12)	0.51
Splice site variant	0	0.000000	0	0.000000		0.00
Deleterious	412	0.002207	44	0.003157	1.22 (0.61 - 2.44)	0.25
Possibly damaging	181	0.000970	22	0.001578	1.39 (0.78 - 2.49)	0.59
Probably damaging	389	0.002084	39	0.002798	1.28 (0.60 - 2.72)	0.29

Table 4F

Results for *SLC30A8.*

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR	SLP
Intronic, etc	7313	0.039171	607	0.043550	0.93 (0.85 - 1.02)	0.88
5 prime UTR	307	0.001644	36	0.002583	1.21 (0.85 - 1.74)	0.55
Synonymous	509	0.002726	65	0.004664	1.40 (1.07 - 1.84)	1.92
Splice region	351	0.001880	40	0.002870	1.12 (0.80 - 1.58)	0.31
3 prime UTR	2262	0.012116	197	0.014134	1.04 (0.88 - 1.23)	0.18
Protein altering	2666	0.014280	257	0.018439	0.96 (0.80 - 1.14)	0.19
InDel, etc	3	0.000016	0	0.000000		0.00
Disruptive	99	0.000530	9	0.000646	0.82 (0.40 - 1.68)	0.23
Splice site variant	59	0.000316	1	0.000072	0.24 (0.03 - 1.77)	-0.82
Deleterious	616	0.003300	30	0.002152	0.61 (0.29 - 1.30)	-0.72
Possibly damaging	455	0.002437	26	0.001865	0.89 (0.57 - 1.39)	-0.23
Probably damaging	441	0.002362	22	0.001578	0.98 (0.41 - 2.34)	-0.02

Table 4G

Results for *PAM.*

Category	Total count in controls	Mean count in controls	Total count in cases	Mean count in cases	OR	SLP
Intronic, etc	12269	0.065717	1169	0.083871	0.97 (0.91 - 1.03)	0.55
5 prime UTR	563	0.003016	49	0.003516	0.93 (0.69 - 1.26)	0.19
Synonymous	2051	0.010986	203	0.014564	1.11 (0.95 - 1.29)	0.78
Splice region	733	0.003926	85	0.006098	0.94 (0.74 - 1.19)	0.23
3 prime UTR	121	0.000648	8	0.000574	0.92 (0.44 - 1.92)	-0.09
Protein altering	6968	0.037323	654	0.046922	1.09 (0.91 - 1.30)	0.47
InDel, etc	199	0.001066	19	0.001363	1.20 (0.74 - 1.95)	0.34
Disruptive	123	0.000659	4	0.000287	0.46 (0.16 - 1.26)	-0.91
Splice site variant	33	0.000177	2	0.000143	0.78 (0.18 - 3.35)	0.00
Deleterious	5168	0.027682	486	0.034869	1.11 (0.81 - 1.50)	0.29
Possibly damaging	459	0.002459	41	0.002942	0.89 (0.59 - 1.35)	0.23
Probably damaging	4890	0.026193	457	0.032788	1.12 (0.81 - 1.54)	0.31

Figure 1

QQ plot of signed log10 p values (SLPs) obtained for weighted burden analysis of association with T2D showing observed against expected SLP for each gene, omitting results for *GCK*, which has SLP = 22.25.