ARTICLE

OPEN

https://doi.org/10.1038/s41586-019-1231-2

Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls

A list of authors and their affiliations appears in the online version of the paper

Protein-coding genetic variants that strongly affect disease risk can yield relevant clues to disease pathogenesis. Here we report exome-sequencing analyses of 20,791 individuals with type 2 diabetes (T2D) and 24,440 non-diabetic control participants from 5 ancestries. We identify gene-level associations of rare variants (with minor allele frequencies of less than 0.5%) in 4 genes at exome-wide significance, including a series of more than 30 *SLC30A8* alleles that conveys protection against T2D, and in 12 gene sets, including those corresponding to T2D drug targets ($P = 6.1 \times 10^{-3}$) and candidate genes from knockout mice ($P = 5.2 \times 10^{-3}$). Within our study, the strongest T2D gene-level signals for rare variants explain at most 25% of the heritability of the strongest common single-variant signals, and the gene-level effect sizes of the rare variants that we observed in established T2D drug targets will require 75,000–185,000 sequenced cases to achieve exome-wide significance. We propose a method to interpret these modest rare-variant associations and to incorporate these associations into future target or gene prioritization efforts.

Human genetics offers a powerful approach for better understanding and treating disease by identifying molecular alterations that are causally associated with physiological traits[1]. Common-variant array-based genome-wide association studies (GWAS) have associated thousands of genomic loci with hundreds of human traits[2], and further analyses indicate that heritability of most complex traits is attributable to modest-effect common regulatory variants[3]. However, non-coding GWAS associations are challenging to assign to causal variants or genes[4].

Protein-coding variants with strong effects on protein function or disease can offer molecular 'probes' into the pathological relevance of a gene[5] and potentially establish a direct causal link[6] between gene gain- or loss-of-function and disease risk[7]—especially when there is evidence of multiple independent variant associations (an 'allelic series') within a gene[8]. Several lines of evidence[9] predict that strong-effect variants (allelic odds ratios > 2) will usually be rare (minor allele frequency (MAF) < 0.5%) and, in many cases, difficult to accurately study through current array-based GWAS and imputation strategies[5]. Whole-genome or whole-exome sequencing, by contrast, allows interrogation of the full spectrum of genetic variation.

Previous exome-sequencing studies have identified relatively few exome-wide significant rare-variant associations for complex diseases such as T2D[10]. This paucity of findings is in part due to the limited sample sizes of previous studies, the largest of which included less than 10,000 disease cases and fall short of the sample sizes that analytic[9] and simulation-based calculations[11] predict are needed to identify rare disease-associated variants under plausible disease models. To increase rare coding variant analysis power, we collected and analysed exome-sequencing data from 20,791 T2D cases and 24,440 controls—one of the largest analyses of exome-sequenced cases for T2D, specifically, and for any disease, more generally.

Genetic discovery from association analysis

Study participants (Supplementary Table 1) were drawn from five self-reported ancestries: (Hispanic/Latino (effective size (n_{eff}) = 14,442; 33.8%), European (n_{eff} = 10,517; 24.6%), African-American (n_{eff} = 5,959; 13.9%), East-Asian (n_{eff} = 6,010; 14.1%) and South-Asian (n_{eff} = 5,833; 13.6%)) and yielded equivalent statistical power to detect associations as a balanced study of around 42,800 individuals or a population-based study (assuming T2D prevalence of 8% and no ascertainment bias) of around 152,000 individuals. Power was improved compared to the previous largest T2D exome-sequencing study[10] of 6,504 cases and 6,436 controls, increasing, for example, from 5% to 90% for a variant with MAF = 0.2% and odds ratio = 2.5 (Extended Data Fig. 1).

Exome sequencing to 40x mean depth, variant calling and quality control (Extended Data Fig. 2, Supplementary Methods, Supplementary Figs. 1–3 and Supplementary Table 2) produced a dataset with 6.33 million variants: 2.3% common (MAF > 5%), 4.2% low-frequency (0.5% < MAF < 5%) and 93.5% rare (MAF < 0.5%) (Supplementary Table 3). These include 2.26 million nonsynonymous variants and 871,000 insertions and deletions (indels), more than twice the number of variants that were analysed in a previous T2D exome-sequencing study[10].

We first tested each variant, regardless of allele frequency, for T2D association ('single-variant' test; Methods and Extended Data Figs. 3, 4). Fifteen variants (in seven loci) exceeded exome-wide significance ($P < 4.3 \times 10^{-7}$ for coding variants[12], $P < 5 \times 10^{-8}$ for synonymous or non-coding variants), including ten nonsynonymous variants (Fig. 1a and Extended Data Table 1). These 15 associations are a substantial increase over the single association that was reported in a previous T2D-exome sequencing study[10] and illustrate again the value of multi-ancestry association analyses[13]—as only 9 out of 15 variants achieved $P < 0.05$ in European samples. However, only two variants were not previously reported by GWAS: a variant in *SFI1* (rs145181683, Arg724Trp; Supplementary Fig. 4) that failed to replicate in an independent cohort ($n = 4{,}522$, $P = 0.90$; Methods) and a low-frequency (in Hispanic/Latino individuals; MAF = 0.89%) moderate-effect (odds ratio = 2.17, 95% confidence interval = 1.63–2.89) *MC4R* variant (rs79783591, Ile269Asn) that has previously been shown to decrease MC4R activity and to be associated with obesity and T2D in smaller studies[14]. Conditioning on body-mass index reduced but did not eliminate the *MC4R* Ile269Asn T2D association ($P = 1.0 \times 10^{-5}$).

Because single-variant analyses have limited power to detect rare-variant associations[9], we next performed association tests for aggregations of variants within genes. Because numerous variant aggregation approaches (that is, 'masks') and gene-level tests are available, we developed a method (Methods, Extended Data Figs. 5, 6 and Supplementary Figs. 5, 6) to consolidate information across 14 analyses

into four results per gene: burden[9] and SKAT[15] analyses, each of which were either summarized as the 'minimum P value' across masks or 'weighted' to estimate the effect of gene haploinsufficiency. We used an exome-wide gene-level significance threshold of $P = 6.57 \times 10^{-7}$ (Methods).

Using this strategy, gene-level associations were exome-wide significant for *MC4R*, *SLC30A8* and *PAM* (Fig. 1b, Extended Data Table 2 and Supplementary Table 4), with variants from multiple ancestries contributing to each signal (Methods). All three genes lie within reported T2D GWAS loci and contain previously identified coding variant signals: the common variant Arg325Trp and 12 rare protective protein-truncating variants (PTVs) for *SLC30A8*[7,16], the low-frequency variants Asp563Gly and Ser539Trp for *PAM*[10,17] and the low-frequency variant Ile269Asn for *MC4R*.

The associations in *MC4R* (combined MAF = 0.79%, minimum $P = 2.7 \times 10^{-10}$, odds ratio = 2.07, 95% confidence interval = 1.65–2.59) and *PAM* (combined MAF = 4.9%, weighted $P = 2.2 \times 10^{-9}$, odds ratio = 1.44, 95% confidence interval = 1.28–1.62) result largely from effects of the previously identified coding variants in these genes, although the *MC4R* signal remained nominally significant after removing Ile269Asn ($P = 8.6 \times 10^{-3}$; Supplementary Fig. 7) and the *PAM* signal remained nominally significant ($P < 0.05$) after removing the 35 strongest individually associated *PAM* variants (Supplementary Fig. 8). As illustrated by a recent study that identified a novel T2D risk mechanism through cellular characterization of PAM Asp563Gly and Ser539Trp[18], variants identified in our study (uniquely from sequencing)[6] could yield further insights into the T2D risk mechanism mediated by *PAM*.

In contrast to *MC4R* and *PAM*, the *SLC30A8* signal (103 variants, combined MAF = 1.4%, weighted $P = 1.3 \times 10^{-8}$, odds ratio = 0.40, 95% confidence interval = 0.28–0.55) was not primarily driven by an individual variant (Arg325Trp (MAF > 1%) was not included in the gene-level analysis). The association was instead driven by 90 missense variants (weighted $P = 3.9 \times 10^{-7}$) and remained nominally significant ($P < 0.05$) even when we removed the 32 strongest individually associated *SLC30A8* variants (Fig. 1c and Supplementary Fig. 9). Although *SLC30A8* was first implicated in T2D over a decade ago[16], the disease-associated molecular mechanism(s) through which SLC30A8 acts remain poorly understood[19]—in part because the common risk-increasing allele Arg325Trp and the rare risk-decreasing PTVs were both initially thought to decrease protein activity[7,19]. The protective allelic series from our analysis argues that decreased T2D risk is the typical effect of *SLC30A8* missense variation—that is, it is not unique to haploinsufficiency—and provides many additional alleles that can be characterized to gain mechanistic insights.

To evaluate association evidence for genes other than *MC4R*, *PAM* and *SLC30A8*, we assessed the 50 most-significant gene-level associations from our study in two independent exome-sequencing datasets: 12,467 European or African-American individuals (3,062 T2D cases) from the CHARGE discovery sequencing project[20] (Supplementary Table 5; 50 genes available) and 49,199 European individuals (12,973 T2D cases) from the Geisinger Health System (Supplementary Table 6; 44 genes available). In a meta-analysis of the three studies (Methods and Supplementary Table 7), *MC4R* ($P = 6.9 \times 10^{-14}$), *PAM* ($P = 3.0 \times 10^{-9}$) and *SLC30A8* ($P = 3.3 \times 10^{-8}$) each became more significant. In addition, one gene, *UBE2NL* ($P = 5.6 \times 10^{-7}$)—which has few prior links to T2D or other complex traits—newly achieved exome-wide significance (http://www.type2diabetesgenetics.org/). All aspects of this association passed quality control (Methods and Supplementary Table 8), although further replication will be important to establish *UBE2NL* as a novel T2D-relevant gene.

More broadly, we observed an excess of directionally consistent associations (both odds ratio > 1 or both odds ratio < 1) between the original and replication analyses (31 out of 46 in CHARGE, one-sided binomial $P = 0.013$; 23 out of 40 in the Geisinger Health System, $P = 0.21$; overall $P = 0.011$; Supplementary Table 7), suggesting that several more of our top gene-level signals will reach exome-wide significance in future studies.

Further insights from gene-level analyses

Even if a gene-level association does not achieve exome-wide significance, it might still be of use to prioritize a gene as relevant to T2D[8] or predict whether loss or gain of protein function increases disease risk[7]. To investigate potential insights that could be obtained by sub-exome-wide significant gene-level associations, we analysed 16 gene sets that were connected to T2D based on a variety of sources of evidence (for example, genes that contained diabetes-associated Mendelian variants, T2D drug targets[21] or genes that have been implicated in diabetes-related phenotypes in mouse models[22]; Methods and Supplementary Table 9).

First, for each gene set, we investigated whether the genes within the set had more significant gene-level associations than expected by chance (Methods). In total, 12 out of 16 gene sets achieved $P < 0.05$ set-level associations (Fig. 2a–e and Supplementary Fig. 10), including T2D drug targets ($P = 2.1 \times 10^{-3}$), genes previously reported in mouse models of non-insulin-dependent diabetes (NIDD; $P = 5.2 \times 10^{-3}$) or impaired glucose tolerance ($P = 7.2 \times 10^{-6}$) and genes that contained common likely causal coding-variant T2D associations[6] ($P = 8.8 \times 10^{-3}$ after conditioning on the common variants nearby these genes). Additionally, as previously described[10], we observed a significant set-level association ($P = 1.2 \times 10^{-3}$) for genes implicated in maturity onset diabetes of the young (MODY; Fig. 2a, Supplementary Table 10), with nominal associations in four genes including *PDX1* (weighted $P = 1.7 \times 10^{-4}$, odds ratio = 3.45, 95% confidence interval = 1.78–6.71, 65 variants). Rare variants in genes associated with MODY also demonstrated aggregate association with lower body-mass index (minimum $P = 5.7 \times 10^{-3}$) and lower fasting insulin (minimum $P = 0.028$), consistent with the known predominant variant risk mechanism of reduced insulin secretion in MODY[23]. Most gene set signals were driven by multiple genes in the set (Supplementary Table 11) and—compared with previous studies focused on PTVs[24]—consisted of substantial contributions from missense variants. Indeed, set-level P values from PTVs alone were >0.05 for almost all gene sets (Supplementary Fig. 11).

Collectively, these results suggest that association strength at the gene level can be used as a potential metric to prioritize candidate genes relevant to T2D. For example, the set of 40 genes within T2D GWAS loci with gene-level $P < 0.05$ had a significant excess of protein–protein interactions among them (Methods and Supplementary Table 12), suggesting that this set may be enriched for 'effector genes' that mediate T2D GWAS associations[6]. Fully evaluating the relevance to T2D of these and other candidate genes will require further experimental work[4].

In addition to prioritizing genes that are potentially relevant to T2D, we assessed whether gene-level analysis could help to predict whether gene inactivation increases or decreases T2D risk, as this is of high interest for the development of therapeutics[8]. We compared the odds ratios that were estimated from a gene-level weighted burden analysis to directional relationships that have been previously reported (Methods). Seven out of eight T2D drug targets showed concordance between genetic and therapeutic directions of effect (three out of four inhibitor targets had an odds ratio < 1, four out of four agonist targets had an odds ratio > 1; one-sided binomial $P = 0.035$; Fig. 2f). The only exception was *KCNJ11* (odds ratio = 1.59, inhibited by sulfonylureas), for which the gene-level signal was driven by a known[25] activating missense mutation (His172Arg); an analysis without this variant predicted the correct (odds ratio < 1) directional relationship. This finding is consistent with the known reciprocal roles of *KCNJ11* in both diabetes and persistent hyperinsulinaemic hypoglycaemia of infancy.

Concordances between gene-level estimates of odds ratios and knockout effects in mice were more equivocal (for example, 7 out of 11 diabetes-associated genes had an odds ratio > 1, binomial $P = 0.27$; 137 out of 240 genes associated with increased circulating glucose had an odds ratio > 1, $P = 0.016$; Supplementary Fig. 12). The lower concordances for these gene sets, despite a trend towards lower-than-expected gene-level P values within them (Supplementary Fig. 10), highlight the

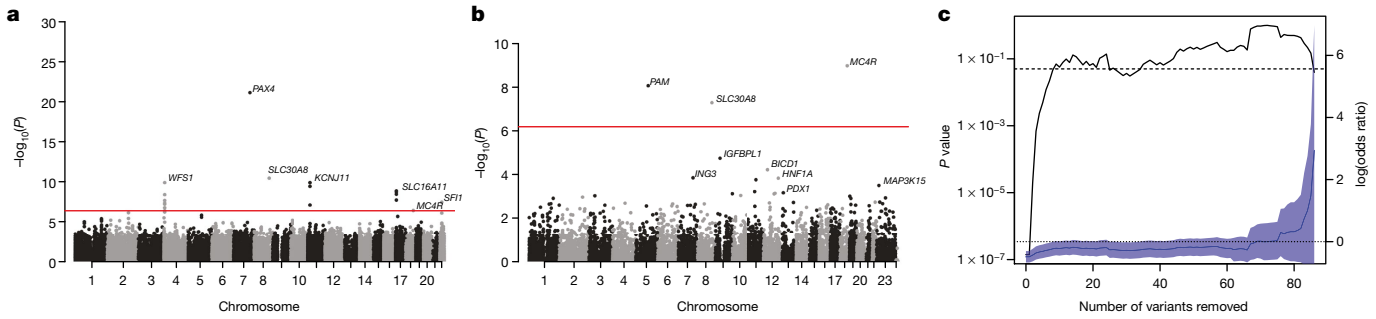


Fig. 1 | Exome-wide association analysis. a, Single-variant associations were calculated using the two-sided EMMAX test. Red line, $P = 4.3 \times 10^{-7}$. **b**, Gene-level *P* values, corrected for four analyses performed using the two-sided minimum *P*-value test. The most-significant genes are labelled. Red line, $P = 6.5 \times 10^{-7}$. **c**, *SLC30A8* gene-level *P* values (left *y* axis, black line), calculated using a two-sided burden test after removing variants (in order of increasing single-variant *P* value) from the 1/5 1% mask (the strongest signal for *SLC30A8*). Dashed line, $P = 0.05$. Right *y* axis (blue line), estimated effect size ($\log_{10}$(odds ratio)). Blue shading, 95% confidence interval. Dotted line, effect size = 0. Single-variant $n = 45{,}231$ individuals. Gene-level $n = 43{,}074$ unrelated individuals.

known limitations of animal models[26], which can be highly dependent on model conditions[27], to predict human physiology. Candidate genes with significant but directionally unexpected gene-level associations may provide valuable insights into seemingly promising preclinical results: for example, the protective gene-level signal for *ATM* in our analysis (burden test of PTVs odds ratio = 0.50, $P = 0.003$) contradicts previous expectations—based on insulin resistance and impaired glucose tolerance in *Atm* knockout mice[28]—that *ATM* loss-of-function should increase T2D risk. Evidence is even less favourable that *ATM* haploinsufficiency strongly increases T2D risk, rejecting an odds ratio > 2 at $P = 1.3 \times 10^{-8}$. These observations could be relevant in the ongoing study of whether *ATM* has a role in metformin response[29] or whether *ATM* activators are considered able to treat cardiovascular disease[30].

Comparison of rare and common variant associations

Despite early arguments that rare-variant studies would considerably advance our understanding of complex diseases[5], most genetic discoveries continue to be provided by studies of common variants, which can be studied in much larger sample sizes through array-based genotyping and imputation[31]. Previous quantitative analyses have similarly emphasized the main contribution of common variants to T2D heritability[6,10], but they have lacked the sequencing data that are needed to fully evaluate the value added by rare variants (that is, direct sequencing in addition to array-based genotyping and imputation) to discover disease-associated loci, explain disease heritability and elucidate allelic series.

To compare discoveries that were possible from sequencing and array-based studies, we collected genome-wide array data within the same individuals that we sequenced (available for 34,529 (76.3%) individuals; 18,233 cases), imputed variants using best-practice reference panels[32,33] and conducted a single-variant association analysis ('imputed GWAS'; Methods and Supplementary Table 13). Out of 10 exome-wide significant nonsynonymous single-variant associations from the sequence analysis, 8 were detected in the imputed GWAS analysis (*PAX4* Arg192His and *MC4R* Ile269Asn were not imputable), together with genome-wide significant non-coding variant associations in 14 additional loci (Fig. 3a and Supplementary Table 14). All 10 variants with significant single-variant sequence associations were also present on the Illumina Exome Array[6]. These results demonstrate the limited power of sequencing to detect single-variant associations beyond array-based genotyping and imputation, even before considering the much larger sample sizes enabled by the substantially lower cost of array-based genotyping.

We next compared the contributions to T2D heritability of the strongest (common) single-variant associations from the imputed GWAS to those of the strongest (mostly rare-variant) gene-level associations from the sequencing analysis (Methods). The three exome-wide significant gene-level signals explain an estimated 0.11% (*MC4R*), 0.092% (*PAM*) and 0.072% (*SLC30A8*) of T2D genetic variance, only 10–20% of the variance explained by the three strongest independent common-variant associations in the imputed GWAS (*TCF7L2*, 0.89%; *KCNQ1*, 0.81%; *CDC123*, 0.35%; Fig. 3b). More broadly, fitting a previous exponential model of heritability[34] to our data (Methods) estimated

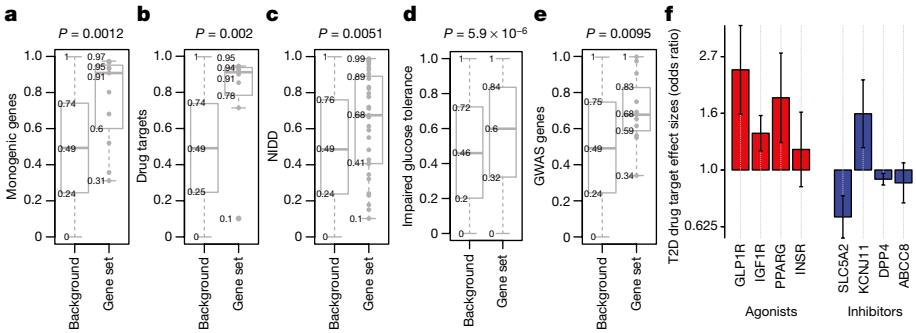


Fig. 2 | Gene set analysis. a–**e**, Rank percentiles (1 = highest) for gene-level associations (compared to matched genes) within 11 monogenic diabetes genes (548 matched genes) (**a**), 8 T2D drug targets (400 matched genes) (**b**), 31 genes linked to NIDD in mice (1,499 matched genes) (**c**), 323 genes linked to impaired glucose tolerance in mice (10,043 matched genes) (**d**) and 11 genes with common likely causal coding variants (537 matched genes) (**e**). *P* values are from a one-sided Wilcoxon rank-sum test between each gene set and comparison set. Labels indicate minimum, 25th percentile, median, 75th percentile and maximum. **f**, Estimated odds ratios, from the weighted burden test of the 5/5 mask, for 8 T2D drug targets (red, agonists; blue, inhibitors). Data are mean ± s.e. of log(odds ratio) from the burden test. $n = 43{,}071$ unrelated individuals.

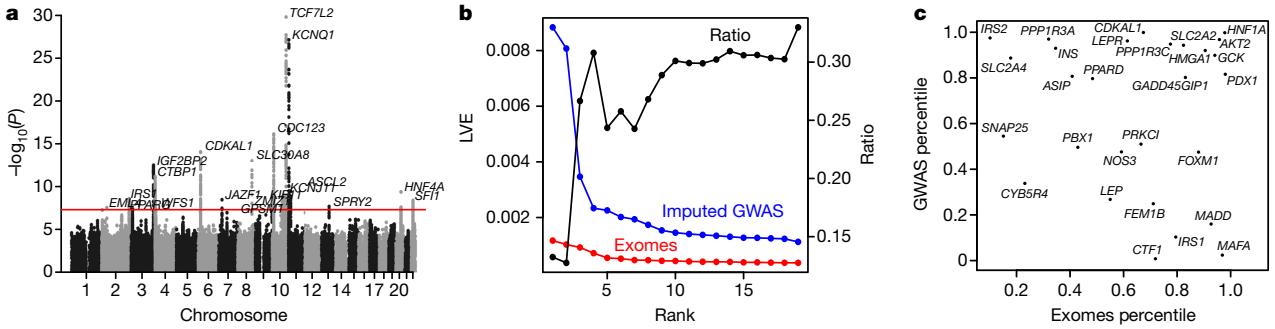


Fig. 3 | Comparison of exome-sequencing to array-based GWAS. **a**, Single-variant associations, calculated by two-sided Firth logistic regression, from an array-based imputed GWAS of the subset ($n = 34{,}529$) of samples in the exome-sequencing analysis for which array data were available. Labels and axes as described in Fig. 1a. **b**, The observed liability variance explained (LVE) by the top 19 exome-sequencing gene-level associations (red) and the top 19 imputed GWAS single-variant associations (maximum of 1 per 250 kb; blue) and their ratio (black). Signals ranked by liability variance explained. **c**, Gene rank percentiles from exome-sequencing gene-level analysis (*x* axis) and a previous multi-ancestry T2D GWAS[13] (*y* axis). Genes are from the mouse NIDD gene set (Fig. 2c).

that the top 100 gene-level signals associated with T2D explained only 1.96% of genetic variance within our study. These results argue against a large contribution to T2D heritability from even the strongest gene-level signals, even after accounting for potential sources of downward bias in our calculations (see Methods).

We finally assessed whether an array-based GWAS could have detected the many potential allelic series that we observed from direct sequencing. Among the variants that contributed to the exome-wide significant gene-level associations in *SLC30A8*, *MC4R* and *PAM*, we estimate that 95.3% of variants are not imputable ($r^2 > 0.4$; Methods) in the 1000 Genomes multi-ancestry reference panel[32], 74.6% of those in Europeans are not imputable in the larger European-focused Haplotype Reference Consortium panel[10] and 90.2% (79.7% of European variants) are absent from the Illumina Exome Array. Additionally, gene set associations (using gene 'scores'; see Methods) from the imputed GWAS showed suggestive associations (four gene sets achieved $P < 0.05$, nine achieved $P < 0.1$; Supplementary Fig. 13) but were weaker than gene set associations from the sequencing analysis. Some of these gene set associations are detectable in larger array-based studies: analysis of a 110,000-sample multi-ancestry GWAS[13] produced $P < 0.05$ for 12 out of 16 gene sets that we studied (Supplementary Fig. 14); however, the genes (and corresponding variants) that are responsible for the array-based gene set associations were mostly different from those responsible for the sequence-based associations, as the two methods often produced uncorrelated rank orderings of genes within gene sets (for example, $r = -0.11$, $P = 0.57$ for the mouse NIDD gene set; Fig. 3c).

Collectively, these results demonstrate the complementarity of array-based GWAS and exome sequencing, with the former favouring locus discovery and the latter enabling full enumeration of potentially informative alleles.

Inferences from nominally significant associations

The T2D drug targets analysed here illustrate the opportunities and challenges of using current exome-sequencing datasets in translational research. Rare-variant gene-level associations are significant across these targets as a set (Fig. 2b) and predict the correct T2D directional relationship for all but one gene (Fig. 2f). However, to detect—at exome-wide significance—the effect sizes estimated from our study with 80% power would require 75,000–185,000 sequenced cases (150,000–370,000 exomes in a balanced study, or 600,000–1,275,000 exomes from a population with a prevalence of T2D of 8%; Fig. 4a and Methods).

As a consequence, many of the modest associations (for example, $P = 0.05$) in current samples may point to clinically or therapeutically relevant variants or genes (Supplementary Fig. 15). The false-positive rate for these associations is expected to be greater than the false-positive rate for exome-wide significant associations[35] and be further influenced by imperfect calibration of association test statistics. If this false-positive rate can be quantified using independent 'truth' data[36], however, then a modest association signal could help to justify further experimentation on a gene based on the likelihood that it is a true association, the cost of the experiment and the benefit of success[37] (Fig. 4b).

We developed and evaluated a method to quantify the false-positive association rate for nonsynonymous variants in our dataset by using independent data, modelling assumptions and prior data to map single-variant *P* values to estimated posterior probabilities of true, causal associations (PPAs) (Methods and Extended Data Fig. 7). Model parameters in the middle of the range that we explored (Methods and Extended Data Fig. 8) predict that 1.5% (95% confidence interval = 0.74–2.2%) of nonsynonymous variants that achieve $P < 0.05$ in our study are truly, causally associated with T2D, increasing to 3.6% (95% confidence interval = 1.4–5.9%) for $P < 0.005$ and 9.7% (95% confidence interval = 3.9–15.0%) for $P < 5 \times 10^{-4}$ (Supplementary Fig. 16). In this model, 541 (95% confidence interval = 270–810) of the 36,604 nonsynonymous variants with $P < 0.05$ in our study represent true, causal associations.

We next applied this method to variants within a curated set of 94 T2D GWAS loci (Methods), which might be expected to show further enrichment of true associations. Our model predicted that nonsynonymous variants within these loci had even higher PPAs: 2.0% (95% confidence interval = 0.048–4.0%) of such variants overall, 8.1% (3.6–12.4%) with $P < 0.05$ in our study and 17.2% (7.7–24.1%) with $P < 0.005$ were estimated to represent true, causal T2D associations. Of particular note are variants in these loci that not only achieve nominal significance ($P < 0.05$) in our analysis but also have moderate-to-large estimated effects on T2D risk (Supplementary Tables 15, 16), as we predict that a substantial number of these variants (for example, 76 (95% confidence interval = 29–117) out of 746 with estimated odds ratio > 2 and 50 (95% confidence interval = 19–77) out of 503 with estimated odds ratio > 3) show true, causal associations.

Outside of GWAS loci, many genes are suspected to be involved in T2D because of prior evidence from non-genetic sources (for example, animal studies[22] or because of implication in related disorders[23]). To evaluate variants in such genes, we extended our PPA estimation approach to incorporate gene prior probabilities (or 'priors')[38] (Methods and Extended Data Fig. 7d) and applied it to two sets of genes.

First, using a prior of 100% for genes associated with MODY—thus assuming that all genes implicated in MODY are relevant to T2D—our model predicts 24 variants (combined MAF = 1.1%) to have PPA ≥ 40% (Supplementary Table 17). Nine have estimated odds ratio > 3 in our study; as none of these were previously reported to be pathogenic MODY variants, they are therefore novel rare-variant

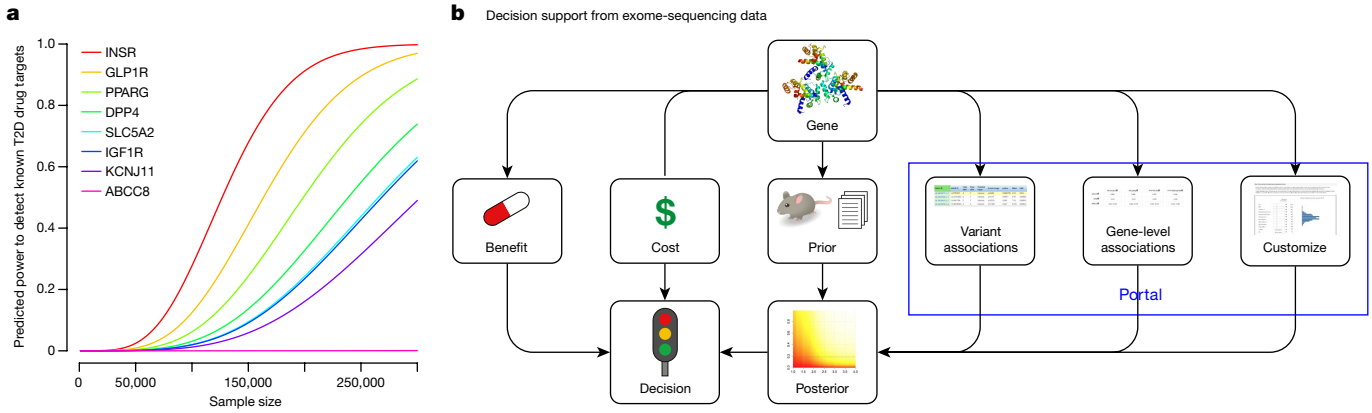


Fig. 4 | Decision support from exome-sequencing data. a, Estimated power, as a function of future sample size (*x* axis), to detect T2D gene-level associations (two-sided test at $P = 6.25 \times 10^{-7}$) with aggregate frequency and odds ratios equal to those estimated from our analysis of 8 T2D drug targets (Fig. 2f). **b**, A proposed workflow for using exome-sequencing data in gene characterization. Depending on the prior belief in the disease relevance of a gene, the cost of experimental characterization and the benefit of gene validation, a decision to conduct the experiment could be informed by the posterior probability of the disease relevance of the gene, as estimated from exome-sequencing association statistics (available through http://www.type2diabetesgenetics.org/).

candidates for use in the prediction of T2D risk. On the other hand, these results show that, once false-positive rates are empirically estimated rather than assumed, nominally significant variants ($P = 0.05$) in genes associated with MODY are still, in absolute terms, more likely to be false-positive rather than true associations[39].

Second, as an example of a gene prior that was derived objectively (rather than subjectively), we used a mixture model approach[40] to estimate the proportion of non-null associations across the mouse NIDD gene set (Methods), leading to a prior of approximately 23% for genes of which knockout causes NIDD in mice. Our model with this prior (Supplementary Table 18) predicts nonsynonymous variants that achieved $P < 0.05$ to have PPAs of 9.9% (PPAs of 24.6% for $P < 0.005$). In particular, we predict several nonsynonymous variants in *MADD* and *NOS3* to have PPA $\geq$ 14% (Supplementary Table 19), suggesting links between variation in these genes and T2D based on combined evidence from human genetic studies and mouse models[41,42].

Although these PPA calculations have limitations (Methods), they present a framework to use suggestive genetic signals to support cost–benefit estimates of 'go/no-go' decisions[43] in the language of decision theory[37] (Fig. 4b). To enable this strategy, we have made our exome-sequencing association results publically available through the AMP T2D Knowledge Portal (http://www.type2diabetesgenetics.org/), which supports queries of precomputed associations and further enables dynamic recomputations of associations with custom covariates and sample- and/or variant-filtering criteria.

Discussion

Our results provide a nuanced description of rare variation and its association with T2D, which might also apply to other complex diseases. Rare-variant gene-level signals are likely to be distributed across numerous genes; however, the vast majority of signals individually explain vanishingly small amounts of T2D heritability: more than one million samples may be required for rare-variant signals in validated therapeutic targets to become significant exome-wide. Even among the four genes that reached exome-wide significance in our analysis, two (*MC4R* and *PAM*) do not include unusually strong rare-variant associations but rather typically modest rare-variant associations that are boosted from nominal to exome-wide significance by low-frequency variants.

Thus, for biological discovery in many complex traits, such as T2D, exome sequencing and array-based GWAS seem complementary: locus discovery and fine mapping are achieved most efficiently using larger array-based GWAS, whereas rare coding variant allelic series—that could aid experimental gene characterization[44] or provide confidence in disease-gene identification—are best discoverable through sequencing. For personalized medicine, exome sequencing may produce some rare variants with sufficient effect sizes (Supplementary Tables 12, 17) to provide viable contributions to the prediction of genetic risk; however, these are sufficiently rare to be best viewed as complements to rather than replacements for GWAS-derived polygenic risk scores[45]. Whole-genome sequencing might soon become sufficiently cost-effective to subsume both array-based GWAS and exome sequencing; even now, it is essential to expand imputation reference panels to power higher-resolution GWAS across all major ethnicities.

Our results suggest that, for now, maximizing the utility of exome sequencing will require drawing insights from associations that do not (yet) reach exome-wide significance. To help to interpret these suggestive associations, we present a principled and empirically calibrated Bayesian approach (Fig. 4, Extended Data Fig. 7 and Supplementary Table 18) to estimate the association probability for any variant in our dataset, highlighting its use to interpret variants in known disease genes and prioritize genes from animal model studies for further investigation. Results and customized analyses from our study can be accessed through a public web portal (http://www.type2diabetesgenetics.org/), advancing the use of exome-sequencing data across many branches of biomedical research.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1231-2.

Received: 6 June 2018; Accepted: 23 April 2019;
Published online: 22 May 2019

1. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
2. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP–trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
3. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
4. Grotz, A. K., Gloyn, A. L. & Thomsen, S. K. Prioritising causal genes at type 2 diabetes risk loci. *Curr. Diab. Rep.* **17**, 76 (2017).
5. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**, 415–425 (2010).
6. Mahajan, A. et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **50**, 559–571 (2018).
7. Flannick, J. et al. Loss-of-function mutations in *SLC30A8* protect against type 2 diabetes. *Nat. Genet.* **46**, 357–363 (2014).
8. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).

9. Zuk, O. et al. Searching for missing heritability: designing rare variant association studies. *Proc. Natl Acad. Sci. USA* **111**, E455–E464 (2014).
10. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
11. Moutsianas, L. et al. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet.* **11**, e1005165 (2015).
12. Sveinbjornsson, G. et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat. Genet.* **48**, 314–317 (2016).
13. Mahajan, A. et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).
14. Tan, K. et al. Functional characterization and structural modeling of obesity associated mutations in the melanocortin 4 receptor. *Endocrinology* **150**, 114–125 (2009).
15. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
16. Sladek, R. et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
17. Steinthorsdottir, V. et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* **46**, 294–298 (2014).
18. Thomsen, S. K. et al. Type 2 diabetes risk alleles in *PAM* impact insulin release from human pancreatic β-cells. *Nat. Genet.* **50**, 1122–1131 (2018).
19. Rutter, G. A. & Chimienti, F. *SLC30A8* mutations in type 2 diabetes. *Diabetologia* **58**, 31–36 (2015).
20. Wessel, J. et al. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat. Commun.* **6**, 5897 (2015).
21. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
22. Blake, J. A. et al. Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res.* **45**, D723–D729 (2017).
23. Flannick, J., Johansson, S. & Njølstad, P. R. Common and rare forms of diabetes mellitus: towards a continuum of diabetes subtypes. *Nat. Rev. Endocrinol.* **12**, 394–406 (2016).
24. Dewey, F. E. et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, aaf6814 (2016).
25. Snider, K. E. et al. Genotype and phenotype correlations in 417 children with congenital hyperinsulinism. *J. Clin. Endocrinol. Metab.* **98**, E355–E363 (2013).
26. Seok, J. et al. Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc. Natl Acad. Sci. USA* **110**, 3507–3512 (2013).
27. Kleiner, S. et al. Mice harboring the human *SLC30A8* R138X loss-of-function mutation have increased insulin secretory capacity. *Proc. Natl Acad. Sci. USA* **115**, E7642–E7649 (2018).
28. Takagi, M. et al. ATM regulates adipocyte differentiation and contributes to glucose homeostasis. *Cell Rep.* **10**, 957–967 (2015).
29. The GoDARTS and UKPDS Diabetes Pharmacogenetics Study Group & The Wellcome Trust Case Control Consortium 2. Common variants near *ATM* are associated with glycemic response to metformin in type 2 diabetes. *Nat. Genet.* **43**, 117–120 (2011).
30. Espach, Y., Lochner, A., Strijdom, H. & Huisamen, B. ATM protein kinase signaling, type 2 diabetes and cardiovascular disease. *Cardiovasc. Drugs Ther.* **29**, 51–58 (2015).
31. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
32. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
33. The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
34. Goldstein, D. B. Common genetic variation and human traits. *N. Engl. J. Med.* **360**, 1696–1698 (2009).
35. Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4**, 45–61 (2002).
36. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81**, 208–227 (2007).
37. Peterson, M. *An Introduction to Decision Theory* (Cambridge Univ. Press, New York, 2009).
38. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10**, 681–690 (2009).
39. Flannick, J. et al. Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. *Nat. Genet.* **45**, 1380–1385 (2013).
40. Zhang, S. D. Towards accurate estimation of the proportion of true null hypotheses in multiple testing. *PLoS ONE* **6**, e18874 (2011).
41. Li, L. C. et al. *IG20/MADD* plays a critical role in glucose-induced insulin secretion. *Diabetes* **63**, 1612–1623 (2014).
42. Nakagawa, T. et al. Diabetic endothelial nitric oxide synthase knockout mice develop advanced diabetic nephropathy. *J. Am. Soc. Nephrol.* **18**, 539–550 (2007).
43. Wagner, J. et al. A dynamic map for learning, communicating, navigating and improving therapeutic development. *Nat. Rev. Drug Discov.* **17**, 150 (2018).
44. Starita, L. M. et al. Variant interpretation: functional assays to the rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).
45. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).

Jason Flannick[1,2,3,4]*, Josep M. Mercader[1,4,5,6,50,158], Christian Fuchsberger[7,8,9,158], Miriam S. Udler[1,4,5,6,50,158], Anubha Mahajan[10,11,158], Jennifer Wessel[12,13,14], Tanya M. Teslovich[15], Lizz Caulkins[1,4], Ryan Koesterer[1,4], Francisco Barajas-Olmos[16], Thomas W. Blackwell[7,9], Eric Boerwinkle[17,18], Jennifer A. Brody[19], Federico Centeno-Cruz[16], Ling Chen[6,50], Siying Chen[7,9], Cecilia Contreras-Cubas[16], Emilio Córdova[16], Adolfo Correa[20], Maria Cortes[21], Ralph A. DeFronzo[22], Lawrence Dolan[23], Kimberly L. Drews[24], Amanda Elliott[1,4,6,50], James S. Floyd[25], Stacey Gabriel[21], Maria Eugenia Garay-Sevilla[26,27], Humberto García-Ortiz[16], Myron Gross[28], Sohee Han[29], Nancy L. Heard-Costa[30,31], Anne U. Jackson[7,9], Marit E. Jørgensen[32,33,34], Hyun Min Kang[7,9], Megan Kelsey[24], Bong-Jo Kim[29], Heikki A. Koistinen[35,36,37], Johanna Kuusisto[38,39], Joseph B. Leader[40], Allan Linneberg[41,42,43], Ching-Ti Liu[44], Jianjun Liu[45,46,47], Valeriya Lyssenko[48,49], Alisa K. Manning[50,51], Anthony Marcketta[15], Juan Manuel Malacara-Hernandez[26,27], Angélica Martínez-Hernández[16], Karen Matsuo[7,9], Elizabeth Mayer-Davis[52], Elvia Mendoza-Caamal[16], Karen L. Mohlke[53], Alanna C. Morrison[54], Anne Ndungu[10], Maggie C. Y. Ng[55,56,57], Colm O'Dushlaine[15], Anthony J. Payne[10], Catherine Pihoker[58], Broad Genomics Platform[59], Wendy S. Post[60], Michael Preuss[61], Bruce M. Psaty[62,63,64,65,66], Ramachandran S. Vasan[31,67], N. William Rayner[10,11,68], Alexander P. Reiner[69], Cristina Revilla-Monsalve[70], Neil R. Robertson[10,11], Nicola Santoro[71], Claudia Schurmann[61], Wing Yee So[72,73,74], Xavier Soberón[16], Heather M. Stringham[7,9], Tim M. Strom[75,76], Claudia H. T. Tam[72,73,74], Farook Thameem[77], Brian Tomlinson[72], Jason M. Torres[10], Russell P. Tracy[78,79], Rob M. van Dam[46,47,80], Marijana Vujkovic[81], Shuai Wang[44], Ryan P. Welch[7,9], Daniel R. Witte[82,83], Tien-Yin Wong[84,85,86], Gil Atzmon[87,88,89], Nir Barzilai[87,89], John Blangero[90,91], Lori L. Bonnycastle[92], Donald W. Bowden[55,56,57], John C. Chambers[93,94,95], Edmund Chan[46], Ching-Yu Cheng[96], Yoon Shin Cho[97], Francis S. Collins[92], Paul S. de Vries[54], Ravindranath Duggirala[90,91], Benjamin Glaser[98], Clicerio Gonzalez[99], Ma Elena Gonzalez[100], Leif Groop[48,101], Jaspal Singh Kooner[102], Soo Heon Kwak[103], Markku Laakso[38,39], Donna M. Lehman[22], Peter Nilsson[104], Timothy D. Spector[105], E. Shyong Tai[46,47,85], Tiinamaija Tuomi[101,106,107,108], Jaakko Tuomilehto[109,110,111,112], James G. Wilson[113], Carlos A. Aguilar-Salinas[114], Erwin Bottinger[61], Brian Burke[24], David J. Carey[40], Juliana C. N. Chan[72,73,74], Josée Dupuis[31,44], Philippe Frossard[115], Susan R. Heckbert[116,117], Mi Yeong Hwang[29], Young Jin Kim[29], H. Lester Kirchner[40], Jong-Young Lee[118], Juyoung Lee[29], Ruth J. F. Loos[61,119], Ronald C. W. Ma[72,73,74], Andrew D. Morris[120], Christopher J. O'Donnell[3,121,122,123], Colin N. A. Palmer[124], James Pankow[125], Kyong Soo Park[102,126,127], Asif Rasheed[115], Danish Saleheen[81,115], Xueling Sim[47], Kerrin S. Small[105], Yik Ying Teo[47,128,129], Christopher Haiman[130], Craig L. Hanis[131], Brian E. Henderson[130], Lorena Orozco[16], Teresa Tusié-Luna[114,132], Frederick E. Dewey[15], Aris Baras[15], Christian Gieger[133,134], Thomas Meitinger[75,76,135], Konstantin Strauch[133,136], Leslie Lange[137], Niels Grarup[138], Torben Hansen[138,139], Oluf Pedersen[138], Philip Zeitler[140], Dana Dabelea[141], Goncalo Abecasis[7,9], Graeme I. Bell[26,27], Nancy J. Cox[142], Mark Seielstad[143,144], Rob Sladek[145,146,147], James B. Meigs[21,50,148], Steve S. Rich[149], Jerome I. Rotter[150,151,152], DiscovEHR Collaboration[59], CHARGE[59], LuCamp[59], ProDiGY[59], GoT2D[59], ESP[59], SIGMA-T2D[59], T2D-GENES[59], AMP-T2D-GENES[59], David Altshuler[1,4,6,50,153,154,155], Noël P. Burtt[1,4], Laura J. Scott[7,9], Andrew P. Morris[10,156], Jose C. Florez[1,4,5,6,50], Mark I. McCarthy[10,11,157] & Michael Boehnke[7,9]

[1]Program in Metabolism, Broad Institute, Cambridge, MA, USA. [2]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. [3]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [4]Program in Medical & Population Genetics, Broad Institute, Cambridge, MA, USA. [5]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [6]Diabetes Research Center (Diabetes Unit), Massachusetts General Hospital, Boston, MA, USA. [7]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA. [8]Institute for Biomedicine, Eurac Research, Bolzano, Italy. [9]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [10]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [11]Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [12]Department of Epidemiology, Fairbanks School of Public Health, Indiana University, Indianapolis, IN, USA. [13]Department of Medicine, School of Medicine, Indiana University, Indianapolis, IN, USA. [14]Diabetes Translational Research Center, Indiana University, Indianapolis, IN, USA. [15]Regeneron Genetics Center, Regeneron Pharmaceuticals, Tarrytown, NY, USA. [16]Instituto Nacional de Medicina Genómica, Mexico City, Mexico. [17]Human Genetics Center, Department of Epidemiology Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA. [18]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [19]Cardiovascular Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA. [20]Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA. [21]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [22]Department of Medicine, University of Texas Health Science Center, San Antonio, TX, USA. [23]Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [24]Biostatistics Center, George Washington University, Rockville, MD, USA. [25]Department of Medicine and Epidemiology, University of Washington, Seattle, WA, USA. [26]Department of Medicine, The University of Chicago, Chicago, IL, USA. [27]Department of Human Genetics, The University of Chicago, Chicago, IL, USA. [28]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN, USA. [29]Division of Genome Research, Center for Genome Science, National Institute of Health, Chungcheongbuk-do, South Korea. [30]Department of Neurology, Boston University School of Medicine, Boston, MA, USA. [31]National Heart Lung and Blood Institute's Framingham Heart Study, Framingham, MA, USA. [32]Steno Diabetes Center Copenhagen, Gentofte, Denmark. [33]National Institute of Public Health, University of Southern Denmark, Copenhagen, Denmark. [34]Greenland Centre for Health Research, University of Greenland, Nuuk, Greenland. [35]Department of Public Health Solutions, National Institute for Health and Welfare, Helsinki, Finland. [36]University of Helsinki and Department of Medicine, Helsinki University Central Hospital, Helsinki, Finland. [37]Minerva Foundation Institute for Medical Research, Helsinki, Finland. [38]Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland, Kuopio, Finland. [39]Department of Medicin, Kuopio University Hospital, Kuopio, Finland. [40]Geisinger Health System, Danville, PA, USA. [41]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [42]Center for Clinical Research and Prevention, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark. [43]Department of Clinical Experimental Research, Rigshospitalet, Copenhagen, Denmark. [44]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [45]Genome Institute of Singapore, Agency for Science Technology and Research, Singapore, Singapore. [46]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, National University Health System, Singapore, Singapore. [47]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore. [48]Department of Clinical Sciences, Diabetes and Endocrinology, Lund University Diabetes Centre, Malmö, Sweden. [49]Department of Clinical Science, University of Bergen, Bergen, Norway. [50]Department of Medicine, Harvard Medical School, Boston, MA, USA. [51]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital, Harvard University, Boston, MA, USA. [52]University of North Carolina Chapel Hill, Chapel Hill, NC, USA. [53]Department of Genetics, University of North Carolina Chapel Hill, Chapel Hill, NC, USA. [54]Human Genetics Center, Department of Epidemiology Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA. [55]Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [56]Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [57]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. [58]Seattle Children's Hospital, Seattle, WA, USA. [59]A list of participants and their affiliations appears in the Supplementary Information. [60]Division of Cardiology, Department of Medicine, Johns Hopkins University, Baltimore, MD, USA. [61]Charles R. Bronfman Institute of Personalized Medicine, Mount Sinai School of Medicine, New York, NY, USA. [62]Cardiovascular Health Research Unit, University of Washington, Seattle, WA, USA. [63]Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA. [64]Department of Medicine, University of Washington, Seattle, WA, USA. [65]Department of Epidemiology, University of Washington, Seattle, WA, USA. [66]Department of Health Services, University of Washington, Seattle, WA, USA. [67]Preventive Medicine & Epidemiology, Medicine, Boston University School of Medicine, Boston, MA, USA. [68]Department of Human Genetics, Wellcome Trust Sanger Institute, Hinxton, UK. [69]Department of Epidemiology, University of Washington, Seattle, WA, USA. [70]Instituto Mexicano del Seguro Social SXXI, Mexico City, Mexico. [71]Department of Pediatrics, Yale University, New Haven, CT, USA. [72]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China. [73]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China. [74]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China. [75]Institute of Human Genetics, Technische Universität München, Munich, Germany. [76]Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [77]Health Science Center, Department of Biochemistry, Faculty of Medicine, Kuwait University, Safat, Kuwait. [78]Department of Pathology and Laboratory Medicine, The Robert Larner M.D. College of Medicine, University of Vermont, Burlington, VT, USA. [79]Department of Biochemistry, The Robert Larner M.D. College of Medicine, University of Vermont, Burlington, VT, USA. [80]Department of Nutrition, Harvard School of Public Health, Boston, MA, USA. [81]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA, USA. [82]Department of Public Health, Aarhus University, Aarhus, Denmark. [83]Danish Diabetes Academy, Odense, Denmark. [84]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore. [85]Duke-NUS Medical School Singapore, Singapore, Singapore. [86]Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, National University Health System, Singapore, Singapore. [87]Department of Medicine, Albert Einstein College of Medicine, New York, NY, USA. [88]Faculty of Natural Science, University of Haifa, Haifa, Israel. [89]Department of Genetics, Albert Einstein College of Medicine, New York, NY, USA. [90]Department of Human Genetics, University of Texas Rio Grande Valley, Edinburg, TX, USA. [91]South Texas Diabetes and Obesity Institute, Brownsville, TX, USA. [92]Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [93]Department of Epidemiology and Biostatistics, Imperial College London, London, UK. [94]Department of Cardiology, Ealing Hospital NHS Trust, Southall, UK. [95]Imperial College Healthcare NHS Trust, Imperial College London, London, UK. [96]Ophthalmology & Visual Sciences Academic Clinical Program (Eye ACP), Duke-NUS Medical School, Singapore, Singapore. [97]Department of Biomedical Science, Hallym University, Chuncheon, South Korea. [98]Endocrinology and Metabolism Service, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [99]Unidad de Diabetes y Riesgo Cardiovascular, Instituto Nacional de Salud Pública, Cuernavaca, Mexico. [100]Centro de Estudios en Diabetes, Mexico City, Mexico. [101]Institute for Molecular Genetics Finland, University of Helsinki, Helsinki, Finland. [102]National Heart and Lung Institute, Cardiovascular Sciences, Imperial College London, London, UK. [103]Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea. [104]Department of Clinical Sciences, Medicine, Lund University, Malmö, Sweden. [105]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [106]Folkhälsan Research Centre, Helsinki, Finland. [107]Department of Endocrinology, Abdominal Centre, Helsinki University Hospital, Helsinki, Finland. [108]Research Programs Unit, Diabetes and Obesity, University of Helsinki, Helsinki, Finland. [109]Diabetes Prevention Unit, National Institute for Health and Welfare, Helsinki, Finland. [110]Center for Vascular Prevention, Danube University Krems, Krems, Austria. [111]Diabetes Research Group, King Abdulaziz

University, Jeddah, Saudi Arabia. [112]Instituto de Investigacion Sanitaria del Hospital Universario LaPaz (IdiPAZ), University Hospital LaPaz, Autonomous University of Madrid, Madrid, Spain. [113]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA. [114]Instituto Nacional de Ciencias Medicas y Nutricion, Mexico City, Mexico. [115]Center for Non-Communicable Diseases, Karachi, Pakistan. [116]Cardiovascular Health Research Unit, University of Washington, Seattle, WA, USA. [117]Department of Epidemiology, University of Washington, Seattle, WA, USA. [118]Department of Business Data Convergence, Chungbuk National University, Gyeonggi-do, South Korea. [119]The Mindich Child Health and Development Insititute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [120]Clinical Research Centre, Centre for Molecular Medicine, Ninewells Hospital and Medical School, Dundee, UK. [121]Section of Cardiology, Department of Medicine, VA Boston Healthcare, Boston, MA, USA. [122]Brigham and Women's Hospital, Boston, MA, USA. [123]Intramural Administration Management Branch, National Heart Lung and Blood Institute, NIH, Framingham, MA, USA. [124]Pat Macpherson Centre for Pharmacogenetics and Pharmacogenomics, Medical Research Institute, Ninewells Hospital and Medical School, Dundee, UK. [125]Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, MN, USA. [126]Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, South Korea. [127]Department of Internal Medicine, Seoul National University College of Medicine, Seoul, South Korea. [128]Life Sciences Institute, National University of Singapore, Singapore, Singapore. [129]Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore. [130]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [131]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA. [132]Instituto de Investigaciones Biomédicas, Departamento de Medicina Genómica y Toxicología, Universidad Nacional Autónoma de México, Mexico City, Mexico. [133]Research Unit of Molecular Epidemiology, Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [134]German Center for Diabetes Research (DZD e.V.), Neuherberg, Germany. [135]Deutsches Forschungszentrum für Herz-Kreislauferkrankungen (DZHK), Partner Site Munich Heart Alliance, Munich, Germany. [136]Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Neuherberg, Germany. [137]Department of Medicine, University of Colorado Denver, Aurora, CO, USA. [138]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [139]Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark. [140]Department of Pediatrics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [141]Department of Epidemiology, Colorado School of Public Health, Aurora, CO, USA. [142]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN, USA. [143]Department of Laboratory Medicine & Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA. [144]Blood Systems Research Institute, San Francisco, CA, USA. [145]Department of Human Genetics, McGill University, Montreal, Quebec, Canada. [146]Division of Endocrinology and Metabolism, Department of Medicine, McGill University, Montreal, Quebec, Canada. [147]McGill University and Génome Québec Innovation Centre, Montreal, Quebec, Canada. [148]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA. [149]Center for Public Health Genomics, University of Virginia School of Medicine, Charlottesville, VA, USA. [150]Department of Pediatrics, Los Angeles BioMedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, USA. [151]Department of Medicine, Los Angeles BioMedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, USA. [152]Institute for Translational Genomics and Population Sciences, Los Angeles BioMedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, USA. [153]Department of Genetics, Harvard Medical School, Boston, MA, USA. [154]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [155]Department of Molecular Biology, Massachusetts General Hospital, Boston, MA, USA. [156]Department of Biostatistics, University of Liverpool, Liverpool, UK. [157]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals Trust, Oxford, UK. [158]These authors contributed equally: Josep M. Mercader, Christian Fuchsberger, Miriam S. Udler, Anubha Mahajan. *e-mail: flannick@broadinstitute.org

METHODS

A full description of the methods used in this study is available as Supplementary Methods.

Data reporting. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Sample selection. We drew samples for exome sequencing from six consortia, most of which consisted of multiple studies and are described fully in Supplementary Table 1. T2D case status was determined according to study-specific criteria described in full in in Supplementary Table 1 and the Supplementary Methods. All individuals provided informed consent and all samples were approved for use by their institution's institutional review board or ethics committee, as previously reported[10,46–48]. Samples that were newly sequenced at The Broad Institute as part of T2D-GENES, SIGMA and ProDiGY are covered under Partners Human Research Committee protocol 2017P000445/PHS 'Diabetes Genetics and Related Traits'.

Data generation. The details of data generation, variant calling, quality control and variant annotation are described in full in the Supplementary Methods. In brief, for each consortium, sequencing data were aggregated (if previously available) or newly generated (if not) and then processed through a standard variant calling pipeline. We then measured samples and variants according to several metrics indicative of sequencing quality, excluding those that were outliers relative to the global distribution (Supplementary Fig. 1, Supplementary Table 2). These exclusions produced a 'clean' dataset of 49,484 samples and 7.02 million variants.

Following initial sample and variant quality control, we performed additional rounds of sample exclusion from association analysis (Extended Data Fig. 2). We also excluded the 3,510 childhood diabetes cases from the SEARCH and TODAY studies based on an analysis that suggested their lack of matched controls would induce artefacts in gene-level association analyses (Supplementary Fig. 17). These exclusions produced an 'analysis' dataset of 45,231 individuals and 6.33 million variants. A power analysis of this dataset is presented in the Supplementary Methods.

After these three rounds of sample exclusions, we estimated—within each ancestry—pairwise identity-by-descent values, genetic relatedness matrices and principal components for use in downstream association analyses. We used the identity-by-descent values to generate lists of unrelated individuals within each ancestry, excluding 2,157 individuals to produce an 'unrelated analysis' set of 43,090 individuals (19,828 cases and 23,262 controls) and 6.29 million non-monomorphic variants. We used this set of individuals and variants for single-variant and gene-level tests (described below) that required an unrelated set of individuals.

We annotated variants with the ENSEMBL Variant Effect Predictor[49] (VEP, version 87). We produced both transcript-level annotations for each variant as well as a 'best guess' gene-level annotation using the –flag-pick-allele option (with ranked criteria described in the Supplementary Methods). We used the VEP LofTee (https://github.com/konradjk/loftee) and dbNSFP (version 3.2)[50] plugins to generate additional bioinformatics predictions of variant deleteriousness; from the dbNSFP plugin, we took annotations from 15 different bioinformatics algorithms (listed in Extended Data Fig. 5) and then added annotations from the mCAP[51] algorithm. As these annotations were not transcript-specific, we assigned them to all transcripts for the purpose of downstream analysis.

Although we incorporated both transcript-level and gene-level annotations into gene-level analyses (see below), all single-variant analyses reported in the manuscript or figures are annotated using the 'best guess' annotation for each variant.

Single-variant association analysis in sequencing data. To perform single-variant association analyses, we first stratified samples by cohort of origin and sequencing technology (with some exceptions described in the Supplementary Methods), yielding 25 distinct sample subgroups (Extended Data Fig. 3). For each subgroup, we performed additional variant quality control beyond that used for the 'clean' dataset, excluding variants according to subgroup-specific criteria described in Extended Data Fig. 3; in general, these criteria were strict—particularly for multiallelic variants and X-chromosome variants. We verified that these filters led to a well-calibrated final analysis through inspection of quantile–quantile plots within and across ancestries (Extended Data Fig. 4).

For each of the 25 sample subgroups, we then conducted two single-variant association analyses: one of all (including related) samples using the (two-sided) EMMAX test[52] and one of unrelated samples using the (two-sided) Firth logistic regression test[53]. Both analyses included covariates for sequencing technology, and the Firth analysis included covariates for principal components of genetic ancestry (those among the first 10 that showed $P < 0.05$ association with T2D).

We then conducted a 25-group fixed-effect inverse-variance weighted meta-analysis for each of the Firth and EMMAX tests, using METAL[54]. We used EMMAX results for association P values and Firth results for effect size estimates.

Additional analysis of rs145181683. To assess whether the rs145181683 variant in *SFI1* ($P = 3.2 \times 10^{-8}$ in the exome-sequencing analysis) represented a true novel association, we obtained association statistics from 4,522 Latinos[55]) who did not overlap with the current study. On the basis of the odds ratio (1.19) estimated in our analysis and the MAF (12.7%) in the replication sample, the power was 91% to achieve $P < 0.05$ under a one-sided association test. The observed evidence ($P = 0.90$, odds ratio = 1.00) did not support rs145181683 as a true T2D association. Further investigation of this lack of replication evidence suggested that, although the association from our sequence analysis is unlikely to be a technical artefact (genotyping quality was high), it could possibly be a proxy for a different (Native American-specific) non-coding causal variant (full details are available in the Supplementary Methods). Further fine-mapping and replication efforts will be necessary to test this hypothesis.

Gene-level analysis. For each gene, following previous studies[10,56,57], we separately tested seven different 'masks' of variants grouped by similar predicted severity (defined in Extended Data Fig. 5). For each gene and each mask, we created up to three groupings of alleles, corresponding to different transcript sets of the gene; for many genes, two or more of these allele groupings were identical.

Before running gene-level tests, we performed additional quality control on sample genotypes. For each of the 25 sample subgroups (the same as used for single-variant analyses), we identified variants that failed subgroup-specific quality control criteria (shown in Extended Data Fig. 5) and set genotypes for these variants in all individuals in the subgroup as 'missing'.

We conducted two gene-level association tests: a burden test, which assumes all analysed variants within a gene are of the same effect, and SKAT[15], which allows variability in variant effect size (and direction); each of these tests is two-sided. We performed each test across all unrelated individuals with 10 principal components of genetic ancestry, sample subgroup and sequencing technology as covariates. As this 'mega-analysis' strategy was different from the meta-analysis strategy that we used for single-variant analyses, as a quality control exercise we conducted a single-variant mega-analysis and found that its results showed broad correlation with those from the original meta-analysis (Supplementary Fig. 18).

We then developed two methods to consolidate the $2 \times 7 = 14$ P values produced for each gene (described in full in Extended Data Fig. 5, Supplementary Methods and Supplementary Figs. 5, 6). First, we corrected the smallest P value for each gene according to the effective number of independent masks tested for the gene (variable, but on average 3.6), based on the gene-specific correlation of variants across masks[58] (referred to as the minimum P-value test; Supplementary Fig. 19). Second, we tested all nonsynonymous variants (that is, missense, splice site and protein-truncating mutations), but weighted each variant according to its estimated probability of causing gene inactivation[9] (referred to as the weighted test, which essentially assessed the effect of gene haploinsufficiency from combined analysis of protein-truncating and missense variants; Supplementary Fig. 6). We verified that these two consolidation methods were well-calibrated (Extended Data Fig. 6) and broadly consistent yet distinct: across the 10 most significantly associated genes, P values were nominally significant using both methods for 8 genes but varied by 1–3 orders of magnitude (Extended Data Table 2).

Because each gene mask could in fact represent up to three sets of alleles (owing to the transcript-specific annotation strategy that we used), for each of the four analyses multiple P values were possible for some genes. To produce a single gene-level P value for each of the four analyses, we thus collapsed (for each gene) the set of P values across transcript sets into a single gene-level P value using the minimum P-value test.

We used a conservative Bonferroni-corrected gene-level exome-wide significance threshold of $P = 0.05/(2 \text{ tests} \times 2 \text{ consolidation methods} \times 19{,}020 \text{ genes}) = 6.57 \times 10^{-7}$. For each gene referenced in the manuscript, we report the P value and odds ratio from the analysis that achieved the lowest P value for the gene.

Gene-level analysis near T2D GWAS signals. In principle, a nearby common-variant association could lead to over- or underestimation of the strength of a gene-level association[59]. To assess whether differential patterns of rare variation across common-variant haplotypes could significantly affect our gene-level results, we conducted two analyses (described in the Supplementary Methods) and found no evidence that confounding from common-variant haplotypes was primarily responsible for the associations that were observed in our gene-level analyses.

Further exploration of significant gene-level associations. For our exome-wide significant gene-level associations (*MC4R*, *PAM* and *SLC30A8*), we conducted additional gene-level analyses to dissect the aggregate signals that were observed. First, we performed tests by progressively removing alleles in order of lowest single-variant analysis P value, in order to understand the (minimum) number of alleles that contributed statistically to the aggregate signal. Second, we performed tests conditional on each allele in the sequence (that is, calculating separate models with each individual allele as a covariate), and we then compared the resulting P values to the full gene-level P value, in order to assess the contribution of each allele individually to the signal. Finally, for *MC4R*, we conducted an analysis with an added sample covariate for body-mass index and found that it, as shown previously[60,61], reduces the significance of both the Ile269Asn single-variant signal ($P = 1.0 \times 10^{-5}$) and the gene-level signal not attributable to Ile269Asn ($P = 0.035$).

To evaluate which ancestries contributed variants to *MC4R*, *SLC30A8*, and *PAM*, we calculated the proportion of variants in each signal unique to an ancestry and also compared the significance and direction of effect of each signal across ancestries. Across the three signals, 68.4% (287 out of 419) of variants in total were unique to one ancestry (63.9% for *MC4R*, 67.0% for *SLC30A8* and 71.6% for *PAM*). Each signal had a direction of effect that was consistent across all five ancestries and each signal achieved $P < 0.05$ in at least two ancestries (*MC4R* in East-Asians and Hispanics; *SLC30A8* in all ancestries other than African-Americans; and *PAM* in Europeans, South-Asians and Hispanics).

Analysis of exomes from the Geisinger Health System. We obtained gene-level association results that were previously computed from an analysis of 49,199 individuals (12,973 T2D cases and 36,226 controls) from the Geisinger Health System (GHS). Association statistics were available for 44 out of the 50 genes with the strongest gene-level associations from our study. A power analysis of the GHS replication analysis is available in the Supplementary Methods.

GHS sequencing data were processed and analysed as previously described[24], and variants were grouped into four (nested) masks (roughly corresponding to the LofTee, 5/5, 1/5 1% and 0/5 1% masks; more details are available in the Supplementary Methods). For each mask, association results were computed using two-sided logistic regression under an additive burden model (with phenotype regressed on the number of variants carried by each individual) with age, age^2 and sex as covariates. To produce a single GHS *P* value for each gene, we applied the minimum *P*-value procedure across the four mask-level results.

Analysis of exomes from the CHARGE consortium. We collaborated with the CHARGE consortium to analyse the 50 genes with the strongest gene-level associations from our study in 12,467 individuals (3,062 T2D cases and 9,405 controls) from their previously described study[62,63]. A power analysis of the CHARGE replication analysis is available in the Supplementary Methods.

Variants in the CHARGE exomes were annotated and grouped into seven masks using the same procedure as for the original exome-sequencing analysis. Burden and SKAT association tests were then performed in the Analysis Commons[64] using a two-sided logistic mixed model[65] assuming an additive genetic model and adjusted for age, sex, study, race and kinship. To produce a single CHARGE *P* value for each gene, we applied the minimum *P*-value procedure across the seven mask-level results, as for the GHS analysis.

Meta-analysis with CHARGE and GHS. We conducted a meta-analysis among our original burden analysis and those of CHARGE and GHS. For each gene, we selected the mask that achieved the lowest *P* value in our original analysis and conducted a two-sided sample-size weighted meta-analysis with the results from CHARGE and GHS for the same mask (or an analogous mask as defined in the Supplementary Methods).

Investigation of the *UBE2NL* association. We investigated the novel association that was found in the gene-level meta-analysis (*UBE2NL*, meta-analysis $P = 5.6 \times 10^{-7}$) in more detail. The *UBE2NL* burden signal was due to five PTVs in the original analysis (observed in 29 cases and 1 control; all of which had high (>45×) sequencing coverage; Supplementary Table 8) and was replicated at $P = 0.02$ in CHARGE; *UBE2NL* results were not available in GHS. As *UBE2NL* lies on the X chromosome, we conducted a sex-stratified analysis of the original samples and observed independent associations in both men ($P = 5.7 \times 10^{-4}$) and women ($P = 1.6 \times 10^{-3}$). *UBE2NL* does not lie near any known GWAS associations (http://www.type2diabetesgenetics.org/) and has few available references[66–68], suggesting that it may be a novel T2D-relevant gene, although further replication will be important to establish its association.

Evaluation of directional consistency between exome-sequencing, CHARGE and GHS analyses. We examined the concordance of direction of effect size estimates (that is, both odds ratios of >1 or <1) between burden tests from our original exome-sequencing analysis and those from CHARGE and GHS. For the 46 genes advanced for replication with burden $P < 0.05$ for at least one mask (that is, ignoring those with evidence for association only under the SKAT model), we compared the direction of effect estimated for the mask with lowest *P*-value mask to that estimated for the same (or analogous) mask in the GHS or CHARGE analysis. We then conducted a one-sided exact binomial test to assess whether the fraction of results with consistent direction of effects was significantly greater than expected by chance.

Gene set analysis in sequencing data. We curated 16 sets of candidate T2D-relevant genes, defined in Supplementary Table 9 with criteria as specified in the Supplementary Methods. For each gene set, we constructed sets of matched genes with similar numbers and frequencies of variants within them (details are provided in the Supplementary Methods). A sensitivity analysis of this matching strategy is presented in the Supplementary Methods.

To conduct a gene set analysis, we then combined the genes in the gene set with the matched genes. Within the combined list of genes, we ranked genes using the *P* values observed for the minimum *P*-value burden test. We then used a one-side Wilcoxon rank-sum test to assess whether genes in the gene set had significantly higher ranks than the comparison genes.

Use of gene-level associations to predict effector genes. To assess whether gene-level associations from exome sequencing—which are composed mostly of rare variants independent of any GWAS associations—could prioritize potential effector genes within known T2D GWAS loci, we first assessed whether predicted effector genes (based on common-variant associations) were also enriched for rare coding variant associations. Our analysis (described in full in the Supplementary Methods) indicated that effector genes predicted from common coding variant associations do show significant enrichments ($P = 8.8 \times 10^{-3}$), but effector genes predicted from transcript-level associations do not ($P = 0.72$).

We then curated a list of 94 T2D GWAS loci, and 595 genes that were within 250 kb of any T2D GWAS index variant, from a 2016 T2D genetics review[69] and observed 40 with a $P < 0.05$ gene-level signal (Supplementary Table 12), greater than the $595 \times 0.05 = 29.75$ expected by chance ($P = 0.038$). Only three (*SLC30A8*, *PAM* and *HNF1A*) were from the list that we curated of 11 genes with causal common coding variants[6]. We found that these 40 genes were significantly more enriched for protein interactions ($P = 0.03$; observed mean = 11.4, expected mean = 4.5) than the 184 genes implicated based on proximity to the index SNP ($P = 0.64$; observed mean = 21.1, expected mean = 21.9), although evaluation of the biological candidacy of these genes will ultimately require in-depth functional studies[70]. Rare coding variants could therefore, in principle, complement common-variant fine-mapping[71,72] and experimental data[4,70] to help to interpret T2D GWAS associations; however, our results indicate that much larger sample sizes and/or orthogonal experimental data will be required to clearly implicate specific effector genes. A full description of this analysis is included in the Supplementary Methods.

Use of gene-level associations to predict direction of effect. To assess whether gene-level association analyses of predicted deleterious variants could be used to predict therapeutic direction of effect, we compared odds ratios estimated from a modified weighted burden test procedure (described in the Supplementary Methods) to those expected for T2D drug targets (assuming agonist targets to have true odds ratios > 1 and inhibitors to have true odds ratios < 1). For a similar comparison to expectations for mouse gene knockouts, we used the relationship between mouse phenotype and human phenotype specified in the Supplementary Methods. Genes present in two gene sets with opposite expected direction of effects were excluded from this analysis.

Collection and analysis of SNP array data. To compare discoveries from our exome-sequencing analyses to discoveries possible from common-variant GWAS of the same samples, we aggregated all available SNP array data for the exome-sequenced samples (18,233 cases and 17,679 controls; Supplementary Table 13). After sample and variant quality control (described in the Supplementary Methods), we imputed variants from the 1000 Genomes Phase 3[32] (1000G) and Haplotype Reference Consortium[33] (HRC) reference panels using the Michigan Imputation Server[73]. We used 1000G-based imputation for all association analyses and HRC-based imputation to assess the number of exome-sequence variants imputable from the largest available European reference panel (details available in the Supplementary Methods).

After imputation, we performed sample and variant quality control, as well as two-sided association tests, analogous to the exome-sequence single-variant analyses. In contrast to the exome-sequencing analyses, a quantile–quantile plot suggested that the associations from the EMMAX test were not well calibrated, and we therefore used only the Firth test (that is, for both *P* values and odds ratios) in the imputed GWAS analysis.

To conduct gene set analysis with the imputed GWAS data, we first used the method implemented in MAGENTA[74] to calculate gene scores from the imputed GWAS single-variant association results. Following the same protocol as for gene set analysis from the exome-sequencing results, we then conducted a one-sided Wilcoxon rank-sum test to compare the gene scores to those of matched comparison genes. We followed the same approach for the gene set analysis that we conducted in a larger, previously published[13] GWAS.

LVE calculations. To calculate LVEs, we used a previously presented formula[75] (equations are available in the Supplementary Methods) to calculate the LVE of a variant with three genotypes (AA, Aa and aa) and corresponding relative risks (1, RR_1 and RR_2). When presenting the strongest LVE values for the imputed GWAS analysis, we only considered variants that were genotyped in at least 10,000 individuals to avoid potential artefacts that result from a spurious association in a small-sample subgroup. For gene-level LVE calculations, we used the variant mask with lowest *P* value to calculate LVEs. We also conducted a sensitivity analysis to bound the extent to which our gene-level LVE estimates might be biased downwards due to their inclusion of benign alleles; this analysis (described in full in the Supplementary Methods) produced upper bounds of gene-level LVEs that were at most twofold higher than the point estimates.

Prediction of LVE explained by the top 100 and top 1,000 gene-level associations. To forecast the LVE that will be explained once 100 (or 1,000) significant T2D gene-level associations are detected, we applied a previously suggested

model[34] in which the LVE of a gene is related to its rank in the overall gene-level *P*-value distribution. Specifically, the model is $LVE_n = e^{an+b}$ where LVE_n is the LVE of the gene with *n*th lowest gene-level *P* value. We fitted this model using linear regression to the top 50 genes in our analysis (Supplementary Fig. 20), yielding estimates of $a = -0.044$ and $b = -7.07$. We then calculated the LVE of the top 100 (or 1,000) genes by summing the actual LVE of the top three signals (which achieved exome-wide significance in our analysis) with the LVE predicted by the model for genes ranked 4–100 (or 4–1,000).

Estimated power to detect gene-level associations with T2D drug targets. To estimate the power of future studies to detect gene-level associations in genes with effect sizes similar to those for established T2D drug targets, we used aggregate allele frequencies and odds ratios estimated from our gene-level analysis and an assumed prevalence of $K = 0.08$ to calculate a proxy for true population frequencies and relative risks. For each gene, we used odds ratios and frequencies from the variant mask that yielded the strongest gene-level association. Because, on average, these drug targets had five effective tests per mask, we used an exome-wide significance threshold of $\alpha = 1.25 \times 10^{-7}$ for power calculations. We calculated power as previously described[76].

The ranges given in the main text (75,000–185,000 disease cases) represent the numbers from the power calculations for *INSR* (the drug target with the highest observed effect size) and *IGF1R* (the drug target with the lowest observed effect size other than *KCNJ11* and *ABCC8*). We excluded *KCNJ11* and *ABCC8* from this reported range, given that a mixture of risk-increasing and risk-decreasing variants in these genes probably diluted their burden signals. We did not account for uncertainty in estimated odds ratios or aggregate variant frequency in these calculations, as no genes had 95% confidence intervals that that did not overlap odds ratio = 1.

Interpretation of suggestive associations. We quantified the PPA for nonsynonymous variants observed in our dataset as a function of association strength measured by single-variant *P* values. We define a true association as a variant that, when studied in larger sample sizes, will eventually achieve statistical significance owing to a true odds ratio ≠ 1. We distinguish true associations from causal associations: causally associated variants are the subset of truly associated variants in which the variant itself is causal for the increase in disease risk, as opposed to being truly associated due to linkage disequilibrium (LD) with a different causally associated variant (that is, an 'LD proxy'). An overview of the method that we developed for PPA calculations is provided in Extended Data Fig. 7, and a full description of the method is included in the Supplementary Methods. Here, we outline the steps in the approach.

First, for various single-variant *P*-value thresholds in the exome-sequencing analysis, we calculated the fraction of variants that reached this threshold with directions of effect concordant with those of an independent exome array study[10]. For example, 61.3% of nonsynonymous variants within T2D GWAS loci that reached $P < 0.05$ in the exome-sequencing analysis had concordant directions of effect with the independent study, a fraction that decreased (as expected) for higher *P*-value thresholds (for example, 49.4% at $P > 0.5$) or when only variants outside of T2D GWAS loci were analysed (51.9% at $P < 0.05$).

Second, we derived an equation to convert the fraction of concordant associations to an estimated proportion of true associations. This value provides a PPA estimate, as a function of *P* value, for an arbitrary variant in the set initially used to calculate direction of effect concordances. We computed separate mappings for arbitrary nonsynonymous variants (using all exome-wide nonsynonymous variants) and one for nonsynonymous variants within GWAS loci (using only nonsynonymous variants within the 94 T2D GWAS loci). We note that the mapping produced from our analysis applies only to the results from the current study: because other studies have different sample sizes and may apply different statistical tests, the mapping would need to be recomputed to interpret the associations of other studies using the same method.

Third, we converted PPA estimates to estimates of the posterior probability of causal associations (PPA_c). This conversion requires estimates of the fraction of coding variant associations that are causal (as opposed to LD proxies). We explored several values for this parameter, as described in the Supplementary Methods and shown in Extended Data Fig. 8.

Fourth, we extended PPA estimates to incorporate gene-specific priors by mapping posterior odds of causal association (PO_c) to a Bayes factor for causal association (BF_c). This calculation requires a set of training variants with a known prior. For this training set, we use nonsynonymous variants within GWAS loci and modelling assumptions for their prior. Details of this model are described in the Supplementary Methods and a sensitivity analysis of its assumptions is shown in Extended Data Fig. 8.

Finally, as a preliminary estimate of a principled prior likelihood for genes in the mouse NIDD gene set, we estimated the proportion of non-null associations across all genes in the set. To use true prior data (rather than associations from the current study), we calculated gene-level *P* values for each gene in the set using the MAGENTA[74] algorithm applied to a recent transethnic T2D GWAS[13]. We then used a previously developed approach[40,77] that models the distribution of observed *P* values as a mixture of uniform (representing the null distribution) and beta (representing the non-null distribution) distributions, yielding a prior value of 23.2%.

Our PPA_c calculations currently have several limitations. They apply only to single-variant associations and not (yet) to gene-level associations; extending them to apply to gene-level associations would avoid the possibility of conflicting results among variants within a gene but would require larger-scale gene-level replication data than that we had available in the current analysis. Additional work will also be needed to generate data and develop methods to estimate objective rather than subjective gene priors (researchers can often overestimate evidence of disease relevance for genes in which they have invested considerable effort), to reduce dependence of our conclusions on modelling assumptions (Extended Data Fig. 8) and to explore the extent to which the large number of variant associations that we predict from our data localize to specific gene or variant functional annotations[78].

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

Data availability

Sequence data and phenotypes for this study are available via the database of Genotypes and Phenotypes (dbGAP) and/or the European Genome-phenome Archive, as indicated in Supplementary Table 1.

Code availability

Available for download are scripts for calculating the minimum *P*-value gene-level test, gene set enrichment analyses and the proportion of true associations as a function of variant *P* values.

46. The SIGMA Type 2 Diabetes Consortium. Association of a low-frequency variant in *HNF1A* with type 2 diabetes in a Latino population. *J. Am. Med. Assoc.* **311**, 2305–2314 (2014).
47. Fu, W. et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
48. Lohmueller, K. E. et al. Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes. *Am. J. Hum. Genet.* **93**, 1072–1086 (2013).
49. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
50. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.* **37**, 235–241 (2016).
51. Jagadeesh, K. A. et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48**, 1581–1586 (2016).
52. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
53. Ma, C., Blackwell, T., Boehnke, M., Scott, L. J. & the GoT2D investigators. Recommended joint and meta-analysis strategies for case–control association testing of single low-count variants. *Genet. Epidemiol.* **37**, 539–550 (2013).
54. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
55. The SIGMA Type 2 Diabetes Consortium. Sequence variants in *SLC16A11* are a common risk factor for type 2 diabetes in Mexico. *Nature* **506**, 97–101 (2014).
56. Do, R. et al. Exome sequencing identifies rare *LDLR* and *APOA5* alleles conferring risk for myocardial infarction. *Nature* **518**, 102–106 (2015).
57. Purcell, S. M. et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
58. Li, M. X., Gui, H. S., Kwan, J. S. & Sham, P. C. GATES: a rapid and powerful gene-based association test using extended Simes procedure. *Am. J. Hum. Genet.* **88**, 283–293 (2011).
59. Mahajan, A. et al. Identification and functional characterization of *G6PC2* coding variants influencing glycemic traits define an effector transcript at the *G6PC2-ABCB11* locus. *PLoS Genet.* **11**, e1004876 (2015).
60. Chambers, J. C. et al. Common genetic variation near MC4R is associated with waist circumference and insulin resistance. *Nat. Genet.* **40**, 716–718 (2008).
61. The DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
62. Psaty, B. M. et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ. Cardiovasc. Genet.* **2**, 73–80 (2009).
63. Yu, B. et al. Rare exome sequence variants in *CLCN6* reduce blood pressure levels and hypertension risk. *Circ. Cardiovasc. Genet.* **9**, 64–70 (2016).
64. Brody, J. A. et al. Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology. *Nat. Genet.* **49**, 1560–1563 (2017).
65. Chen, H. et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Hum. Genet.* **98**, 653–666 (2016).
66. Ramatenki, V. et al. Identification of new lead molecules against UBE2NL enzyme for cancer therapy. *Appl. Biochem. Biotechnol.* **182**, 1497–1517 (2017).
67. Gómez-Ramos, A., Podlesniy, P., Soriano, E. & Avila, J. Distinct X-chromosome SNVs from some sporadic AD samples. *Sci. Rep.* **5**, 18012 (2015).

68. Jiang, Y. et al. Six novel rare non-synonymous mutations for migraine without aura identified by exome sequencing. *J. Neurogenet.* **29**, 188–194 (2015).
69. Flannick, J. & Florez, J. C. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat. Rev. Genet.* **17**, 535–549 (2016).
70. Thomsen, S. K. et al. Systematic functional characterization of candidate causal genes for type 2 diabetes risk variants. *Diabetes* **65**, 3805–3811 (2016).
71. Gaulton, K. J. et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).
72. Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 2018).
73. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
74. Segrè, A. V., Groop, L., Mootha, V. K., Daly, M. J. & Altshuler, D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6**, e1001058 (2010).
75. So, H. C., Gui, A. H., Cherny, S. S. & Sham, P. C. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genet. Epidemiol.* **35**, 310–317 (2011).
76. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Optimal designs for two-stage genome-wide association studies. *Genet. Epidemiol.* **31**, 776–788 (2007).
77. Pounds, S. & Morris, S. W. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of *P*-values. *Bioinformatics* **19**, 1236–1242 (2003).
78. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
79. Scott, R. A. et al. An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes* **66**, 2888–2902 (2017).
80. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).

Acknowledgements Studies at the Broad Institute were funded as follows. Sequencing for T2D-GENES cohorts was funded by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) grant U01DK085526 (Multiethnic Study of Type Diabetes Genes) and National Human Genome Research Institute (NHGRI) grant U54HG003067 (Large Scale Sequencing and Analysis of Genomes). Sequencing for GoT2D cohorts was funded by National Institutes of Health (NIH) 1RC2DK088389 (Low-Pass Sequencing and High Density SNP Genotyping in Type 2 Diabetes). Sequencing for ProDiGY cohorts was funded by NIDDK U01DK085526. Sequencing for SIGMA cohorts was funded by the Carlos Slim Foundation (Slim Initiative in Genomic Medicine for the Americas (SIGMA)). Analysis was supported by NIDDK grant U01DK105554 (AMP T2D-GENES Data Coordination Center and Web Portal). The Mount Sinai IPM Biobank Program is supported by The Andrea and Charles Bronfman Philanthropies. The Wake Forest study was supported by NIH R01 DK066358. Oxford cohorts and analysis is funded by The European Commission (ENGAGE: HEALTH-F4-2007-201413); MRC (G0601261, G0900747-91070); NIH (RC2-DK088389, DK085545, R01-DK098032 and U01DK105535); Wellcome Trust (064890, 083948, 085475, 086596, 090367, 090532, 092447, 095101, 095552, 098017, 098381, 100956, 101630 and 203141). The FUSION study is supported by NIH grants DK062370 and DK072193. The research from the Korean cohort was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, South Korea (grant numbers HI14C0060, HI15C1595). The Malmö Preventive Project and the Scania Diabetes Registry were supported by a Swedish Research Council grant (Linné) to the Lund University Diabetes Centre. The Botnia and The PPP-Botnia studies (L.G. and T.T.) have been financially supported by grants from Folkhälsan Research Foundation, the Sigrid Juselius Foundation, The Academy of Finland (grants 263401, 267882 and 312063 to L.G. and 312072 to T.T.), Nordic Center of Excellence in Disease Genetics, EU (EXGENESIS, EUFP7-MOSAIC FP7-600914), Ollqvist Foundation, Swedish Cultural Foundation in Finland, Finnish Diabetes Research Foundation, Foundation for Life and Health in Finland, Signe and Ane Gyllenberg Foundation, Finnish Medical Society, Paavo Nurmi Foundation, Helsinki University Central Hospital Research Foundation, Perklén Foundation, Närpes Health Care Foundation and Ahokas Foundation. The study has also been supported by the Ministry of Education in Finland, Municipal Heath Care Center and Hospital in Jakobstad and Health Care Centers in Vasa, Närpes and Korsholm. The assistance of the Botnia Study Group is acknowledged. This research was supported by contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083 and N01HC85086, and grants U01HL080295 and U01HL130114 from the National Heart, Lung and Blood Institute (NHLBI), with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. The Jackson Heart Study (JHS) is supported by contracts HHSN268201300046C, HHSN268201300047C, HHSN268201300048C, HHSN268201300049C and HHSN268201300050C from the NHLBI and the National Institute on Minority Health and Health Disparities. J.G.W. is supported by U54GM115428 from the National Institute of General Medical Sciences. The Diabetic Cohort (DC) and Multi-Ethnic Cohort (MEC) were supported by individual research grants and clinician scientist award schemes from the National Medical Research Council (NMRC) and the Biomedical Research Council (BMRC) of Singapore. The DC, MEC, Singapore Indian Eye Study (SINDI) and Singapore Prospective Study Program (SP2) were supported by individual research grants and clinician scientist award schemes from the NMRC and the BMRC of Singapore. The Longevity study at Albert Einstein College of Medicine, USA was funded by The American Federation for Aging Research, the Einstein Glenn Center and the NIA (PO1AG027734, R01AG046949, 1R01AG042188 and P30AG038072). The TwinsUK study was funded by the Wellcome Trust and European Community's Seventh Framework Programme (FP7/2007-2013) and received support from the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. Framingham Heart Study is supported by NIH contract NHLBI N01-HC-25195 and HHSN268201500001I. This research was also supported by NIA AG08122 and AG033193, NIDDK U01DK085526, U01DK078616 and K24 DK080140, NHLBI R01 HL105756, and grant supplement R01 HL092577-06S1 for this research. We also acknowledge the dedication of the FHS study participants without whom this research would not be possible. The Mexico City Diabetes Study has been supported by the following grants: RO1HL 24799 from the NHLBI; Consejo Nacional de Ciencia y Tecnología 2092, M9303, F677-M9407, 251M, 2005-C01-14502 and SALUD 2010-2151165; and Consejo Nacional de Ciencia y Tecnología (CONACyT) (Fondo de Cooperación Internacional en Ciencia y Tecnología (FONCICYT) C0012-2014-01-247974). The KARE cohort was supported by grants from Korea Centers for Disease Control and Prevention (4845–301, 4851–302, 4851–307) and an intramural grant from the Korea National Institute of Health (2016-NI73001-00). The Diabetes in Mexico Study was supported by Consejo Nacional de Ciencia y Tecnología grant number S008-2014-1-233970 and by Instituto Carlos Slim de la Salud, AC. The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the NHLBI, NIH, Department of Health and Human Services (contract numbers HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I). We thank the staff and participants of the ARIC study for their important contributions. Funding support for 'Building on GWAS for NHLBI-diseases: the U.S. CHARGE consortium' was provided by the NIH through the American Recovery and Reinvestment Act of 2009 (ARRA) (5RC2HL102419). CHARGE sequencing was carried out at the Baylor College of Medicine Human Genome Sequencing Center (U54 HG003273 and R01HL086694). Funding for GO ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO) and exome sequencing was performed through NHLBI grants RC2 HL-102925 (BroadGO) and RC2 HL-102926 (SeattleGO). The infrastructure for the Analysis Commons is supported by R01HL105756 (NHLBI, to B.M.P.), U01HL130114 (NHLBI, to B.M.P.) and 5RC2HL102419 (NHLBI, to E. Boerwinkle). The LuCAMP project was funded by the Lundbeck Foundation and produced by The Lundbeck Foundation Centre for Applied Medical Genomics in Personalised Disease Prediction, Prevention and Care (http://www.lucamp.org/). The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent Research Center at the University of Copenhagen partially funded by an unrestricted donation from the Novo Nordisk Foundation (https://cbmr.ku.dk/). Further funding came from the Danish Council for Independent Research Medical Sciences. The Inter99 was initiated by T. Jørgensen (principal investigator), K. Borch-Johnsen (co-principal investigator), H. Ibsen and T. F. Thomsen. The steering committee comprises the former two and C. Pisinger. The study was financially supported by research grants from the Danish Research Council, the Danish Centre for Health Technology Assessment, Novo Nordisk, the Research Foundation of Copenhagen County, the Ministry of Internal Affairs and Health, the Danish Heart Foundation, the Danish Pharmaceutical Association, the Augustinus Foundation, the Ib Henriksen Foundation, the Becket Foundation and the Danish Diabetes Association. D.R.W. is supported by the Danish Diabetes Academy, which is funded by the Novo Nordisk Foundation. The KORA study was initiated and financed by the Helmholtz Zentrum München—German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. The NHLBI Exome Sequencing Project (ESP) was supported through the NHLBI Grand Opportunity (GO) program and funded by grants RC2 HL103010 (HeartGO), RC2 HL102923 (LungGO) and RC2 HL102924 (WHISP) for providing data and DNA samples for analysis. The exome sequencing for the NHLBI ESP was supported by NHLBI grants RC2 HL102925 (BroadGO) and RC2 HL102926 (SeattleGO). This research was supported by the Multi-Ethnic Study of Atherosclerosis (MESA) contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079 and UL1-TR-001420. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, TSCI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research (DRC) grant DK063491. The San Antonio Mexican American Family Studies (SAMAFS) are supported by the following grants/institutes. The San Antonio Family Heart Study (SAFHS) and San Antonio Family Diabetes/Gallbladder Study (SAFDGS) were supported by U01DK085524, R01 HL0113323, P01 HL045222, R01 DK047482 and R01 DK053889. The Veterans Administration Genetic Epidemiology Study (VAGES) study was supported by a Veterans Administration Epidemiologic grant. The Family Investigation of Nephropathy and Diabetes - San Antonio (FIND-SA)

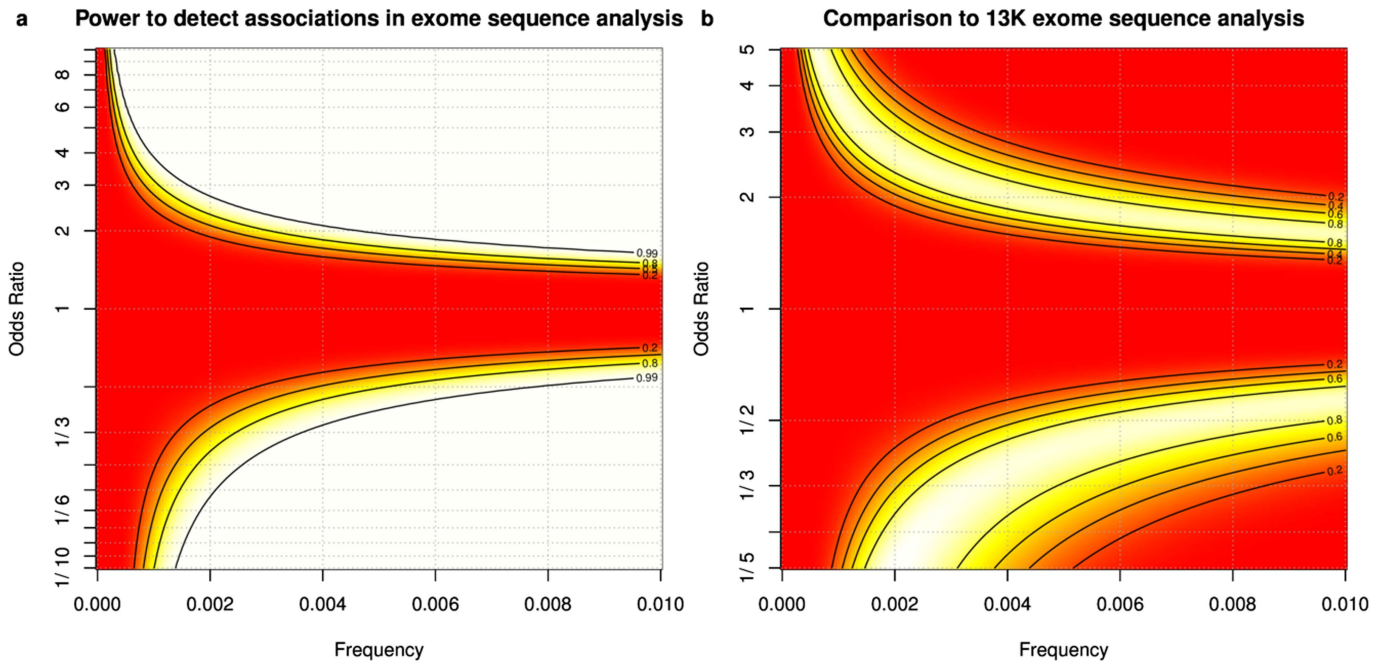
study was supported by NIH grant U01DK57295. The SAMAFS research team acknowledges the contributions of H. E. Abboud to the research activities of the SAMAFS. Sample collection, research and analysis from the Hong Kong Diabetes Register (HKDR) at the Chinese University of Hong Kong (CUHK) were supported by the Hong Kong Foundation for Research and Development in Diabetes established under the auspices of the Chinese University of Hong Kong, the Hong Kong Government Research Grants Committee Central Allocation Scheme (CUHK 1/04C), a Research Grants Council Earmarked Research Grant (CUHK4724/07M), the Innovation and Technology Fund (ITS/088/08 and ITS/487/09FP) and the Research Grants Committee Theme-based Research Scheme (T12-402/13N). The TODAY contribution to this study was completed with funding from NIDDK and the NIH Office of the Director (OD) through grants U01DK61212, U01DK61230, U01DK61239, U01DK61242 and U01DK61254; from the National Center for Research Resources General Clinical Research Centers Program grants M01-RR00036 (Washington University School of Medicine), M01-RR00043-45 (Children's Hospital Los Angeles), M01-RR00069 (University of Colorado Denver), M01-RR00084 (Children's Hospital of Pittsburgh), M01-RR01066 (Massachusetts General Hospital), M01-RR00125 (Yale University) and M01-RR14467 (University of Oklahoma Health Sciences Center); and from the NCRR Clinical and Translational Science Awards grants UL1-RR024134 (Children's Hospital of Philadelphia), UL1-RR024139 (Yale University), UL1-RR024153 (Children's Hospital of Pittsburgh), UL1-RR024989 (Case Western Reserve University), UL1-RR024992 (Washington University in St Louis), UL1-RR025758 (Massachusetts General Hospital) and UL1-RR025780 (University of Colorado Denver). The Pakistan Genetic Resource (PGR) is funded through endowments awarded to CNCD, Pakistan. J.F. is supported by BADERC DK057521. R.L. is supported by the NIH (R01DK110113, U01HG007417, R01DK101855 and R01DK107786). A.P.M. is supported by the NIH-NIDDK (U01DK105535); and a Wellcome Trust Senior Fellow in Basic Biomedical Science (award WT098017). J.C.F. is supported by NIDDK K24 DK110550 and P30 DK057521. G.I.B. is supported by P30 DK020595. Y.S.C. acknowledges support from the National Research Foundation of Korea (NRF) grant (NRF-2017R1A2B4006508). C.-Y.C. is supported by Clinician Scientist Award (NMRC/CSA-SI/0012/2017) of the Singapore Ministry of Health's National Medical Research Council. R.C.W.M. and J.C. acknowledges support from the Hong Kong Research Grants Council Theme-based Research Scheme (T12-402/13N), Research Grants Council General Research Fund (14110415), the Focused Innovation Scheme, the Vice-Chancellor One-off Discretionary Fund, the Postdoctoral Fellowship Scheme of the Chinese University of Hong Kong, as well as the Chinese University of Hong Kong-Shanghai Jiao Tong University Joint Research Collaboration Fund. We thank all medical and nursing staff of the Prince of Wales Hospital Diabetes Mellitus Education Centre, Hong Kong. LuCAMP thanks A. Forman, T. H. Lorentzen and G. J. Klavsen for laboratory assistance, P. Sandbeck for data management, G. Lademann for secretarial support and T. F. Toldsted for grant management. We thank study participants of the DC, MEC, SINDI and SP2 for their contributions and the National University Hospital Tissue Repository (NUHTR). We thank the Jackson Heart Study (JHS) participants and staff for their contributions to this work. This study was provided with biospecimens and data from the Korean Genome Analysis Project (4845-301), the Korean Genome and Epidemiology Study (4851-302) and the Korea Biobank Project (4851-307, KBP-2013-11 and KBP-2014-68) that were supported by the Korea Centers for Disease Control and Prevention, South Korea. The Pakistan Genomic Resource (PGR) thank all the study participants for their participation. For this publication, biosamples from the KORA Biobank as part of the Joint Biobank Munich (JBM) have been used. M.I.M. is a Wellcome Trust Senior Investigator (WT098381) and a National Institute of Health Research (NIHR) Senior Investigator; the views expressed in this article are his views and not necessarily those of the NHS, the NIHR, or the Department of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions J.C.F., M.I.M. and M.B. contributed equally to this work. J.F., N.P.B., J.C.F., M.I.M. and M.B. provided leadership. J.F., J.M.M., C.F., M.S.U., A. Mahajan, J.W., T.M.T., T.W.B., L. Chen, S.C., A.E., A.U.J., K.M., A.N., A.J.P., N.W.R., N.R.R., H.M.S., J.M.T., R.P.W., L.J.S. and A.P.M. analysed data. L. Caulkins, R.K. and M.C. provided project management and support. Members of the Broad Genomics Platform Consortium contributed to the data generation for the indicated studies. A.C., R.A.D., S.G., S.H., H.M.K., B.-J.K., H.A.K., J.K., J. Liu, K.L.M., M.C.Y.N., M.P., R.S.V., C.S., W.Y.S., C.H.T.T., F.T., B.T., R.M.v.D., M.V., T.-Y.W., G. Atzmon, N.B., J.B., D.W.B., J.C.C., E. Chan, C.-Y.C., Y.S.C., F.S.C., R.D., B.G., J.S.K., S.H.K., M.L., D.M.L., E.S.T., J.T., J.G.W., E. Bottinger, J.C., J.D., P.F., M.Y.H., Y.J.K., J.-Y.L., J. Lee, R.L., R.C.W.M., A.D.M., C.N.A.P., K.S.P., A.R., D.S., X. Sim, Y.Y.T., C.L.H., G. Abecasis, G.I.B., N.J.C., M.S., R.S., J.B.M. and D.A. provided data and analysis from the T2D-GENES study. V.L., L.L.B., L.G., P.N., T.D.S., T.T. and K.S.S. provided data and analysis from the GoT2D study. M.E.J., A.L., D.R.W., N.G., T.H. and O.P. provided data and analysis from the LuCAMP study. L.D., K.L.D., M.K., E.M.-D., C.P., N.S., B.B., P.Z. and D.D. provided data and analysis from the ProDiGY study. F.B.-O., F.C.-C., C.C.-C., E. Córdova, M.E.G.-S., H.G.-O., J.M.M.-H., A.M.-H., E.M.-C., C.R.-M., C. Gonzalez, M.E.G., C.A.A.-S., C.H., B.E.H., L.O., X. Soberón and T.T.-L. provided data and analysis from the SIGMA study. J.W., E. Boerwinkle, J.A.B., J.S.F., N.L.H.-C., C.-T.L., A.K.M., A.C.M., B.M.P., S.W., P.S.d.V., J.D., S.R.H., C.J.O., J.P. and J.B.M. provided data and analysis from the CHARGE study. T.M.T., J.B.L., A. Marcketta, C.O., D.J.C., H.L.K., F.E.D., A.B. and D.J.C. provided data and analysis from the Regeneron study. T.M.S., C. Gieger, T.M. and K.S. provided data and analysis from the KORA study. E. Boerwinkle, M.G., N.L.H.-C., A.C.M., W.S.P., B.M.P., A.P.R., R.P.T., C.J.O., L.L., S.R. and J.I.R. provided data and analysis from the ESP study.

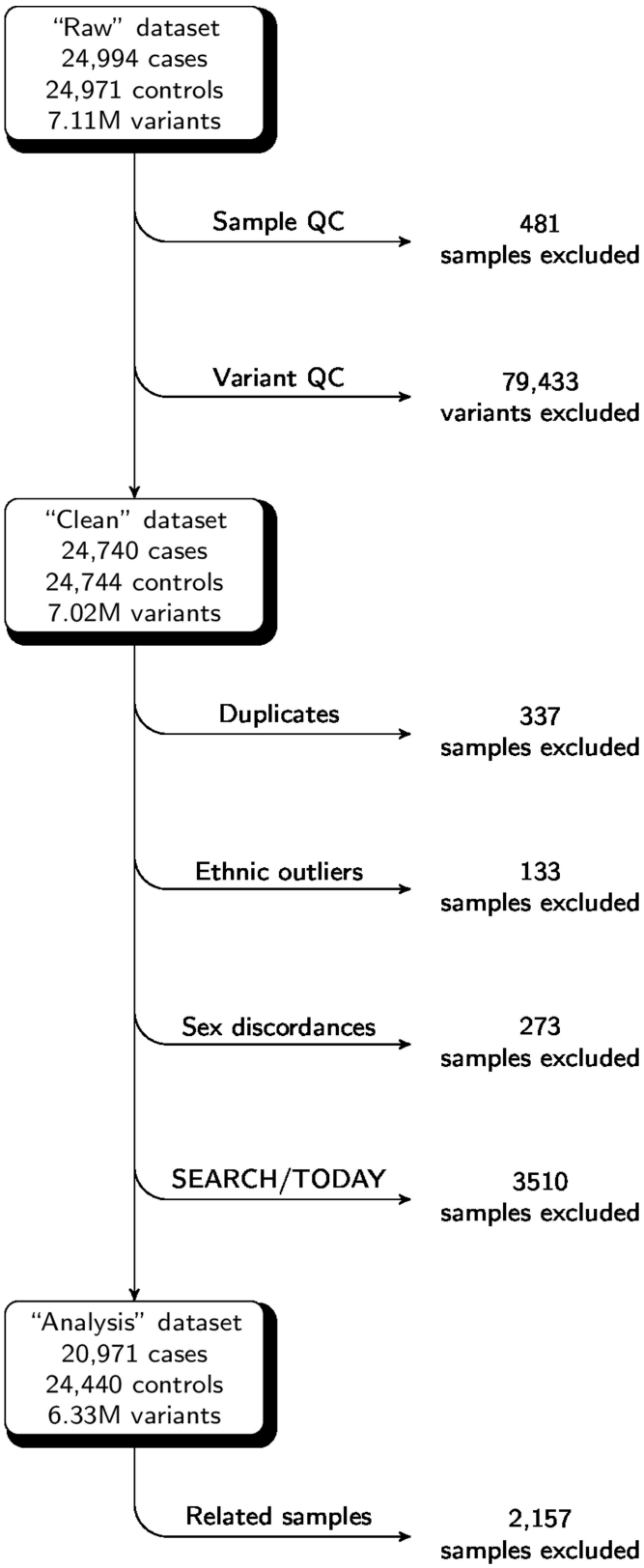
Competing interests P.Z. is a consultant for Merck, Daichii-Sankyo, Boerhinger-Ingelheim and Janssen; B.M.P. serves on the DSMB of a clinical trial funded by Zoll LifeCor and on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson.

Additional information

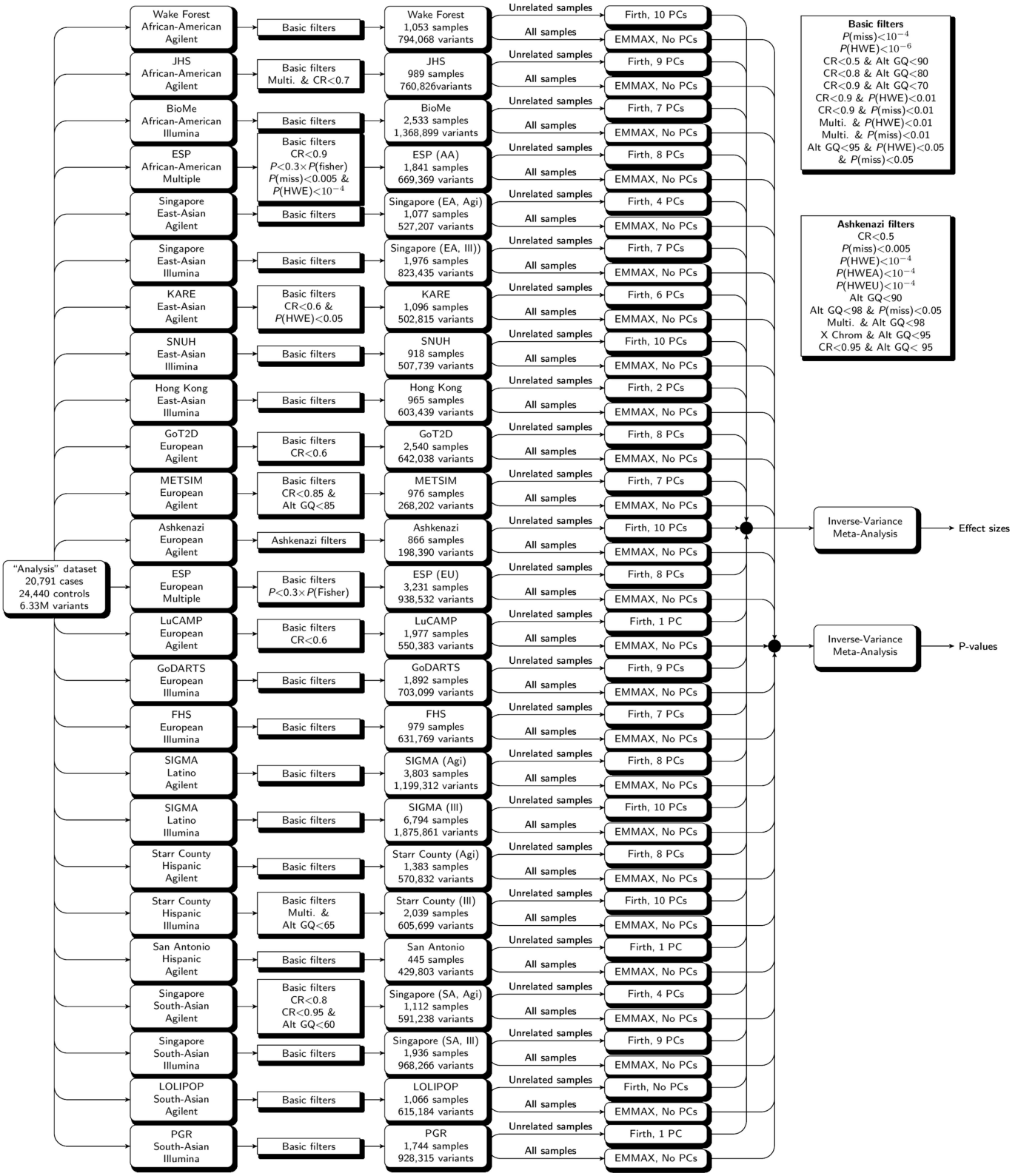
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1231-2.
Peer review information *Nature* thanks Braxton Mitchell and the other anonymous reviewer(s) for their contribution to the peer review of this work.



Extended Data Fig. 1 | Power analysis. The power to detect associations (using a two-sided test) at $P < 5 \times 10^{-8}$ for variants (or collections of variants) with a given minor allele frequency (x axis) and odds ratio (y axis) measured as the average across all ancestries. **a**, Cells are shaded according to the power of the current study of 20,791 T2D cases and 24,440 controls, with white indicating high power and red indicating low power. The 20%, 50%, 80% and 99% contour lines are labelled. **b**, Cells are shaded according to the difference in power between the current study and a previously published study of 12,940 individuals[10], with yellow–white indicating a large increase in power and red indicating a small increase in power. The 20%, 40%, 60% and 80% contour lines are labelled.

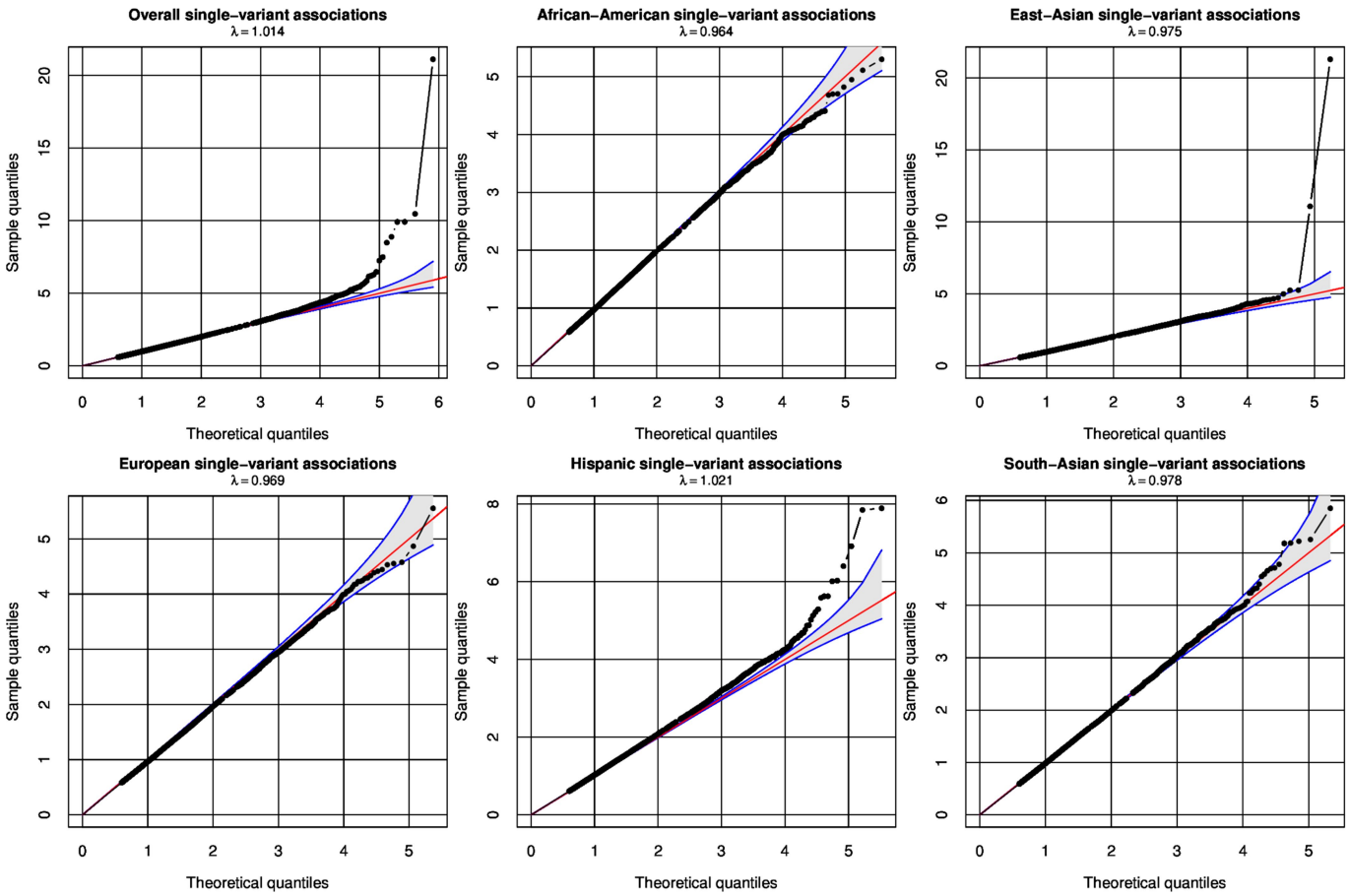


Extended Data Fig. 2 | Data quality control workflow. A schematic of the steps involved in sample and variant quality control is shown. Quality control was conducted as described in the Methods to construct a set of samples and variants included in the association analysis. Each step is depicted as an arrow, with the number of samples or variants excluded by the step shown at the end of the arrow. The final set of samples and variants analysed are represented by the 'Analysis' dataset; we further excluded samples of high relatedness to other samples in the dataset from some but not all analyses. After each step that removed samples, we also removed newly monomorphic variants (hence the decrease in variants between the 'Clean' and 'Analysis' datasets).

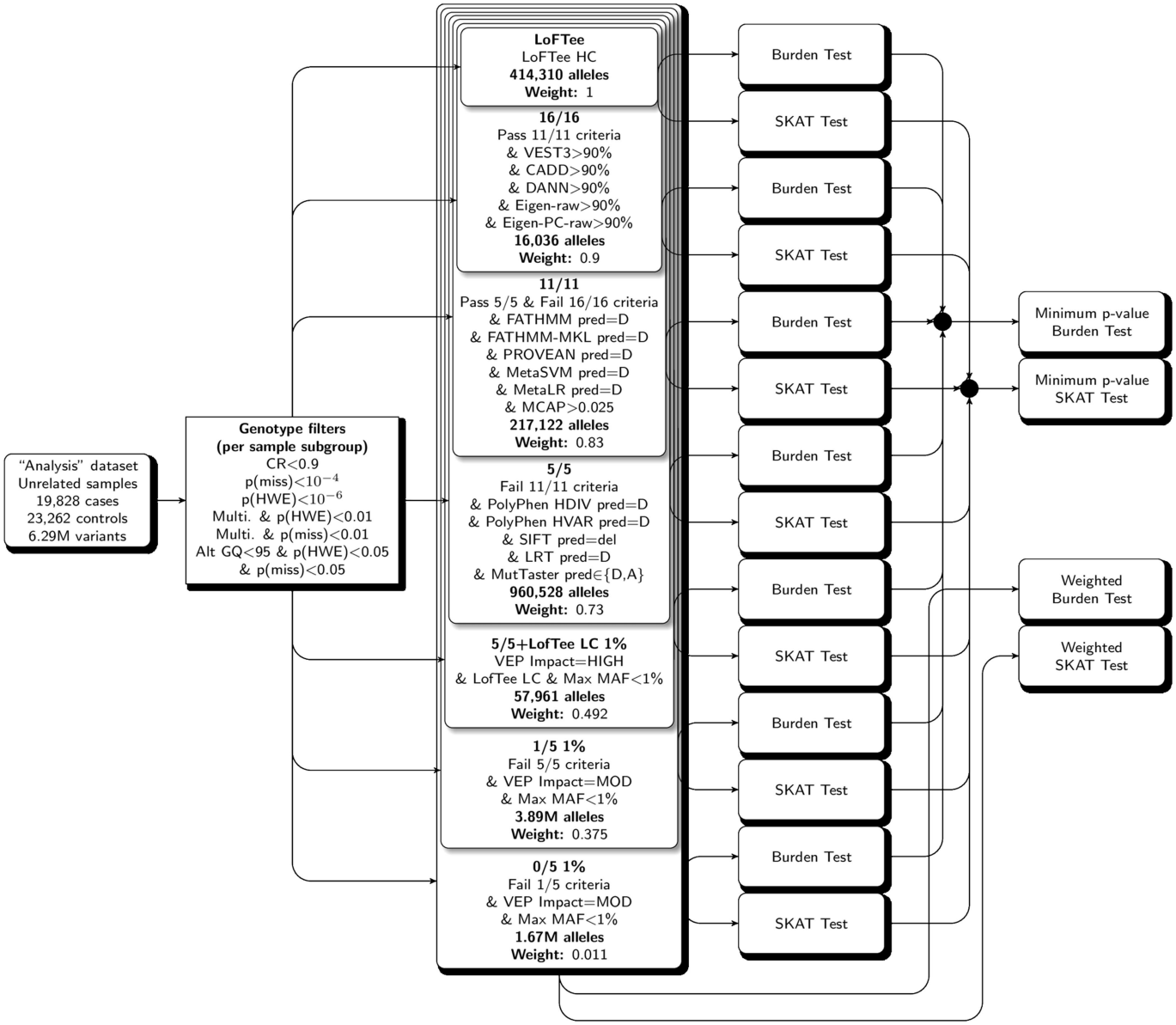


Extended Data Fig. 3 | See next page for caption.

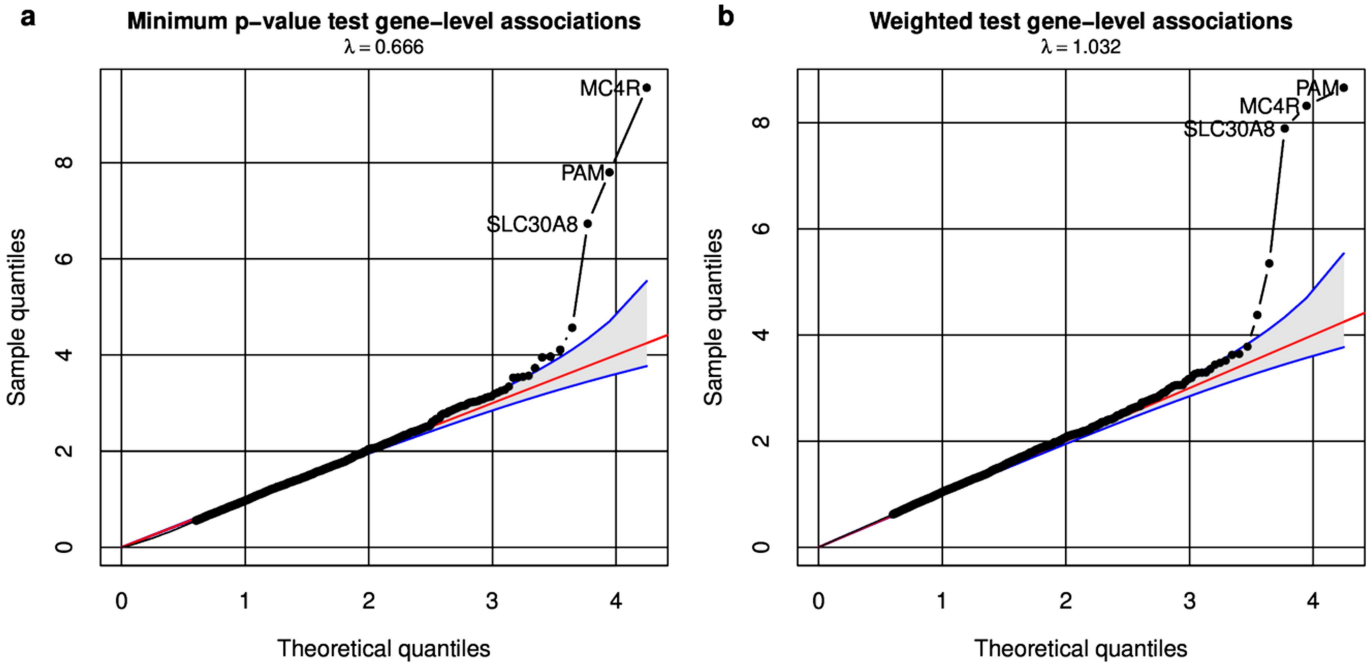
Extended Data Fig. 3 | Single-variant association analysis workflow. A schematic of the steps involved in single-variant exome-sequencing association analysis is shown, as described in the Methods. We began analysis with a division of samples in the 'Analysis' dataset (leftmost column) into 25 different subgroups (second column from the left) based on cohort, ancestry and sequencing technology. We then filtered variants according to metrics computed separately for each subgroup; we applied the filters listed in the 'Basic filters' box to all subgroups and for some subgroups we applied additional (more stringent) filters as indicated by boxes in the third column from the left. The resulting number of variants and samples advanced for analysis in each subgroup are indicated in the fourth column from the left. We analysed each subgroup with both the EMMAX test (to measure association strength) and the Firth test (to measure allelic odds ratios), each of which are two-sided; the number of principal components included as covariates in the Firth test is shown in the fifth column from the left. Finally, we combined each of the EMMAX and Firth subgroup-level results using a 25-group meta-analysis to produce the final *P* values and odds ratios reported for each variant. Multi, variant is multiallelic; CR, call rate; *P*, variant subgroup-level *P* value; *P*(Fisher), variant subgroup-level *P* value from Fisher's exact test; *P*(miss), *P* value for subgroup-level variant differential missingness between T2D cases and controls; *P*(HWE), *P* value for deviation from subgroup-level Hardy–Weinberg equilibrium; Alt GQ, mean genotype quality of non-reference genotypes (across all samples); X Chrom, variant is on X chromosome.



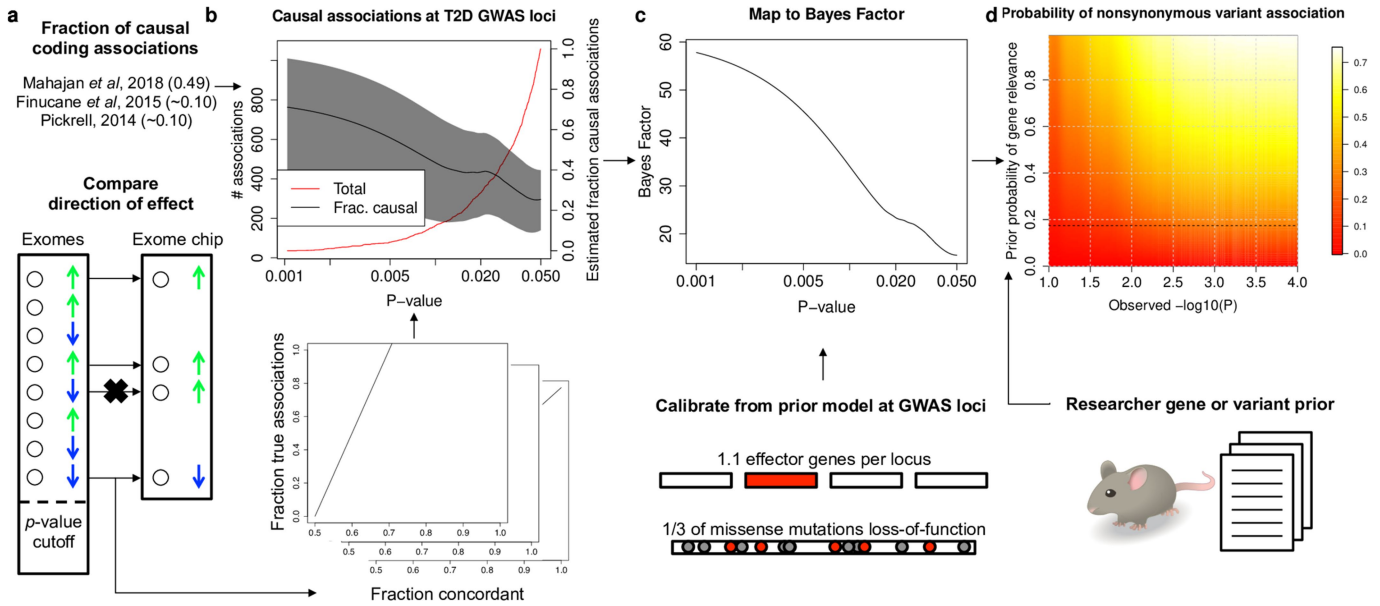
Extended Data Fig. 4 | Calibration of single-variant analysis. To assess whether our single-variant association statistics (two-sided, calculated by the EMMAX test) were well-calibrated, we computed quantile–quantile plots of associations across all samples (Overall) and within each ancestry (total $n = 45{,}231$ individuals). To avoid deflation of the quantile–quantile plot from rare variants (for which the expected P values are discrete rather than uniformly distributed), only variants with minor allele counts of 20 or greater (either overall or within the relevant ancestry) are shown. Variants were also LD-pruned before plotting, to avoid induced variance from correlated P values of these variants, using the 'clump' method implemented in PLINK. The λ values indicate genomic control, as measured by the ratio in observed median χ^2 statistic to that expected under the null hypothesis. Red line, expectation of P values under the null distribution. Blue lines (and grey region), 95% confidence interval of expectations under the null distribution.



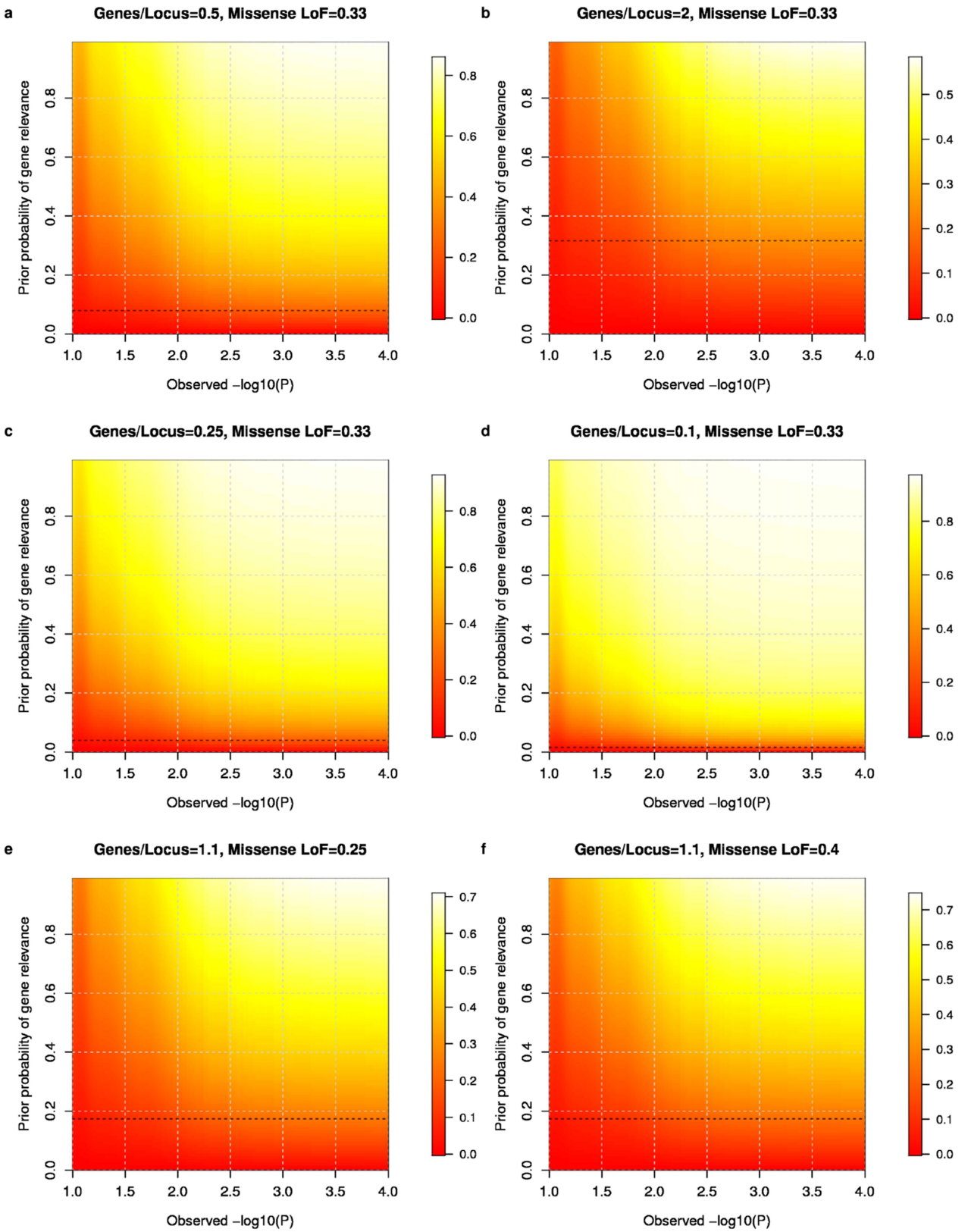
Extended Data Fig. 5 | Gene-level association analysis workflow. A schematic of the steps involved in gene-level exome-sequencing association analysis, as described in Methods, is shown. We began analysis with subgroup-level genotype filtering (second column from the left) of unrelated samples in the 'Analysis' dataset (leftmost column); we then applied genotype filters for each subgroup (filtering genotypes for either all or no samples in each subgroup), similar to those used in subgroup-level single-variant analyses. We then annotated each non-reference variant allele with 16 different bioinformatics algorithms to assess allele deleteriousness, and we grouped alleles into one of seven nested masks (third column from the left; the number of variants and weights shown correspond to alleles absent from 'higher', or more stringent, nested masks). We computed burden and SKAT analyses (both of which are two-sided) using one of two approaches to combine alleles across masks (Methods): first, by analysing all alleles at once with weights assigned according to the most stringent mask containing the allele (weighted test); and second, by analysing each mask independently and then calculating the lowest *P* value corrected for the effective number of tests (minimum *P*-value test). Multi, variant is multiallelic; CR, call rate; *P*(miss), *P* value for subgroup-level variant differential missingness between T2D cases and controls; *P*(HWE), *P* value for deviation from subgroup-level Hardy–Weinberg equilibrium; Alt GQ, mean genotype quality of non-reference genotypes (across all samples).



Extended Data Fig. 6 | Calibration of gene-level association analyses. For both the burden and SKAT tests, we tested for gene-level association within seven different allelic masks. As this produced seven *P* values for each test, we developed two means to consolidate these results (Methods). **a**, **b**, The quantile–quantile plots of associations are shown for the minimum *P*-value burden test (**a**) and the weighted burden test (**b**). Each test is two-sided and consists of $n = 43{,}071$ unrelated individuals. Only genes with combined minor allele count of 20 or greater are shown in the quantile–quantile plots, to avoid deflation from genes with too few variants to produce *P* values asymptotically uniform under the null distribution. The λ values indicate genomic control, as measured by the ratio in observed median χ^2 statistic to that expected under the null hypothesis. The three genes with exome-wide significant associations are labelled. Red line, expectation of *P* values under the null distribution. Blue lines (and grey region), 95% confidence interval of expectations under the null distribution.



Extended Data Fig. 7 | PPA calculation workflow. a, We estimated the PPAs for nonsynonymous variants in our sequence analysis based on concordance with independent exome array data and previously published[6,78,80] estimates of the fraction of causal coding associations (Methods). **b**, PPA estimates for nonsynonymous variants within T2D GWAS loci are shown as a function of *P* value (right *y* axis, black line; 95% confidence interval, grey shading) together with the total number of such variants (left *y* axis, red line). **c**, For variants outside of T2D GWAS loci, we developed a method to further compute Bayes factors, which measure the odds of true and causal association, as a function of *P* value, using a model of the prior odds of true and causal association for variants in GWAS loci (Methods). **d**, These Bayes factors can be combined with a subjective prior belief in the T2D relevance of a gene (*y* axis) to produce the estimated posterior probability of true and causal association for any nonsynonymous variant in the exome-sequence dataset based on its observed $\log_{10}(P)$ (*x* axis). Posterior estimates are shaded proportional to value (red, low; white, high). Values are shown for the default modelling assumptions of 33% of missense variants that caused gene inactivation and 30% of true missense associations that represented the causal variant.



Extended Data Fig. 8 | Estimated posterior probability of associations for different prior hypotheses. We estimated the posterior probability of association for nonsynonymous variants that met various single-variant *P*-value thresholds (two-sided EMMAX test, $n = 45{,}231$ individuals) in our analysis, as described in the Methods and shown in Extended Data Fig. 7. To perform the needed calculations, we assumed that, on average, 1.1 genes that are found within each T2D GWAS locus are relevant to T2D and 33% of missense mutations within these genes cause gene loss-of-function. **a**–**f**, To assess the sensitivity of our analysis to these assumptions, we repeated the calculations with different assumptions of 0.5 (**a**), 2.0 (**b**), 0.25 (**c**) and 0.1 (**d**) T2D-relevant genes within each GWAS locus, as well as 25% (**e**) and 40% (**f**) of missense variants leading to loss-of-function. All analyses assume the default modelling parameters that 30% of true nonsynonymous associations are causal associations; different values for this parameter would scale posterior probability estimates linearly.

Extended Data Table 1 | Most significant single-variant associations from exome-sequencing analysis

Gene	Variant	Consequence	Impact	Change	MAF	Case	Ctrl	OR	P	Ref	Ref P
PAX4	rs2233580	missense_variant	Med.	p.Arg192His	0.12	890	563	1.7	7.6e-22	[6]	1.8e-12
SLC30A8	rs13266634	missense_variant	Med.	p.Arg325Trp	0.43	12258	13756	0.897	3.4e-11	[6]	1.8e-47
WFS1	rs1801212	missense_variant	Med.	p.Val333Ile	0.27	7101	8456	1.13	1.2e-10	[6]	1.1e-24
KCNJ11	rs5219	missense_variant	Med.	p.Lys23Glu	0.39	16471	15959	0.898	1.2e-10	[6]	5.7e-22
KCNJ11	rs5215	missense_variant	Med.	p.Val250Ile	0.39	16687	16132	0.901	3.4e-10	[6]	4.5e-21
SLC16A11	rs2292351	5_prime_UTR_variant	Low	-	0.34	5244	4249	1.25	1.3e-09	[80]	0.26
SLC16A11	rs13342692	missense_variant	Med.	p.Asp127Gly	0.34	9468	7492	1.12	1.7e-09	[6]	5.7e-05
SLC16A11	rs75493593	missense_variant	Med.	p.Pro443Thr	0.3	6262	4929	1.24	3.2e-09	[80]	0.24
WFS1	rs1801213	synonymous_variant	Low	p.Arg228Arg	0.33	10641	11689	1.1	3.7e-09	[80]	1.e-10
SLC16A13	rs76070643	synonymous_variant	Low	p.Tyr166Tyr	0.3	6357	5028	1.2	1.8e-08	[80]	0.29
WFS1	rs1046317	3_prime_UTR_variant	Mod.	-	0.32	9957	11005	1.09	2.0e-08	[13]	1.3e-09
SFI1	rs145181683	missense_variant	Med.	p.Arg724Trp	0.16	2861	2144	1.19	3.2e-08	[6]	0.3
WFS1	rs9998519	intron_variant	Mod.	-	0.39	13395	14741	1.08	4.3e-08	[80]	2.7e-13
WFS1	rs10010131	intron_variant	Mod.	-	0.39	13046	14406	1.08	5.6e-08	[13]	4.e-09
ABCC8	rs757110	missense_variant	Med.	p.Ala1369Ser	0.39	16626	16237	0.913	7.1e-08	[6]	8.1e-19
WFS1	rs1801214	synonymous_variant	Low	p.Asn500Asn	0.38	12841	14187	1.08	1.6e-07	[6]	1.e-22
MC4R	rs79783591	missense_variant	Med.	p.Ile269Asn	0.0089	195	83	2.17	3.4e-07	[6]	0.075
WFS1	rs1801206	synonymous_variant	Low	p.Val395Val	0.53	15408	16499	1.08	3.7e-07	[13]	0.0055
WFS1	rs1046316	synonymous_variant	Low	p.Ser855Ser	0.32	10412	11572	1.08	5.3e-07	[80]	2.3e-11
COBLL1	rs7607980	missense_variant	Med.	p.Asn939Asp	0.15	4010	4651	0.857	6.3e-07	[6]	8.6e-20
PISD	rs12171042	downstream_gene_variant	Mod.	-	0.53	15797	15264	1.09	7.0e-07	[6]	0.00037
PAM	rs35658696	missense_variant	Med.	p.Asp563Gly	0.05	1038	944	1.29	1.3e-06	[6]	1.2e-16
PPIP5K2	rs36046591	missense_variant	Med.	p.Ser1207Gly	0.049	986	905	1.3	1.4e-06	[6]	1.4e-16
RAI1	rs3818717	synonymous_variant	Low	p.Ile1867Ile	0.55	14691	16514	0.927	1.8e-06	[6]	0.00049
PPIP5K2	rs116234738	3_prime_UTR_variant	Mod.	-	0.055	956	894	1.29	2.3e-06	[80]	7.2e-05
MAEA	rs2272481	intron_variant	Mod.	-	0.42	10249	10165	0.898	2.7e-06	[80]	0.02
COBLL1	rs34305002	intron_variant	Mod.	-	0.25	3646	4625	0.863	3.3e-06	-	-
PIK3C2B	rs1553921	synonymous_variant	Low	p.Leu96Leu	0.64	10215	8702	0.912	3.5e-06	[6]	0.023
WDR13	X_48460357	splice_acceptor_variant	Low	-	0.0006	21	2	3.48	3.6e-06	-	-
TMCC2	rs1768586	synonymous_variant	Low	p.Ala315Ala	0.36	9113	10374	0.924	4.0e-06	[13]	0.95
MDM4	rs4252717	intron_variant	Mod.	-	0.42	19786	19287	0.934	4.4e-06	[6]	0.02
MDM4	rs2290854	intron_variant	Mod.	-	0.72	20466	19245	0.935	4.5e-06	[13]	0.51
GCKR	rs1260326	missense_variant	Med.	p.Leu446Pro	0.5	15010	16627	1.07	5.4e-06	[6]	5.3e-25
CDC123	rs12590	3_prime_UTR_variant	Mod.	-	0.24	9078	8543	1.11	5.6e-06	[80]	0.00046
TMCC2	rs1668870	intron_variant	Mod.	-	0.36	8368	9585	0.919	5.7e-06	[80]	0.85
ANKRD36C	rs188178234	missense_variant	Med.	p.Arg786Trp	0.012	265	364	0.713	6.2e-06	[80]	0.27
TMCC2	rs1668867	synonymous_variant	Low	p.Tyr562Tyr	0.36	9505	10664	0.926	6.8e-06	[13]	0.51
PIK3C2B	rs1124777	synonymous_variant	Low	p.Pro199Pro	0.64	10186	8702	0.915	6.9e-06	[6]	0.019
SIN3A	rs4886696	intron_variant	Mod.	-	0.47	13368	15391	1.08	7.9e-06	[80]	0.1
ZRANB2	rs11556475	synonymous_variant	Low	p.Tyr114Tyr	0.16	3365	3068	1.12	8.3e-06	[80]	0.021
CDC123	rs1051055	3_prime_UTR_variant	Mod.	-	0.38	14476	13738	0.902	9.2e-06	[13]	0.0071
TGFB1	rs11466334	intron_variant	Mod.	-	0.082	877	555	1.31	9.5e-06	-	-
WFS1	rs1046314	synonymous_variant	Low	p.Lys811Lys	0.49	15003	16045	1.06	9.5e-06	[13]	4.3e-09
SLC26A3	rs117703371	splice_acceptor_variant	Low	p.Ser438Ser	0.0084	236	169	1.46	1.0e-05	[80]	0.082
BICD1	rs183649090	intron_variant	Mod.	-	0.0019	28	73	0.426	1.1e-05	-	-
RNGTT	rs6937994	3_prime_UTR_variant	Mod.	-	0.075	598	664	0.725	1.1e-05	-	-
SNX9	rs80009789	intron_variant	Mod.	-	0.069	740	816	0.783	1.2e-05	[80]	0.26
PRR14L	rs131224	upstream_gene_variant	Mod.	-	0.19	3986	2813	1.14	1.2e-05	[13]	0.59
SCN9A	rs12478318	downstream_gene_variant	Mod.	-	0.22	4367	3522	1.13	1.2e-05	[6]	0.48
PTGER3	rs5671	synonymous_variant	Low	p.Thr206Thr	0.16	3386	3081	1.15	1.3e-05	[13]	0.076

The most significant results from exome-sequencing single-variant association analyses are shown ($n = 45{,}231$ individuals). Gene, the closest gene to the variant. Variant, a unique identifier for the variant within our exome-sequencing analysis. Consequence, the predicted consequence of the variant, defined by sequence ontology annotation and produced using VEP. Impact, the effect of the variant, as predicted using VEP (Med, medium; Mod, modifier). Change, the predicted protein change, defined according to the 'best guess' transcript as described in the Methods. MAFs are calculated as the maximum across all ancestries. Case, the number of samples with T2D carrying the variant. Ctrl, the number of samples without T2D carrying the variant. OR, the odds ratio, calculated from the Firth analysis. P, the *P* value, calculated from the (two-sided) EMMAX analysis. Ref P: the *P* value of the variant in one of three previous GWAS or exome array analyses, referenced in the Ref column[6,13,79].

Extended Data Table 2 | Most significant gene-level associations from exome-sequencing analysis

			Burden				SKAT	
	Best Result		Min P		Weighted		Min P	Weighted
Gene	Var	CAF	OR	P	OR	P	P	P
MC4R	41	0.00795	2.07	2.74e-10	2.2	4.81e-09	7.74e-08	3.48e-08
PAM	79	0.0493	1.31	1.58e-08	1.44	2.2e-09	1.53e-07	7.03e-08
SLC30A8	86	0.0116	0.598	1.85e-07	0.397	1.29e-08	0.00011	0.000221
IGFBPL1	33	0.00522	0.564	0.000108	0.208	4.5e-06	0.0222	0.00114
BICD1	188	0.0163	0.857	0.214	0.85	0.575	1.49e-05	0.632
UBE2NL	5	0.000417	12.7	2.71e-05	1.66	0.115	0.00963	0.29
ING3	55	0.00255	2.29	0.000112	7.03	3.47e-05	0.0268	0.0135
HNF1A	131	0.0184	1.23	0.0219	1.47	0.0125	3.62e-05	0.00106
NUMA1	147	0.0459	1.14	0.0249	1.08	0.202	0.000129	4.27e-05
MAP3K15	256	0.0392	0.85	7.76e-05	0.777	0.00239	0.0035	0.12
PDX1	11	0.000371	4.71	0.0214	3.46	0.000166	0.165	0.0573
DPH7	31	0.00631	1.27	0.186	1.3	0.0874	0.000818	0.000184
MGAT4C	28	0.00134	3.13	0.000187	1.9	0.00303	0.00823	0.000886
TMEM216	3	0.301	1.08	0.000294	1.08	0.000207	0.573	0.584
HYAL2	74	0.00308	0.503	0.000728	0.292	0.000228	0.0554	0.024
MBD3	9	0.000649	3.94	0.00415	4.9	0.000239	0.0137	0.00672
SOCS2	37	0.00334	1.96	0.000271	5.16	0.000352	0.00539	0.00334
SLC16A2	57	0.262	0.935	0.000285	0.945	0.000639	0.411	0.575
PTPRC	274	0.033	1.23	0.000297	1.48	0.0162	0.0363	0.11
ANGPTL6	38	0.00598	0.73	0.0585	0.744	0.0443	0.00102	0.0003
STAT4	6	0.00102	0.677	0.00624	0.364	0.000305	0.00294	0.000527
PCBP1	13	0.000371	0.0983	0.000568	0.00288	0.000333	0.243	0.165
MAGEB5	25	0.00183	0.53	0.000454	0.138	0.000885	1.	1.
ARHGEF7	107	0.0201	1.26	0.608	1.18	0.499	0.00165	0.000458
SLC48A1	12	0.000951	3.22	0.000964	22.6	0.000519	0.936	0.777
DGKK	147	0.0121	0.862	0.0519	1.01	0.975	0.00052	0.179
TDRD5	20	0.00805	0.674	0.00203	0.617	0.000521	0.00331	0.00114
TCEA1	12	0.000441	0.0944	0.00053	0.531	0.0331	0.0612	0.354
RXRG	22	0.000788	3.39	0.00291	3.6	0.000546	0.195	0.126
KDM7A	110	0.0136	1.37	0.000559	2.03	0.0017	0.394	0.22
C19orf66	49	0.00413	1.7	0.00165	3.86	0.000566	0.131	0.0712
PDF	27	0.0029	0.578	0.00399	0.776	0.641	0.000584	0.75
OR2M4	10	0.00415	0.554	0.000607	0.699	0.00607	0.0411	0.0263
RHBDD2	8	0.000557	0.721	0.0642	0.458	0.0181	0.00559	0.000626
HLCS	66	0.00364	1.91	0.000644	1.42	0.0278	0.161	0.167
NT5DC1	92	0.0123	1.21	0.0905	1.46	0.0633	0.00276	0.000653
TMEM161B	84	0.00487	1.56	0.00856	2.97	0.000664	0.166	0.0599
LRRTM3	80	0.0083	1.44	0.00078	2.54	0.000884	0.000678	0.000673
PRSS54	8	0.000719	0.569	0.295	0.923	0.646	0.000717	0.0802
MEST	29	0.00148	2.48	0.00341	5.64	0.000743	0.216	0.0942
MIEF1	102	0.0501	1.17	0.000751	1.28	0.000891	0.00996	0.0305
RUVBL2	84	0.00582	1.58	0.00122	2.49	0.000803	0.246	0.0842
FLCN	124	0.0111	0.702	0.00245	0.485	0.000816	0.706	0.945
DHX9	11	0.000441	0.0993	0.000829	0.569	0.153	0.037	0.00564
TRDN	131	0.0227	1.28	0.000854	1.17	0.598	0.00107	0.0517
ABCB11	226	0.0231	1.21	0.0161	1.55	0.000881	0.261	0.0236
MAGEE2	119	0.278	1.07	0.000907	1.06	0.00241	0.473	0.229
ANKS3	186	0.0255	0.808	0.00287	0.664	0.000908	0.614	0.232
DHRS13	73	0.00853	1.66	0.0899	1.3	0.101	0.000917	0.0054
PPP3CA	41	0.00445	0.745	0.0489	0.517	0.0712	0.000921	0.000992

The most significant results from gene-level analyses are shown ($n = 43{,}071$ unrelated individuals). Genes are ranked according the (two-sided) minimum *P* value achieved across the four gene-level analyses. Gene, a unique identifier for the gene within our exome-sequencing analysis. Var (CAF), the number of alleles (combined allele frequency (CAF) of variants) in the mask achieving the strongest association across the four tests (that is, the 0/5 1% mask for the weighted test, or the mask with the minimum *P* value for the minimum *P*-value test). Burden, results from the (two-sided) burden analysis. SKAT, results from the (two-sided) SKAT analysis. Min P, results from the (two-sided) minimum *P*-value analysis. Weighted, results from the (two-sided) weighted analysis. OR, the odds ratio as estimated from the burden analysis. P, the (two-sided) *P* value for the indicated analyses.

nature research

Corresponding author(s): Jason Flannick

Last updated by author(s): Mar 5, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used in data collection
Data analysis	All software used in the analysis was open source and is described in the Methods section of the manuscript. Existing software packages used were: RTA v2.7.3, BWA v0.7 , Picard v1, GATK v3.4, VEP v87, Plink 1.9, EPACTS v3.2.4, MetaXcan v0.3, DAPPLE, MAGENTA v2.4, R v3.4, Michigan Imputation Server. Custom scripts (available for download as a zip file) were written to conduct the minimum p-value test, perform the Wilcoxon rank sum test for gene sets, and estimate the fraction of true associations as a function of variant p-value.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequence data and phenotypes for this study are available via the database of Genotypes and Phenotypes (dbGAP) and/or the European Genome-phenome Archive, as indicated in Supplementary Table 1.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We conducted a power analysis (described in the text) to demonstrate that the 45,231 samples analyzed in the study provided a significant increase in power to detect rare variant associations compared to previous studies. Power calculations were performed for frequency and effect size combinations in the range of those previously hypothesized to exist for complex diseases like T2D.
Data exclusions	At the time of sample selection for sequencing, samples were excluded if they matched predetermined criteria for T1D or MODY as described in Supplementary Table 1. At the analysis stage, excluded data were of three types. (a) Samples and (b) variants were excluded if they failed quality control analyses (described in Methods). (c) ~3600 cases from the PRODiGY study were excluded because they did not have suitably matched controls, resulting in inflated tests statistics as described in the Methods section. Exclusion criteria during the analysis stage were determined based on inspection of the distribution of data.
Replication	We replicated our significant associations in independent datasets from CHARGE and GHS, as described in the main text. All three exome-wide significant associations were replicated.
Randomization	Samples were allocated according to the cohort in which they were collected. Further control for confounding factors (imprecise ancestry matching even within cohort, technical confounders) were controlled for by including covariates in the regression model used for association analysis
Blinding	As our analysis involved a regression of phenotype on genotype, neither of which can be influenced by the analyst or data collector, blinding was not relevant to our study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Population characteristics are described in Supplementary Table 1. Relevant characteristics include age, sex, and BMI. For most cohorts, patients with T2D had higher BMI and age, except for cohorts from the GoT2D study where lean, young cases and old, obese controls were preferentially selected. Gender distribution varied by cohort.
Recruitment	Patients were recruited originally as a part of numerous cohort studies, described in Supplementary Table 1, each of which had different selection criteria. For most cohorts, patients were recruited over a long period of time and then cases and controls were selected for sequencing based on DNA and phenotyping quality. T2D diagnosis was determined by clinical data and not the participants themselves. Some bias may have occurred in terms of patient response to recruitment but these are unlikely to be correlated with genotype or have a significant effect on our analysis.
Ethics oversight	All samples were approved for use by their home institution's institutional review board or ethics committee, as previously reported (see references in the Methods section of the manuscript). Samples newly sequenced at The Broad Institute as part of T2D-GENES, SIGMA, and ProDiGY are covered under Partners Human Research Committee protocol # 2017P000445/PHS "Diabetes Genetics and Related Traits".

Note that full information on the approval of the study protocol must also be provided in the manuscript.