The genetic architecture of type 2 diabetes

There is compelling evidence that the individual risk of type 2 diabetes (T2D) is strongly influenced by genetic factors. Progress in characterizing the specific T2D-risk alleles responsible has been catalysed by the ability to perform genome-wide association studies (GWAS). Over the past decade, successive waves of T2D GWAS—featuring ever larger samples, progressively denser genotyping arrays supplemented by imputation against more complete reference panels, and richer ethnic diversity—have delivered more than 80 robust association signals. However, in these studies, the alleles interrogated for association were predominantly common (minor allele frequency (MAF) >5%), and with limited exceptions, the variants driving known association signals were also common, with individually modest impacts on T2D risk. Variation at known loci explains only a minority of observed T2D heritability.

Residual genetic variance is partly explained by a long tail of common variant signals of lesser effect. However, the contribution to T2D risk that is attributable to lower-frequency variants remains a matter of considerable debate, not least because of the relevance of disease risk that is attributable to lower-frequency variants remains a matter of considerable debate, not least because of the relevance of disease pathophysiology, but large-scale sequencing does not support the idea that lower-frequency variants have a major role in predisposition to type 2 diabetes.

The genetic architecture of common traits, including the number, frequency, and effect sizes of inherited variants that contribute to individual risk, has been long debated. Genome-wide association studies have identified scores of common variants associated with type 2 diabetes, but in aggregate, these explain only a fraction of theheritability of this disease. Here, to test the hypothesis that lower-frequency variants explain much of the remainder, the GoT2D and T2D-GENES consortia performed whole-genome sequencing in 2,657 European individuals with and without diabetes, and exome sequencing in 12,940 individuals from five ancestry groups. To increase statistical power, we expanded the sample size via genotyping and imputation in a further 111,548 subjects. Variants associated with type 2 diabetes after sequencing were overwhelmingly common and most fell within regions previously identified by genome-wide association studies. Comprehensive enumeration of sequence variation is necessary to identify functional alleles that provide important clues to disease pathophysiology, but large-scale sequencing does not support the idea that lower-frequency variants have a major role in predisposition to type 2 diabetes.

Analysis of genome-wide variation

The GoT2D consortium selected for whole-genome sequencing cases of type 2 diabetes (T2D) and ancestry-matched normoglycaemic controls from northern and central Europe (Methods and Supplementary Table 1). To increase power to identify low-frequency (0.5% < MAF < 5%) and rare (MAF < 0.5%) T2D variants with large effects, we preferentially identified individuals from the extremes of genetic risk (Methods). The genome sequence of 1,326 cases and 1,331 control individuals was determined through joint statistical analysis of low-coverage whole-genome sequence (~5×), deep-coverage exome sequence (~82×), and array-based genotypes at 2.5 million single nucleotide variants (SNVs) (Extended Data Fig. 1 and Extended Data Table 1).

We detected, genotyped, and estimated haplotype phase for 26.7 million genetic variants (Extended Data Fig. 1 and Extended Data Table 2), including 1.5 million short insertion-deletion variants (indels) and 8,876 large deletions. Individual diploid genomes carried a mean of 3.30 million variants (range: 3.20 million–3.35 million), including 271,245 indels (262,201–327,077), and 669 (579–747) large deletions. These data include many variants not directly studied by previous GWAS, including all of the indels as well as 420,473 common and 2.4 million low-frequency SNVs that were poorly tagged (r² ≤ 0.30) by genotype arrays. We estimate near-complete ascertainment (98.2%) of SNVs with minor allele counts of greater than 5 (MAF > 0.1%), and high accuracy (over 99.1%) at heterozygous genotypes (Methods and Fig. 1a). As half of the sequenced individuals were T2D cases, ascertainment was enhanced for any rare or low-frequency variants that substantially increase T2D risk (Fig. 1a).

Specifically, we estimate ≥80% power to detect (at genome-wide significance, α = 5 ×10⁻⁸) T2D risk variants with MAF ≥ 5% and odds ratio (OR) ≥ 1.87, or MAF ≥ 0.5% and OR ≥ 4.70 (Extended Data Fig. 2).

We tested all 26.7 million variants for T2D association by logistic regression assuming an additive genetic model (Supplementary Table 2). Analyses using a mixed-model framework to account for population structure and relatedness generated almost identical results. At genome-wide significance, 126 variants at four loci were associated with T2D (Fig. 1b). These included two previously reported common-variant loci (TCF7L2 and ADCY5), a previously reported low-frequency variant in CCND2 (ref. 7) (rs76895963, MAF = 2.6%, Pseq = 4.2 × 10⁻⁸), and a novel common-variant association near EML4 (MAF = 34.8%, Pseq = 1.0 × 10⁻⁸). There was no significant evidence of association with T2D for sets of low-frequency or rare variants within coding regions, nor within specified non-coding regulatory elements (Methods).

Power to detect association with low-frequency and rare variants of modest effect is limited in a sample of 2,657 individuals. To increase power for variants discovered via genome sequencing, we imputed sequence-based genotypes into 44,414 additional individuals of European origin (11,645 T2D cases and 32,769 controls; Methods) from 13 studies (Supplementary Table 3). We estimated power in the combined sequence plus imputed data, adjusting for imputation
quality, to be $\geq 80\%$ for variants with MAF $\geq 5\%$ and OR $\geq 1.23$, or MAF $\geq 0.5\%$ and OR $\geq 1.92$ (Extended Data Fig. 2). A meta-analysis combining results for the sequence and imputed data identified 674 variants across 14 loci associated with T2D at genome-wide significance (Fig. 1c). We observed a previously undescribed association with a common variant near CENPW ($r^2 = 1.759026$, MAF = 23.2\%, $P_{\text{meta}} = 3.5 \times 10^{-8}$; Fig. 1c) and replicated this association in an additional 14,201 cases and 100,964 controls from the DIAGRAM consortium ($P = 2.5 \times 10^{-8}$, $P_{\text{combined}} = 1.1 \times 10^{-11}$; Methods). The EML4 signal detected in the sequence data was not replicated in the imputed data ($P = 0.59$; $P_{\text{meta}} = 0.26$; Fig. 1c).

To test for additional association signals, we performed conditional analysis at loci previously associated with risk of T2D (Methods). We identified two previously unreported association signals, both involving low-frequency variants, at a corrected significance threshold ($\alpha < 1.8 \times 10^{-6}$; Methods): one at the IRS1 locus ($r^2 = 0.712464$, MAF = 2.2\%, $P_{\text{conditional}} = 2.5 \times 10^{-7}$) and one upstream of PPARG ($r^2 = 0.79856023$, MAF = 2.2\%, $P_{\text{conditional}} = 9.2 \times 10^{-10}$) (Extended Data Table 3). The PPARG signal overlaps regulatory elements in hASC pre-adipose and HepG2 cells, consistent with evidence that altered adipose regulation drives the primary PPARG signal.

**Analysis of coding variation**

The T2D-GENES consortium adopted a complementary strategy, focused on variants in protein-coding sequence, and seeking to improve power to detect rare-variant association by exploiting the more robust functional annotation of coding variation and the potential to aggregate multiple alleles of presumed similar impact in the same gene\(^{12,15}\). We combined exome sequence data from 10,437 T2D cases and controls of diverse ancestry generated by T2D-GENES with the equivalent data from GoT2D. This created a joint data set (after all quality control) comprising 12,940 individuals (6,504 cases and 6,436 controls) drawn from five ancestry groups: 4,541 of European origin, and around 2,000 (range: 1,943–2,217) each of South Asian, East Asian, Hispanic and African American origin (Extended Data Fig. 1, Extended Data Table 1 and Supplementary Table 4). Mean coverage was 82\% across the coding sequence of 18,281 genes, identifying 3.04 million variants (1.19 million protein-altering; Supplementary Figs 5, 6). Each diploid focused on variants in protein-coding sequence, and seeking to improve power to detect rare-variant association by exploiting the more robust functional annotation of coding variation and the potential to aggregate multiple alleles of presumed similar impact in the same gene\(^{12,15}\). We combined exome sequence data from 10,437 T2D cases and controls of diverse ancestry generated by T2D-GENES with the equivalent data from GoT2D. This created a joint data set (after all quality control) comprising 12,940 individuals (6,504 cases and 6,436 controls) drawn from five ancestry groups: 4,541 of European origin, and around 2,000 (range: 1,943–2,217) each of South Asian, East Asian, Hispanic and African American origin (Extended Data Fig. 1, Extended Data Table 1 and Supplementary Table 4). Mean coverage was 82\% across the coding sequence of 18,281 genes, identifying 3.04 million variants (1.19 million protein-altering; Supplementary Figs 5, 6). Each diploid genome carried a mean of 9,243 (range: 8,423–11,487) synonymous, 7,636 (6,935–9,271) missense, and 250 (183–358) protein-truncating alleles (Supplementary Table 7).

We tested for T2D association within the five ancestry groups, assuming an additive genetic model, using mixed-model approaches that account for population structure and relatedness\(^{16,18}\) and combined ancestry-specific results by trans-ethnic meta-analysis (Methods). We estimate $\geq 80\%$ power to detect (at genome-wide significance) T2D risk variants with MAF $\geq 5\%$ and OR $\geq 1.36$, or MAF $\geq 0.5\%$ and OR $\geq 2.29$ (Methods and Extended Data Fig. 2). Only one variant reached genome-wide significance ($\text{PAX4 Arg192His}$, rs2233580, $P = 9.3 \times 10^{-8}$) (Table 1, Extended Data Figs 3, 4 and Supplementary Fig. 8). This association was exclusive to East Asian individuals, in whom the 192His allele is common (MAF $\approx 10\%$) with a substantial effect size (allelic OR = 1.79 (1.47–2.19)); 192His is virtually absent in individuals from other ancestries (MAF = 0.014\%). The rs2233580 association was replicated in independent East Asian case–control data ($n = 3,301; P = 5.9 \times 10^{-7}$; Supplementary Table 9) and was distinct ($r^2 < 0.05$) from previously reported GWAS SNVs at the GCC1–PAX4 locus\(^{6,8}\). $\text{PAX4}$ encodes a transcription factor involved in islet differentiation and function\(^{17}\) (Supplementary Table 10), and $\text{PAX4}$ variants have been implicated in early-onset monogenic diabetes\(^{18}\). However, in East Asian cases, 192His was not associated with age of diabetes diagnosis ($P = 0.64$), indicating that this variant influences risk of type 2 diabetes rather than early-onset monogenic diabetes (Supplementary Table 9).

To increase power to detect the association of rare variants that cluster in individual genes, we deployed gene-level variant aggregation tests\(^{19}\) across the exome sequence data (Methods and Supplementary Table 11). We observed no deviation from the null distribution of association statistics, and no single gene reached exome-wide significance ($\text{PAX4} \times 10^{-6}$; Methods and Supplementary Figs 12, 13). When we focused on 634 genes that mapped to known GWAS regions, only $\text{FES}$ exceeded a reduced significance threshold of $\alpha = 2.5 \times 10^{-6}; \text{Methods and Supplementary Fig. 14}$). This aggregate signal was driven entirely by the South Asian-specific Pro536Ser variant (MAF = 0.9\%, OR = 6.7 (2.6–17.3), $P = 7.5 \times 10^{-6}$), indicating that $\text{FES}$ is likely to be the effector gene at the PRC1 GWAS locus\(^{4}\).

To increase power to detect coding variant associations (Extended Data Fig. 2), we contributed early T2D-GENES exome data to the design of an Illumina exome array\(^{25}\), and then collected genotypes from an additional 28,305 T2D cases and 51,549 controls of European ancestry from 13 studies (Extended Data Fig. 1, Extended Data Table 1 and Supplementary Table 15). Of 27,904 protein-variants with MAF $\geq 0.5\%$ detected in exome sequence data from 4,541 European individuals, variation at 81.6\% was captured on the array (Supplementary Fig. 16).

Association analysis in the combined sequence and array data from more than 90,000 individuals identified 18 coding variants (17 non-synonymous) at 13 loci that exceeded genome-wide significance ($\alpha = 5 \times 10^{-8}$; Table 1 and Extended Data Figs 3, 4). All of these were common (MAF $> 5\%$) and all but one mapped within established common-variant GWAS regions\(^{23}\). The exception, which we replicated in the INTERACT study\(^{19}\) ($n = 9,292; P_{\text{INTERACT}} = 2.4 \times 10^{-4}$, $P_{\text{meta}} = 2.2 \times 10^{-11}$), involved a common haplotype of four strongly
These loci were identified through single-variant analyses of exome sequence data from 6,504 cases and 6,436 controls and exome arrays from 28,305 cases and 51,549 controls. RAF, risk allele genome-wide significance \((P < 5 \times 10^{-8})\) in the combined analysis. Three variants at ASCC2 did not achieve genome-wide significance themselves, but are included because they fall into a region of established common variant GWAS regions, GCKR and PPARG. For five loci \((SLC30A8, GCKR, PPARG, KCNJ11--ABCC8, and PAM)\), the coding variants identified had previously been nominated as causal for their respective GWAS signals \([2,7,13]\). For the other seven loci, GWAS meta-analyses had previously highlighted a lead variant in non-coding sequence \([2,5,6]\). We re-evaluated these relationships with conditional and credible set analyses and found that, at most, the evidence supported a direct causal role for the coding variants concerned (Extended Data Table 3, Extended Data Figs 3, 4, Supplementary Table 10 and Supplementary Fig. 17). For example, at the CILP2 locus\(^2\), previous GWAS had identified the non-coding variant rs10401969 as the lead SNV. However, direct genotyping of TM6SF2 Glu167Lys on the exome array revealed complete linkage disequilibrium with rs10401969, and reciprocal signal extinction in conditional analyses (Extended Data Table 3 and Extended Data Figs 3, 4). In previous GWAS, the association at Glu167Lys had been obscured by incomplete genotyping and poor imputation.

**Table 1** Nonsynonymous coding variants achieving genome-wide significance

| Locus       | Gene     | Variant                              | Exomes \((n = 12,940)\) | Exome-chip \((n = 79,854)\) | Combined \((n = 92,794)\) |
|-------------|----------|--------------------------------------|--------------------------|---------------------------|--------------------------|
|             |          |                                      | OR (95% CI)              | OR (95% CI)               | OR (95% CI)               |
| Established common causal coding variant signals |          |                                      | **P**                    | **P**                     | **P**                     |
| GCKR        | GCKR     | rs1260326 Pro446Leu                   | 0.075                    | 4.8 \times 10^{-9}        | 1.07 \times 10^{-9}       |
| PPARG       | PPARG    | rs1801282 Pro124Ala                   | 0.0030                   | 1.16 \times 10^{-7}       | 1.10 \times 10^{-7}       |
| PAM/PPIPS2K | PAM      | rs35658696 Asp563Gly                   | 0.00045                  | 3.16 \times 10^{-7}       | 1.15 \times 10^{-7}       |
|             | PPIPS2K  | rs36046591 Ser1207Gly                 | 0.00099                  | 1.34 \times 10^{-6}       | 1.17 \times 10^{-6}       |
| SLC30A8     | SLC30A8  | rs13266634 Arg325Trp                  | 0.11                     | 3.4 \times 10^{-9}        | 1.07 \times 10^{-9}       |
| KCNJ11/ABCC8| KCNJ1    | rs5215 Val337Ile                     | 0.06                     | 5.1 \times 10^{-9}        | 1.07 \times 10^{-9}       |
| ABCC8       | ABCC8    | rs757110 Ala1369Ser                   | 0.20                     | 2.3 \times 10^{-8}        | 1.07 \times 10^{-8}       |

Other coding variant associations within established common variant GWAS regions

| Locus | Gene     | Variant                              | Exomes \((n = 12,940)\) | Exome-chip \((n = 79,854)\) | Combined \((n = 92,794)\) |
|-------|----------|--------------------------------------|--------------------------|---------------------------|--------------------------|
|       |          |                                      | **P**                    | **P**                     | **P**                     |
| THADA | THADA    | rs35720761 Cys1605Tyr                 | 0.11                     | 3.5 \times 10^{-8}        | 1.11 \times 10^{-6}       |
| COBLL1| COBLL1   | rs7607980 Asn939Asp                   | 1.21                     | 4.7 \times 10^{-11}       | 1.14 \times 10^{-11}      |
| WFS1  | WFS1     | rs18012121 Val333Ile                 | 0.14                     | 9.3 \times 10^{-12}       | 1.08 \times 10^{-12}      |
|       |          |                                      | 0.0026                   | 2.0 \times 10^{-12}       | 1.08 \times 10^{-12}      |
|       |          |                                      | 0.0019                   | 1.08 \times 10^{-11}      | 1.05 \times 10^{-11}      |
| RREB1 | RREB1    | rs9379084 Asp1171Asn                  | 0.10                     | 1.3 \times 10^{-10}       | 1.07 \times 10^{-10}      |
|       |          |                                      | 0.08                     | 1.3 \times 10^{-11}       | 1.03 \times 10^{-11}      |
| PAX4  | PAX4     | rs2233580 Arg192His                   | 0.17                     | 1.10^{-10}                | 1.12 \times 10^{-10}      |
|       |          |                                      | 0.00                      | 1.09 \times 10^{-9}       | 1.06 \times 10^{-11}      |
| GPM51 | GPM51*   | rs60980157 Ser391Leu                  | 0.17                     | 9.3 \times 10^{-9}        | 1.09 \times 10^{-12}      |
|       |          |                                      | 0.26                     | 1.09 \times 10^{-9}       | 1.06 \times 10^{-12}      |
| CILP2 | TM6SF2   | rs58542926 Glu167Lys                  | 1.22                     | 1.9 \times 10^{-7}        | 1.13 \times 10^{-10}      |
|       |          |                                      | (1.10^{-13})             | (1.08^{-18})              | (1.14^{-10})              |

Coding variant associations outside established common variant GWAS regions

| Locus       | Gene     | Variant                              | Exomes \((n = 12,940)\) | Exome-chip \((n = 79,854)\) | Combined \((n = 92,794)\) |
|-------------|----------|--------------------------------------|--------------------------|---------------------------|--------------------------|
| MTRM3/b     | MTRM3   | rs41278853 Asn960Ser                  | 0.26                     | 9.2 \times 10^{-5}        | 1.12 \times 10^{-5}       |
| ASCC2       | ASCC2   | rs11549795 Val123Ile                  | 0.23                     | 3.2 \times 10^{-6}        | 1.12 \times 10^{-6}       |
|             |          |                                      | 0.00040                  | 2.0 \times 10^{-5}        | 1.07 \times 10^{-7}       |
|             |          |                                      | 0.00050                  | 1.9 \times 10^{-5}        | 1.06 \times 10^{-7}       |
|             |          |                                      | 0.0023                   | 1.23 \times 10^{-5}       | 1.11 \times 10^{-7}       |

These loci were identified through single-variant analyses of exome sequence data from 6,504 cases and 6,436 controls and exome arrays from 28,305 cases and 51,549 controls. RAF, risk allele frequency; Eur MAF, minor allele frequency in Europeans; OR, odds ratio; CI, confidence interval; n, total number of individuals analysed. Genome-wide significance defined as \(P < 5 \times 10^{-8}\). GPM51 variant failed quality control in exome sequence: association \(P\) values derive only from exome-array analysis. The synonymous variant Thr515Thr \((rs55834942)\) in GCKR values derive only from exome-array analysis. The synonymous variant Thr515Thr \((rs55834942)\) in GCKR values derive only from exome-array analysis.
that have been more directly implicated in monogenic and syndromic forms of diabetes. We extended gene-based tests for rare-variant associations to gene-sets highlighted these two genes as probable effectors in T2D risk. TSPAN8 was implicated in monogenic diabetes. Additional coding variant associations in these genes, alleles of penetrance sufficient to drive familial segregation of early-onset diabetes coexist alongside those with more modest effects predisposing carriers to later-onset T2D. No other compelling signals of early-onset diabetes were detected using gene-set enrichment or protein–protein interaction analysis in other pre-defined gene-sets.

**No evidence for synthetic association**

In 2010, Goldstein and colleagues proposed that common-variant GWAS signals may be the consequence of low-frequency and rare variants that by chance cluster on common haplotypes. Although this hypothesis has been debated and assessed indirectly, we used the near-complete ascertainment of genetic variation in 2,657 genome-sequenced individuals to directly test the importance of ‘synthetic’ associations. We focused on the ten T2D GWAS loci at which our sample provided the strongest statistical evidence for association (P < 0.001), implementing a conditional analysis procedure to assess whether combinations of SNVs within a 5-Mb window could explain the common-variant signal (Extended Data Table 4 and Methods).

We first focused on missense variants, finding that none of the ten signals could be explained by low-frequency and rare variants within 2.5 Mb of the common index SNV (Extended Data Fig. 5). For example,
at the IRS1 locus, including the five observed missense IRS1 alleles in the model did not meaningfully diminish the index SNP association ($P_{\text{unconditional}} = 2.8 \times 10^{-6}$; $P_{\text{conditional}} = 4.3 \times 10^{-6}$). With 99.7% ascertainment of low-frequency coding variants (Methods), these results rule out synthetic associations produced by missense variants at these ten loci.

We expanded the search to include all low-frequency and rare variants, non-coding and coding, within 2.5 Mb of index SNVs. At no locus was a single low-frequency or rare variant sufficient to explain the GWAS signal (Extended Data Fig. 5). At eight of the ten loci, ten or more low-frequency and rare variants were needed to reverse the direction of effect at the common index SNP; at TCF7L2, even 50 were insufficient (Extended Data Fig. 5). We note that the statistical procedure we developed and deployed is biased in favour of the synthetic association hypothesis, since it is highly prone to over-fitting. Nonetheless, at 8 of the 10 loci the data were indistinguishable from a null model of no synthetic association (Extended Data Table 4 and Supplementary Fig. 27).

Nominating candidate functional alleles

Using the GoT2D whole-genome sequence data, we constructed 99% ‘credible sets’ for each T2D GWAS locus on the assumption that there is one causal variant per locus30 (Methods). Across 78 published autosomal loci at which the reported index SNP had MAF > 1%, 99% credible set sizes ranged from 2 (CDK2A/B) to around 1,000 (POUSF1) variants; at 71 loci, the credible set contained more than 10 variants (Extended Data Fig. 5 and Supplementary Table 28). The GoT2D data set provides near-complete ascertainment of common and low-frequency variants, which support more comprehensive credible set analysis than studies based on genotyping or imputation alone31; of the credible set variants identified from whole-genome sequence data, about 60% are absent from HapMap and ~5% from 1000G Phase 1 (Extended Data Fig. 5).

Genomic maps of chromatin state or transcription factor binding32,35 have been used to prioritize causal variants within credible sets36,37. We jointly modelled genetic association and genomic annotation data at T2D GWAS loci using lGwas38. Consistent with previous reports34,35, associated variants were enriched in coding exons, transcription factor binding sites, and enhancers active in pancreatic islets and adipose tissue (Extended Data Fig. 6). Overall, including the functional annotation data reduced the credible set size by 35%. At several loci, access to complete sequence data prioritized variants that overlapped with relevant regulatory annotations and had been previously overlooked. For example, at the CCND2 locus, three variants not present in HapMap Phase 2 have combined a probability of 90.0% of explaining the common-variant signal3 (Extended Data Fig. 6); one of these (rs3217801) is a 2-base pair (bp) indel overlapping an islet enhancer element.

Modelling disease architecture

To evaluate the overall contribution of low-frequency coding variation to T2D risk, we estimated the proportion of variance in T2D liability attributable to each such variant39 (Methods and Extended Data Fig. 7). We focused on exome array data to maximize sample size, and on variants with MAF > 0.1%; the sensitivity of variant ascertainment and accuracy of OR estimation decline below this threshold. Among the 31,701 variants on the exome array with 0.1% < MAF < 5%, there was a progressive increase in the maximum OR estimates with decreasing frequency. However, the liability variance explained for these variants rarely exceeded 0.05%, limiting the power to detect association in the sample size available (Extended Data Fig. 7). We estimated (Methods) that the liability variance that was collectively attributable to coding variants in the 0.1% < MAF < 5% range was 2.9%, compared to 6.3% for common variants.

Finally, we compared our whole-genome T2D association results with predictions from population genetic simulations40 under twelve models that varied widely with respect to the proportion of heritability explained by common, low-frequency, and rare variants. We mirrored the GoT2D study design (with imputation) and performed in parallel the same association analysis on empirical and simulated data, focusing on variants with MAF > 0.1% and allowing for power loss due to imperfect imputation (Methods).

Figure 3 displays results for three representative models: a ‘purifying selection’ model in which low-frequency and rare variants explain approximately 75% of T2D heritability; an intermediate model in which both common and lower-frequency variants contribute substantially; and a ‘neutral’ model in which common variants explain about 75% of T2D heritability. The predictions of the first two models differ markedly from the empirical data with respect to the numbers of low-frequency and rare risk variants that are associated with T2D. Specifically, these two models predict a larger number and greater effect size of low-frequency variants should be found in our whole-genome sequencing study as compared to those observed in the empirical data. By contrast, the empirical data are consistent with predictions under the ‘neutral’ common-variant model.

The century-old Mendelian–biometrician debate pitted those who attributed trait variation to rare variants of large effect against those who argued that trait variation was largely due to many common variants of small effect. The debate today is about whether the ‘missing heritability’ after GWAS is due largely to individually rare, highly penetrant variants41 or to a large universe of common alleles with modest effects42. The results are of more than academic interest, as genetic architecture plays out powerfully in relation to the power of genetic diagnosis and the application of precision medicine.

Figure 3 | Empirical T2D association results compared to results under different simulated disease models. Observed number of rare and low-frequency (MAF < 5%) genetic association signals for T2D detected genome-wide after imputation compared to the numbers seen under three simulated disease models for T2D which were plausible given results (T2D recurrence risks, GWAS, linkage) before large-scale sequencing.
Our data and analysis indicate that, for T2D, nearly all common-variant associations detectable by whole genome sequencing were previously found by GWAS based on genotyping arrays and imputation; concerns about incomplete coverage due to ‘holes’ in HamMap coverage were, we show, unfounded. Of more lasting interest, the combination of genome and exome sequencing in large samples provides limited evidence of a role for lower-frequency variants—coding or genome-wide—in T2D predisposition. Of course, rare risk alleles have long been known to contribute in families with early-onset forms of diabetes, and sequencing of Mendelian and GWAS genes has identified rare variants that influence disease risk. Sequencing of T2D cases in much larger samples will undoubtedly uncover additional low-frequency and rare variants that provide biological and potentially clinical value. Nonetheless, our empirical and simulated data suggest that these lower-frequency variants contribute much less to T2D heritability than do common variants. Moreover, the frequency spectrum of variant association signals is consistent with a model in which limited selective pressure distributes most of the genetic variance influencing T2D risk among common alleles, in line with the frequency distribution of inter-individual sequence variation. Similar large-scale sequencing-based exploration of other complex traits will be required to determine the extent to which the genetic architecture of T2D is representative of other late-onset diseases.

Our results further strengthen the case for sequencing of diverse samples; the population-enriched T2D risk variant in PAX4 dovetails with similar findings involving SLC16A11 (ref. 45) in East Asian and Native American populations and TBC1D4 (ref. 46) in Greenland Inuits. Studies involving populations that have been subject to bottlenecks or extreme selective pressures may be particularly fruitful.

Understanding the inherited basis of T2D will require much further progress in identifying the mechanisms whereby common, mostly non-coding, variants influence disease risk. The combination of global progress in identifying the mechanisms whereby common, mostly non-coding, variants influence disease risk. The combination of global progress in identifying the mechanisms whereby common, mostly non-coding variants influence disease risk. The combination of global progress in identifying the mechanisms whereby common, mostly non-coding variants influence disease risk. The combination of global progress in identifying the mechanisms whereby common, mostly non-coding variants influence disease risk. The combination of global progress in identifying the mechanisms whereby common, mostly non-coding variants influence disease risk. The combination of global progress in identifying the mechanisms whereby common, mostly non-coding variants influence disease risk. The combination of global progress in identifying the mechanisms whereby common, mostly non-coding variants influence disease risk. The combination of global progress in identifying the mechanisms whereby common, mostly non-coding variants influence disease risk. The combination of global progress in identifying the mechanisms whereby common, mostly non-coding variants influence disease risk.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 May 2015; accepted 12 June 2016.

Published online 11 July 2016.

1. Willemsen, G. et al. The concordance and heritability of type 2 diabetes in 34,166 twin pairs from international twin registers: the discordant twin (DISCOTWIN) consortium. Twin Res. Hum. Genet. 18, 762–771 (2015).  
2. Morris, A. P. et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. 44, 981–990 (2012).  
3. Mahajan, A. et al. Genome-wide trans-ancestry meta-analysis provides insight into human population variation, physiology, and disease. Integration of biological insights gleaned from common and rare variant associations with T2D into a unified picture of disease pathophysiology will be required to fully understand the basis of this common but challenging disease.
41. McClellan, J. & King, M. C. Genetic heterogeneity in human disease. Cell **141**, 210–217 (2010).
42. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. **42**, 565–569 (2010).
43. Flannick, J. et al. Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. Nat. Genet. **46**, 357–363 (2014).
44. Bonnefond, A. et al. Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. Nat. Genet. **44**, 297–301 (2012).
45. Sigma Type 2 Diabetes Consortium et al. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. Nature **506**, 97–101 (2014).
46. Moltke, I. et al. A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. Nature **512**, 190–193 (2014).
47. Sigma Type 2 Diabetes Consortium et al. Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. JAMA **311**, 2305–2314 (2014).
48. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. Science **343**, 80–84 (2014).
49. Majithia, A. R. et al. Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. Proc. Natl Acad. Sci. USA **111**, 13127–13132 (2014).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** Grant support and acknowledgments are listed in the Supplementary Information.

**Author Contributions** Author contributions are described in the Supplementary Information.

**Author Information** Whole-genome sequence data from the GoT2D project are available by application to the European Genome-Phenome Archive (https://www.ebi.ac.uk/ega/home) under accession number EGAS00001001459 and from dbGAP (http://www.ncbi.nlm.nih.gov/gap) under accession number phs000840.v1.p1. Whole-exome sequence data from the T2D-GENES project are available from the European Genome-Phenome Archive (https://www.ebi.ac.uk/ega/home) under accession number EGAS00001001460 and from dbGAP (http://www.ncbi.nlm.nih.gov/gap) under accession numbers phs000847.v1.p1, phs001095.v1.p1, phs001096.v1.p1, phs001097.v1.p1, phs001098.v1.p1, phs001099.v1.p1, phs001100.v1.p1 and phs001102.v1.p1. Summary-level data from the exome array component of this project (and from the exome and genome sequences) can be freely accessed at the Accelerating Medicines Partnership T2D portal (http://www.type2diabetesgenetics.org), and similar data from the GoT2D-imputed data at http://www.diagram-consortium.org. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.I.M. (mark.mccarthy@drl.ox.ac.uk) or M.B. (boehnke@umich.edu).
METHODS

Ethics statement. All human research was approved by the relevant institutional review boards and conducted according to the Declaration of Helsinki. All participants provided written informed consent.

Data generation

GoT2D integrated panel generation

GoT2D sequenced samples. Here we describe how we generated, processed, and carried out quality control (QC) on sequence and genotype data for the 2,891 individuals included in the GoT2D from four studies, and how this resulted in 2,657 individuals (1,326 T2D cases and 1,331 non-diabetic controls) for analysis (Extended Data Fig. 1). We preferentially sampled early-onset, lean, and/or familial T2D cases and overweight controls with low fasting glucose levels60. Specific details of selected samples are provided in Extended Data Table 1 and Supplementary Table 1.

DNA sample preparation. De-identified DNA samples were sent to the Broad Institute (DGI, FUSION), Welcome Trust Centre for Human Genetics in Oxford (UKT2D), and Helmholtz Zentrum München (KORA) and prepared for genetic analysis. DNA quantity was measured by Picogreen (all), and samples with sufficient total DNA and minimum concentrations for downstream experiments were genotyped for a set of 24 SNVs using the Sequenom iPLEX assay (DGI, FUSION, UKT2D); one gender assay and 23 SNVs located across the autosomes. The genotypes for these SNVs were used as a quality filter to advance samples and a technical fingerprint for subsequent sequencing and genome-wide array genotypes.

Exome sequencing. Genomic DNA was sheared, end repaired, ligated with barcoded Illumina sequencing adapters, amplified, size selected, and subjected to in-solution hybrid capture using the Agilent SureSelect Human All Exon 44Mb v2.0 (DGI, FUSION, UKT2D) and v3.0 (KORA) bait set (Agilent Technologies, USA). Resulting Illumina exome sequencing libraries were QPCR quantified, pooled, and sequenced with 76-bp paired-end reads using Illumina GAII or HiSeq 2000 sequencers. The resulting BAM and VCF files were sub-mitmed to NCBI and are available in dbGaP (accession number phs000840.v1.p1).

Genome sequencing. Whole-genome Illumina sequencing library construction was performed as described for exome capture above, except that genomic DNA was sheared to a larger target size and hybrid capture was not performed. The resulting libraries were size selected to contain fragment insert size of 380 bp ± 20% (DGI, FUSION, KORA) and 420 bp ± 25% (UKT2D) using gel electrophoresis or the SAGE Pippin Prep (Sage Science, USA). Libraries were QPCR quantified, pooled, and sequenced with 101-bp paired-end reads using Illumina GAII or HiSeq 2000 sequencers to ~5-fold mean coverage.

Genotype likelihood calculation. To calculate genotype likelihoods across each of the 2,874 individuals, we evaluated accuracy for the low-coverage data using the strategy developed for 1000G Phase 1 (ref. 55). Specifically, we phased the integrated likelihoods using Beagle with 10,000 SNVs per chunk and 1,000 overlapping SNVs between consecutive chunks. We refined phased genotypes using GotCloud as implemented in GotCloud (http://genome.sph.umich.edu/wiki/GotCloud) with 400 states to improve genotype and haplotype quality.

GoT2D integrated panel haplotype integration

Genotype likelihood calculation. We merged SNVs discovered from the three experimental platforms into one site list and calculated genotype likelihoods across all sites separately by platform. Because exome sequence data have substantial off-target coverage, we calculated likelihoods across the genome combining data from the genome and exome sequence experiments. For genome sequence, we calculated likelihoods using GotCloud; for exomes, we used GATK UnifiedGenotyper; for HumanOmni2.5 genotypes, we converted hard genotype calls into genotype likelyhods assuming a genotype error rate of 10⁻⁶. For indels, we calculated likelihoods in a similar way except that the HumanOmni2.5 data could not be used. For structural variants (SVs), genotype likelihoods were calculated from GenomESTRIP using the whole-genome sequence data. Integration of genotype and sequence data. We calculated combined genotype likelihoods across each of the 2,874 individuals as the product of the corresponding genome, exome, and HumanOmni2.5 likelihoods assuming independent data across platforms (Extended Data Fig. 1). We then phased the genotype data using the strategy developed for 1000G Phase 1 (ref. 55). Specifically, we phased the integrated likelihoods using Beagle with 10,000 SNVs per chunk and 1,000 overlapping SNVs between consecutive chunks. We refined phased sequences using Thunder as implemented in GotCloud (http://genome.sph.umich.edu/wiki/GotCloud) with 400 states to improve genotype and haplotype quality.

GoT2D integrated panel QC.

GoT2D integrated panel QC. 2,874 individuals were available in the integrated haplotype panel. To identify population outliers, we carried out principal components analysis (PCA). We computed PCs for each of the three variant types (SNVs, short indels, large deletions) using PCAxTS on an LD-pruned (r² < 0.20) set of variants with MAF > 0.01. Variants with PC1 score > 2.5 were considered outliers and were removed from further analysis. We performed quality control using the GATK UnifiedGenotyper to call short indels (<50bp). Because short indels are known to have high false positive rates due to systematic sequencing and alignment errors, we performed stringent filtering criteria in SVM and VQSR and removed indels that failed either algorithm. For exome sequencing, we used GATK UnifiedGenotyper to call short indels, following best practices described elsewhere.

Large deletion identification. We used GenomeSTRIP to call large (>100-bp) deletions in the whole-genome sequence data. After initial discovery of large deletions in 2,764 QC-passed individuals, we merged the discovered sites with deletions identified in 1,092 sequenced individuals from the 1000G Project to increase sensitivity and then genotyped the merged site lists across the 2,764 individuals. After applying the default filtering implemented in GenomeSTRIP, pass-filtered sites variable in any of the samples were identified as candidate variant sites. Among these candidate sites, we excluded variants in known immunoglobulin loci to reduce the impact of possible cell-line artefacts. We then excluded 136 more individuals owing to an unusually large number of variants per sample (> median + 3 × mean absolute deviation). Variants present only in these excluded individuals were removed from further analysis.

GoT2D integrated panel haplotype integration

Genotype likelihood calculation. We merged SNVs discovered from the three experimental platforms into one site list and calculated genotype likelihoods across all sites separately by platform. Because exome sequence data have substantial off-target coverage, we calculated likelihoods across the genome combining data from the genome and exome sequence experiments. For genome sequence, we calculated likelihoods using GotCloud; for exomes, we used GATK UnifiedGenotyper; for HumanOmni2.5 genotypes, we converted hard genotype calls into genotype likelihoods assuming a genotype error rate of 10⁻⁶. For indels, we calculated likelihoods in a similar way except that the HumanOmni2.5 data could not be used. For structural variants (SVs), genotype likelihoods were calculated from GenomESTRIP using the whole-genome sequence data. Integration of genotype and sequence data. We calculated combined genotype likelihoods across each of the 2,874 individuals as the product of the corresponding genome, exome, and HumanOmni2.5 likelihoods assuming independent data across platforms (Extended Data Fig. 1). We then phased the genotype data using the strategy developed for 1000G Phase 1 (ref. 55). Specifically, we phased the integrated likelihoods using Beagle with 10,000 SNVs per chunk and 1,000 overlapping SNVs between consecutive chunks. We refined phased sequences using Thunder as implemented in GotCloud (http://genome.sph.umich.edu/wiki/GotCloud) with 400 states to improve genotype and haplotype quality.

GoT2D integrated panel QC. 2,874 individuals were available in the integrated haplotype panel. To identify population outliers, we carried out principal components analysis (PCA). We computed PCs for each of the three variant types (SNVs, short indels, large deletions) using PCAxTS on an LD-pruned (r² < 0.20) set of variants with MAF > 0.01. Variants with PC1 score > 2.5 were considered outliers and were removed from further analysis. We performed quality control using the GATK UnifiedGenotyper to call short indels (<50bp). Because short indels are known to have high false positive rates due to systematic sequencing and alignment errors, we performed stringent filtering criteria in SVM and VQSR and removed indels that failed either algorithm. For exome sequencing, we used GATK UnifiedGenotyper to call short indels, following best practices described elsewhere.

Large deletion identification. We used GenomeSTRIP to call large (>100-bp) deletions in the whole-genome sequence data. After initial discovery of large deletions in 2,764 QC-passed individuals, we merged the discovered sites with deletions identified in 1,092 sequenced individuals from the 1000G Project to increase sensitivity and then genotyped the merged site lists across the 2,764 individuals. After applying the default filtering implemented in GenomeSTRIP, pass-filtered sites variable in any of the samples were identified as candidate variant sites. Among these candidate sites, we excluded variants in known immunoglobulin loci to reduce the impact of possible cell-line artefacts. We then excluded 136 more individuals owing to an unusually large number of variants per sample (> median + 3 × mean absolute deviation). Variants present only in these excluded individuals were removed from further analysis.
We estimated the sensitivity of low-pass genome sequence data to detect true SNVs by calculating the proportion of exome-sequencing-detected SNVs detected by low-pass genome sequencing in the 2,538 individuals with data for all three experimental platforms. For exome sequence allele counts < 1,000, we merged adjacent allele count bins until the number of alleles was > 1,000. We estimated the sensitivity of low-pass genome sequencing to detect common, low-frequency, and rare SNVs as 99.8%, 99.0%, and 48.2%, respectively. Similarly, we estimated the sensitivity of low-pass genome sequence to detect true short indels by calculating the proportion in all downstream-sequencing-detected short indels detected by low-pass genome sequencing. Sensitivity estimates were > 99.9%, 93.8%, and 17.9% for common, low-frequency, and rare short indels, respectively.

To estimate the sensitivity of the combined low-pass genome and exome sequence data, we focused on coding SNVs and calculated the proportion of HumanOmni2.5 SNVs detected by either sequencing platform. Because HumanOmni2.5 SNVs are enriched for common variants, we calculated a weighted average sensitivity at each allele count, weighted by the number of exome-detected variants given the allele count. Sensitivity estimates were 99.9%, 99.7%, and 83.9% for common, low-frequency, and rare variants.

GoT2D integrated panel evaluation of genotype accuracy. To evaluate genotype accuracy for SNVs, we focused on chromosome 20, and compared the concordance of low-pass whole-genome-sequence-based genotypes with those based on exome sequence. Overall genotype concordance was 99.86%. Homozygous reference, heterozygous, and homozygous non-reference concordances were 99.97%, 98.34%, and 99.72%, respectively. We also compared genotype concordance between exome sequence and HumanOmni2.5 genotypes. Overall concordance was 99.4%. When the HumanOmni2.5 genotypes were homozygous reference, heterozygous, and homozygous non-reference, concordances were 99.97%, 99.69%, and 99.88%, respectively.

We evaluated the genotype accuracy of indels for the 210 chromosome-20 indels that overlapped between those discovered by exome and genome sequencing. Overall genotype concordance was 99.4%. When the exome genotypes were homozygous reference, heterozygous, and homozygous non-reference, concordances were 99.8%, 95.8%, and 98.6%, respectively.

To evaluate the genotype accuracy of our low-pass genome sequence data to detect true structural variants, we took advantage of the 181 individuals in our study who were previously included in the WTCCC array-CGH based structural variant detection experiment. Taking the WTCCC data as a gold standard, we estimated genotype accuracy across 1,047 overlapping structural variants (with reciprocal overlap > 0.8) genome-wide. The overall genotype concordance was 99.8%. When the WTCCC genotypes were homozygous reference, heterozygous, and homozygous non-reference, concordances were 99.9%, 96.6%, and 99.7%, respectively.

GoT2D + T2D-GENES multiethnic exome panel generation and QC

Samples. We considered 6,504 T2D cases and 6,436 controls from 14 studies of African American, East Asian, South Asian, Hispanic, and European ancestry. In contrast to the GoT2D whole-genome integrated panel, this data set also includes GoT2D individuals for whom whole-genome data were not available. Sample characteristics are provided in Extended Data Table 1 and Supplementary Table 4.

Sequence reads were processed and aligned to the reference genome (hg19) with Picard (http://broadinstitute.github.io/picard/). Polymorphic sites and genotypes were called with GATK, with filtering of sites performed using Variant Quality Score Recalibration (VSQR) for SNVs, and hard filters for indels. Genotype likelihoods were computed controlling for contamination.

Hard calls (the GATK-called genotypes but set as missing at a genotype quality (GG) < 20 threshold) and dosages (the expected value of the genotype, defined as Pr(RX|data) + Pr(X|data), where X is the alternative allele) were computed for each sample at each variant site. Hard calls were used only for quality control, while dosages were used downstream for QC and filtering.

GoT2D integrated panel association analysis

Single-variant association analysis. We tested for single-variant association in a logistic regression framework assuming an additive genetic model. We used the Firth bias-corrected likelihood ratio test, as our primary analysis strategy; we repeated association analysis using the score test for inclusion in sample-size-weighted meta-analysis (Supplementary Table 2). Tests were adjusted for sex, the first two genotype-based PCs to account for population stratification, and an indicator function for observed temporal stratification based on sequencing date and centre. PCs were calculated using linkage-disequilibrium (LD) pruned (r2 < 0.20) HumanOmni2.5M array variants with MAF > 1% after removing large-LD region96,97.

Aggregate association analysis. To test for aggregate association within coding regions of the genome, we used the approach described in Gene-based analysis below. For every gene and mask tested, P values were greater than 2.5 × 10−6.

We also tested for aggregate association among variants in non-coding regions of the genome. We aggregated variants in individual pancreatic islet enhancer elements (see Genomic annotation below), as these elements collectively demonstrated strongest genome-wide enrichment of T2D association. We performed both the burden and SKAT tests using genotypes from the integrated panel on variants with MAF < 5% in each islet enhancer element. We used a Bonferroni threshold P < 1.68 × 10−7 based on a nominal significance level of α = 0.05 corrected for 298,240 elements with at least one variant. All elements tested in this manner had P values greater than 2.5 × 10−6.

GoT2D + T2D-GENES multiethnic association analysis

Kinship analysis. Within each ancestry group, we considered autosomal variants in exonic and non-exonic regions. We calculated IBS between each pair of samples in the ancestry group based on independent variants (ethnic-specific r2 < 0.05) and constructed a kinship matrix to account for intra-ethnic population structure and relatedness in downstream mixed-model (EMMAX)-based association analyses16,86. We also used IBS to identify pairs of related individuals within each ancestry group (defined by π-hat > 0.3). We then defined intra-ethnic related exclusion lists for downstream non-EMMAX association analyses using the following steps: (i) remove the control from each T2D-status discordant pair; and (ii) remove the sample with lowest call rate from each T2D-status discordant pair. Within each of the 2,000 individuals (2,000 individuals) (Extended Data Fig. 2). We repeated our calculations for combined exome sequence and exome array data, assuming a fixed-effect across all ethnicities, for an effective total sample size of 82,758 individuals (Extended Data Fig. 2).
Single-variant association analysis. Within each ancestry group, we performed a score test of T2D association with each variant passing ethnic-specific QC in a linear regression framework under an additive model in EMMAX. We also performed a Wald test of T2D association with each variant passing ethnic-specific QC in a logistic regression framework under an additive model with adjustment for ethnic-specific axes of genetic variation after exclusion of related samples (Supplementary Table 30). Within each ancestry group, we calculated genomic control inflation factors (score EMMAX and Wald) based on independent variants used for the ethnic-specific kinship analyses and corrected association summary statistics (P value and s.e.) to account for residual population structure.

Subsequently, we performed trans-ethnic fixed-effects meta-analysis of ancestry-specific association summary statistics at each variant based on: (i) sample size weighting of score EMMAX directed P values; and (ii) inverse-variance weighting of Wald beta/s.e. (to obtain unbiased estimates of allelic odds ratios and confidence intervals that cannot be constructed from EMMAX effect estimates). We also performed trans-ethnic meta-analysis of ancestry-specific association summary statistics (score EMMAX beta/s.e.) at each variant using MANTRA, using pair-wise mean allele frequency differences at the subset of independent variants used for trans-ethnic kinship analyses as a prior for relatedness between ancestry groups.

Validation of PAX4 association signal in additional East Asian studies. We validated the PAX4 Arg192His (rs2233580) association signal in an additional 1,789 T2D cases and 1,509 controls of East Asian ancestry from Hong Kong, Korea, and Singapore (Supplementary Table 9). Within each study, we tested for association with T2D in a logistic regression model, and combined association summary statistics across studies through fixed-effects meta-analysis (Supplementary Table 9). Among T2D cases, we also tested for association with age of diagnosis in a linear regression model, and combined association summary statistics across studies through fixed-effects meta-analysis (Supplementary Table 9).

Admixture analysis. Admixed populations can offer greater statistical power to detect association because diverse ancestry increases genetic variation. However, admixture can also introduce false-positive signals due to population stratification and heterogeneity of effects because of differential LD. To assess the contribution of ancestral background in the two admixed groups (African American and Hispanic), we inferred local ancestry based on SNVs in available GWAS data using two approaches. For African Americans, we ran HAPMIX using CEU and YRI haplotypes from HapMap as reference, and estimated the proportion of European ancestry at each genomic position. For Hispanics, we ran Multinmix using European, West African, and Native American haplotypes from HapMap as reference, and estimated the proportion of European ancestry at each genomic position, since we observe only a very low West African contribution (1.1–3.2%, Supplementary Fig. 31). We then repeated our intra-ethnic EMMAX-based analyses within African American and Hispanic ancestry groups, this time adjusting for local ancestry by including the estimated proportion of European ancestry at each variant as a covariate. Adjustment for local ancestry resulted in numerically similar association statistics as those from unadjusted analyses in the African American and Hispanic samples.

Gene-based analysis. We generated four variant lists (‘masks’) based on MAF and functional annotation. We mapped variants to transcripts in Ensembl 66 (GRCh37.66). Using annotations from CHAoS v0.6.3, SnEff v3.1, and VEP v2.7, we identified variants predicted to be protein-truncating (for example, nonsense, frameshift, essential splice site) denoted PTV-only or ‘Mask 1’; or protein-altering (for example, missense, in-frame indel, non-essential splice site) in at least one mapped transcript (by at least one of the three algorithms) with MAF <1%, denoted PTV+missense or ‘Mask 2’. We additionally used the procedure described by Purcell et al. to identify subsets of missense variants with MAF <1% meeting ‘strict’ criteria for being deleterious using annotations from Polyphen-2-HumDiv, Polyphen-2-HumVar, LRT, Mutation Taster, and SIFT; variants predicted deleterious by all five algorithms or by at least one algorithm were denoted PTV+NSmiss or ‘Mask 3’ and PTV+NSsmall or ‘Mask 4’, respectively. Indels predicted by CHAoS, SnEff, or VEP to introduce frameshifts were included in the ‘strict’ category. We calculated MAFs for each ancestry using high-quality genotype calls (QC >20) for all samples passing extended QC. We considered a variant to have MAF <1% if MAF estimates for every ancestry group were <1%.

We used the MetaSKAT R package (v0.33) with the SKAT v0.93 library to perform SKAT-O analysis within each ancestry, and in meta-analysis. Within each ancestry group, we analysed genotype dosages with adjustment for ethnic-specific axes of genetic variation after exclusion of 96 related individuals. We assumed homogenous allele frequencies and genetic effects across all studies within an ancestry group. We performed meta-analysis using genotype-level data, allowing for heterogeneity of allele frequencies and genetic effects between (but homogeneity within) ancestry groups. All analyses were completed using the recommended rho vector for SKAT-O: (0, 0.12, 0.22, 0.32, 0.52, 0.5, 1).

Imputed data

Samples. We carried out genotype imputation into 44,414 individuals (11,645 T2D cases and 32,769 controls) from 13 studies using the GoT2D integrated haplotypes as reference panel. Characteristics of the imputed studies are provided in Extended Data Table 1 and Supplementary Table 3. Single-variant association meta-analysis. The one sequenced and thirteen imputed studies totalled 12,971 T2D cases and 34,100 controls. Each study performed its own sample- and variant-based QC. In each study, SNVs with minor allele count (MAC) ≥1 passing QC were tested for T2D association assuming an additive genetic model adjusting for study-specific covariates. Association testing was performed using logistic regression Firth bias-corrected, likelihood ratio, or score tests as implemented in EPACTS (genome.sph.umich.edu/wiki/EPACTS) or SNPTEST. To account for related samples in the Framingham Heart Study, generalized estimating equations (GEE) were used, as implemented in R. Residual population stratification for each study was accounted for using genomic control. We then carried out fixed-effects sample-size weighted meta-analysis as implemented in METAL.

Conditional analyses in established GWAS loci. We compiled a list of 143 previously reported genome-wide significant SNVs in 81 T2D autosomal loci (i) from Morris et al. and Voight et al.; (ii) from papers they referenced; and (iii) from references in the NHGRI GWAS catalogue. We LD pruned these SNVs (r2 <0.95), yielding a list of 129 SNVs. We deleted the CILP2 locus (and two SNVs) from subsequent whole-genome analyses owing to large regions in which no variants passed QC, resulting in a list of 127 index SNVs at 80 autosomal loci. To identify additional T2D-associated variants within these 80 T2D autosomal loci in the genome-wide data, we repeated GWA analysis for 12 of the 13 studies (conditional analysis results for FHS were unavailable), conditioning on the 127 index SNVs. We performed fixed-effects inverse-variance meta-analysis to combine conditional analysis results from the studies totalling 12,298 cases and 26,440 controls. For each known locus, we analysed all SNVs within 500kb of the known index SNVs; if there were multiple known index SNVs, we analysed all SNVs within 500kb of the most proximal and distal index SNVs. We imposed a conditional-analysis significance threshold of α = 1.8 × 10−8 based on a proportional number of multiple tests for ∼83 Mb of the ∼3,000-Mb genome.

Exome array data

Samples. We estimated 28,305 T2D cases and 51,549 controls from 13 studies of European ancestry, genotyped with the Illumina exome array. Characteristics of the studies are provided in Extended Data Table 1 and Supplementary Table 15. Overlap of exome sequence variation with exome array. We assessed overlap of variants present on the exome array with those observed in our trans-ethnic exome-sequence data. As exome array primarily contains SNVs that are predicted to be protein altering, we focused on nonsense, essential splice site, and missense variants. Only variants passing QC in both sequence and array data were included in our overlap assessment.

Data processing, QC, and kinship analysis. Within each study, exome array data were initially called using Illumina GenCall (https://www.illumina.com/Documents/products/technotes/technote_gencall_data_analysis_software.pdf) and Birdseed. Sample and variant QC was then undertaken within each study based on several quality control filters. Criteria for sample exclusion included low call rate (<99%), mean heterozygosity, high singleton counts, non-European ancestry, sex discrepancy, GWAS discordance (where data were available), genotyping platform fingerprint discordance, and duplicate discordance. Variants were excluded based on call rate (<99%), deviation from Hardy–Weinberg equilibrium (P >10−6), duplicate, chromosome or allele mismatch, GenTrain score <0.6, cluster separation score <0.4, and manual cluster checks. Missing genotypes were imputed on a per-sample basis using xCall, with a second round of QC to exclude poor quality samples (call rate <99% and mean heterozygosity) and variants (call rate <99%). Within each study, we considered independent autosomal variants that passed QC with MAF >1% for kinship analyses, and calculated IBS between each pair of samples. We used these statistics to: (i) identify non-European ancestry samples to be excluded from all downstream analyses; (ii) construct a kinship matrix to account for fine-scale population structure and relatedness in downstream EMMAX-based association analyses; (iii) identify related samples to be excluded from downstream non-EMMAX association analyses; and (iv) calculate axes of genetic variation for inclusion as covariates in downstream non-EMMAX association analyses. We compiled a list of 143 previously reported genome-wide significant SNVs in 81 T2D autosomal loci (i) from Morris et al. and Voight et al.; (ii) from papers they referenced; and (iii) from references in the NHGRI GWAS catalogue. We LD pruned these SNVs (r2 <0.95), yielding a list of 129 SNVs. We deleted the CILP2 locus (and two SNVs) from subsequent whole-genome analyses owing to large regions in which no variants passed QC, resulting in a list of 127 index SNVs at 80 autosomal loci. To identify additional T2D-associated variants within these 80 T2D autosomal loci in the genome-wide data, we repeated GWA analysis for 12 of the 13 studies (conditional analysis results for FHS were unavailable), conditioning on the 127 index SNVs. We performed fixed-effects inverse-variance meta-analysis to combine conditional analysis results from the studies totalling 12,298 cases and 26,440 controls. For each known locus, we analysed all SNVs within 500kb of the known index SNVs; if there were multiple known index SNVs, we analysed all SNVs within 500kb of the most proximal and distal index SNVs. We imposed a conditional-analysis significance threshold of α = 1.8 × 10−8 based on a proportional number of multiple tests for ∼83 Mb of the ∼3,000-Mb genome.
related samples. For each test, we corrected s.e. and P value for the genomic control inflation factor (\( \hat{\lambda} \geq 1 \)) calculated on the basis of the independent autosomal variants used for kinship analysis.

Across studies, we performed fixed-effects meta-analysis of association summary statistics at each variant based on: (i) inverse-variance weighting of score EMMAΧ beta/s.e.; (ii) sample size weighting of score EMMAΧ directed P values; and (iii) inverse-variance weighting of Wald beta/s.e. For each of these meta-analyses, we applied a second round of correction of s.e. and P value by genomic control, again calculated based on the independent autosomal SNVs's used for kinship analyses.

**Combined exome sequence and exome array single-variant analysis.** We considered variants that were represented both in the exome sequence and on the exome chip. We began by performing fixed-effects meta-analysis of association summary statistics (after correction for genomic control, as described above) from the exome-chip meta-analysis and the European ancestry sequenced samples using: (i) inverse-variance weighting of score EMMAΧ beta/s.e.; (ii) sample size weighting of score EMMAΧ directed P values; and (iii) inverse-variance weighting of Wald beta/s.e. Subsequently, we performed trans-ethnic fixed-effects meta-analysis of ancestry-specific association summary statistics (after correction for genomic control, as described above) at each variant based on: (i) sample size weighting of score EMMAΧ directed P values; and (ii) inverse-variance weighting of Wald beta/s.e.

**Gene-based analyses.** We made use of the four variant masks defined for exome sequence gene-based analyses, but with MAF calculated across all exome array studies. Within each study, we performed SKAT-O analyses\(^\text{31}\), with adjustment for axes of genetic variation after exclusion of related samples. We combined P values for association across studies via meta-analysis with Stouffer’s method\(^\text{32}\),

\[
\frac{P}{\sum_{i=1}^{k} \frac{1}{P_i}} < \frac{0.05}{3,147} < \frac{0.05}{634} \times 1.6 < 7.9 < 10^{-5}
\]

We note that by reducing the stringency of the significance threshold for variants used for kinship analysis.

**Enrichment of exome association signals in GWAS.** To define T2D-associated intervals, we first identified all SNVs associated with T2D in published genome-wide association studies (GWAS) by searching literature and the NHGRI GWAS Catalogue (see also Conditional analyses in established GWAS loci above). We identified 143 autosomal SNVs, with some associated in more than one ancestry (167 SNV-ancestry pairs). For each SNV-ancestry pair, we identified the most distant SNV-ancestry pairs. We included all low-frequency and rare variants occurring on T2D-associated haplotypes (haplotypes on which the T2D-associated GWAS index SNV minor allele is present); and (ii) asked whether any combination of these low-frequency or rare variants could explain the effect observed at the T2D GWAS index SNV. We carried out these analyses restricting attention to protein-coding variants within the window and then again for all low-frequency and rare SNVs in the 5-Mb window.

To define common haplotypes at each locus, we used the phased whole-genome sequence data. We first employed the phased genotypes for common (MAF > 5%) variants segregating in the interval between recombination hotspots at the locus (to minimize the number of recombinant haplotypes identified). We next identified the haplotypes on which the T2D-associated (risk or protective) GWAS index SNV minor allele was present. We then assembled the set of low-frequency and rare variants from across the 5-Mb interval which occurred on the background of these T2D-associated common-variant haplotypes. Owing to recombination and imperfect phasing, low-frequency or rare (excluding singletons) variants are often observed on more than one haplotype background. We included all low-frequency or rare variants that occurred more frequently on a T2D-associated haplotype than on other haplotypes.

From this pool of low-frequency and rare variants, we considered only variants with the same direction of effect as the common GWAS index SNV minor allele, as required by the synthetic association hypothesis, which posits that low-frequency or rare variants of larger effect than the common SNV could induce a weaker association signal. We then used a greedy algorithm to select the low-frequency or rare variant which, when added to the index GWAS SNV's dosage in a logistic regression, most reduced the residual effect remaining at the index SNV, as measured by the estimated additional effect. We repeated this process, adding variants to the model, until the estimated effect at the index GWAS genotype or gene dosage changed sign, representing no residual effect of the index SNV. At each locus, we also counted the number of variants required to increase the association P value at the GWAS index SNV beyond the nominal \( P = 0.05 \) significance threshold (Extended Data Table 4). **Credible set analysis of GoT2D genome sequence data.** At 78 of the 80 T2D GWAS loci (see Conditional analyses in established GWAS loci above), the previously reported index SNV had MAF > 1% in our GoT2D genome-sequenced sample. At these 78 loci, we constructed credible sets of common variants that, with some minimum specified probability (for example, \( > 99% \)), contain the variant which, when added to the index GWAS SNV's dosage in a logistic regression, most reduced the residual effect remaining at the index SNV, as measured by the estimated additional effect. We repeated this process, adding variants to the model, until the estimated effect at the index GWAS genotype or gene dosage changed sign, representing no residual effect of the index SNV. At each locus, we also counted the number of variants required to increase the association P value at the GWAS index SNV beyond the nominal \( P = 0.05 \) significance threshold (Extended Data Table 4).

For each GWAS index SNV, we identified the set of common variants with \( r^2 \geq 0.10 \) with the index SNV within a 5-Mb window centred on the index SNV. For each variant in this set, we calculated the posterior probability of being causal\(^\text{31}\). We first calculated an approximate Bayes' factor (ABF) for each variant as:

\[
\text{ABF} = \frac{1}{\sqrt{\pi}} r^{n/2}
\]

where \( r = 0.04 \) (s.e. = 0.04), \( z = \beta / \text{s.e.} \), and \( \beta \) and s.e. are the estimated effect size (log OR) and its standard error from logistic regression. We then calculated the posterior probability for each variant as \( \text{ABF}/T \), where \( T \) is the sum of the ABF values over all candidate variants across the interval. This calculation assumes a
Gaussian prior with mean 0 and variance 0.04 for $\beta$, the same prior employed in the commonly used single-variant association program SNPTEST32.

We based the analysis on the genome-wide meta-analysis results, since most common variants were included in this analysis, and sample sizes were significantly larger than for the genome sequence data alone.

We calculated the effective imputed sample size for each variant in the meta-analysis data as $N_{\text{eff}} = \sum_{j} r_j N_j^{\text{im}}$, where $r_j$ is the imputation quality and $N_j^{\text{im}}$ is the effective sample size for imputation cohort $j$. To ensure approximately sample size across variants, we considered to be well-imputed only those variants with effective imputed sample size ($N_{\text{eff}}$) $> 80\%$ of the maximum observed across all variants in the window.

Indels were not imputed or meta-analysed in this study, and $< 2\%$ of common SNVs were not well-imputed by the above effective sample size criterion. To include these common variants while using the most precise estimates available, we calculated posterior probabilities separately for each genome-wide data source. Where an indel from the sequence data set had a SNV proxy in high LD ($r^2 > 0.80$) in the meta-analysis data set, we used the proxy's information instead. Where a common SNV that poorly imputed had high-quality association data from the genome sequence data set alone, the posterior probability from the genome sequence data set was used instead. In each case, the final posterior probabilities for all SNVs were re-scaled such that their sum across a locus equaled one.

We used these final posterior probabilities to rank variants in decreasing order. To define credible sets of a specified level (for example, 99%), we included variants with highest final posterior probabilities until their sum reached or exceeded that level (Supplementary Table 28).

**Genomic enrichment analyses of the GoT2D genome sequence data**

**Genomic annotation.** We collected genome annotation data from several sources. First, we obtained gene transcript masks from GENCODEv14 (ref. 80). For protein-coding genes, we included transcripts with a protein-coding tag that either were present in the conserved coding DNA sequence (CCDS) database or had experimentally confirmed mRNA start and end; we then included 5′ UTR, exon, and 3′ UTR regions from the resulting transcripts. For non-coding genes, we included transcripts with a lncRNA, miRNA, snRNA, or snRNA tag.

Second, we defined regulatory chromatin states in 12 cell types. We collected sequence reads generated for the following assays: H3K4me1, H3K4me3, H3K27ac, H3K27me3, H3K36me3, and CTCF ChIP, in nine ENCODE cell types (GM12878, K562, HePG2, H1mM, HUVEC, NHEK, NHLF, hESC, HMEC)32, pancreatic islets35, and hASC (adipose stromal cell) pre- and mature adipocytes33.

We mapped reads to hg19 using BWA51 and used the resulting mapped reads for all cell types to call regulatory states using ChromHMM41, assuming ten states. We then assigned names to the resulting state definitions: (i) H3K4me3, H3K27ac (active promoter); (ii) H3K4me3, H3K27ac, H3K4me1 (active enhancer 1); (iii) H3K27ac, H3K4me1 (active enhancer 2); (iv) H3K4me1 (weak enhancer); (v) H3K27me3, H3K4me3, H3K4me1 (poised promoter); (vi) H3K27me3 (repressed); (vii) low/no signal 1; (viii) CTFC (insulator); (ix) low/no signal 2; and (x) H3K36me3 (transcription).

Third, we obtained transcription factor-binding ChIP sites from three sources: 141 proteins from ENCODE35, 5 from Pasquali et al35, and 1 from Mikkelson et al33.

From gene transcript data we defined CDS (protein coding transcript exons); ncRNA (non-coding RNA transcripts); and 3′ and 5′ UTR (UTR regions of coding transcripts). From chromatin state data for each of the 12 cell types we identified active enhancers (pooled active enhancer 1 and 2 elements; weak enhancers; and active promoters. From transcription factor binding sites we defined transcription factor-binding sites (TFBS) sites pooled across all factors). This resulted in a total of 41 annotation categories (Extended Data Fig. 6).

**Enrichment of association signal enriched sub-networks.** We identified p-values for individual gene mutations using the Firth penalized likelihood method. Analysis in the exome array data set was performed with default parameter settings (0.3 as density threshold, 0.8 as permutation p-value node), and with the–fluff option.

The first approach consists of two steps. First, the entire human PPI network was searched for protein complexes (clusters) using the algorithm implemented in clusterONE34, which identifies protein complexes with high cohesiveness. The method was run with default parameter settings (0.3 as density threshold, 0.8 as merging threshold, and 2 as the penalty-value node), and with the–fluff option activated, which allows the addition of highly connected boundary nodes to the cluster. Second, gene-based association $P$ values derived from SKAT-O analyses of the 12,940 individual gene mutations were aggregated, using Fisher's method, for the genes encoding each of the proteins within a cluster to generate a ‘cluster association’ statistic.

An empirical $P$ value for the significance of these aggregated cluster association statistics was derived by comparing each cluster to a large number of complexes of the same topology, but composed of randomly sampled proteins. Specifically,
a background distribution was obtained for each protein complex as follows: each protein in the cluster was randomly substituted by a different protein represented in the InWeb database, matched for number of minor allele carriers in the data set. SKAT-O P values were assigned to each protein from the exome sequencing results, and an aggregated P value was obtained for each pseudo-complex using Fisher’s method, as above. This process was repeated 100,000 times, and the empirical P value for each complex was calculated as the proportion of the iterations for which the Fisher’s P value of the observed complex was more significant than that of P value of the maximum number of complexes. This procedure was repeated for all gene-level masks (PTV only, PTV + missense, PTV + NSmissense and PTV + NSmissense).

To test the study-wide significance of apparently associated clusters, we used two permutation designs. In the first design, we generated 100,000 pseudo-complexes for each cluster, replacing each protein within each cluster with one protein from InWeb3, matched for the number of minor allele carriers in the data set. We calculated the number of permuted data sets which generated any ‘pseudo-cluster’ association P value more significant than our most enriched cluster. In the second design, we used a Monte-Carlo algorithm to generate 10,000 random PPI networks, with the same degree as observed in the InWeb database, ran clusterONE on each, and once again compared the distribution of ‘best’ cluster association P value with that observed in the real data.

The second approach uses the dense module searching algorithm (a heuristic ‘greedy’ method) described in dmGWAS45, where a module is defined as a sub-network within the whole network if it contains a locally increased proportion of low P value genes. This method differs from the earlier method in using the association P values, in combination with the PPI data, to construct the networks. The module is grown for each protein in the PPI by adding the neighbouring nodes within a pre-defined distance (d = 2) that can yield a maximum increment of the module score \( Z_m = \sum Z_k \sqrt{k} \) for module m, where k is the number of genes in the module, and \( Z_k \) is calculated from the P value of exome gene-based tests using an empirical normal distribution function. The addition of neighbourhood nodes is stopped when the increment is less than 10% of \( Z_m \) (that is, \( Z_m \times (1 - 0.1) \)). As with the clusterONE approach, this procedure was conducted for all four exome gene-level based masks.

To evaluate whether the top ranked-modules are significantly associated with T2D, we permuted case-control status across the 12,940 exomes (maintaining ethnic strata) 10,000 times and generated 10,000 SKAT-O gene-based association tests on all genes in the top 15 modules (once for each gene-based variant mask, 40,000 in total). During each permutation, \( Z_m \) was re-calculated for each module, and a set of empirical P values was obtained by comparing the P value of the original module to these modules with the SKAT-O results from the swapped labels. Following the above procedure, all 15 top modules were found significantly enriched for the PTV + NSmissense and PTV + NSmissense gene-based variant masks (P < 10^-4, after the 10,000 case-control permutations).

Modelling disease architecture T2D liability risk and architecture bounding in the exome array data. We used a Bayesian framework implemented in R to compute the probability that each variant explains more than a defined amount of the T2D risk liability-scale variance (LVE). The joint distribution in the MAF-OR space is computed by assuming a T2D prevalence of 8% and beta and normal distributions for the MAF and OR, respectively. The OR is calculated with reference to the minor allele. The MAF is adjusted to take account of apparent allele frequency heterogeneity between cohorts (subjects from missing cohorts are excluded from calculations). Analyses are restricted to variants with MAF ≥ 0.1% because the representation of variants with MAF below this threshold on the exome array is poor. The probability is obtained by numerically integrating over the joint distribution for MAF–OR combinations that explain more than the defined amount of liability-scale variance. For bounding the maximum number of variants that could contribute to T2D risk variance, we performed a sensitivity analysis on the 88 known T2D index SNVs. Ranges of OR and MAF consistent with 80% power to detect single-variant association in this population simulated with both phenotypes and genotypes, we selected appropriate numbers of cases and controls and conducted single-variant association tests in order to compare the distribution of P values from simulation to that observed in the current study. Results shown are the average of 25 independent simulation replicates for each disease model.

Comparison of simulated outcomes to empirical T2D results. We focused on comparing simulated outcomes under three disease models, each of which were previously found to be consistent with sibling relative risk, GWAS, and linkage results for T2D, but vary widely in causal variant properties (Fig. 3): a rare-variant model in which rare variants explain ~75% of T2D heritability (small target size T = 750 kb and moderate dependence between effect size and selection \( \tau = 0.5 \)), an intermediate model in which rare, low-frequency, and common variants all contribute significantly to T2D heritability (T = 2.0 Mb and \( \tau = 0.3 \)), and a common polygenic model in which common variants explain ~75% of T2D heritability (T = 3.75 Mb and weak dependence \( \tau = 0.1 \)). We first compared the simulated outcomes of a whole-genome sequencing study in ~3,000 samples under each model. All three models predicted similar distributions of variant association test statistics using the sequenced individuals alone (data not shown).

However, the predictions began to diverge when we simulated imputation into GWAS samples and studied the distribution of test statistics after meta-analysis. For each simulated model, we sampled 14,175 cases and 14,175 controls (to match the effective sample size of the actual imputation cohorts used for meta-analysis). Because genotyping accuracy in simulated samples is perfect (unlike in imputation), for both imputed and unphased imputation quality as a function of MAC in the empirical data (using the \( r^2 \) value reported by the imputation software that was used in each cohort). We then corrected, for each variant, the association test statistic in simulated data by multiplying the \( \chi^2 \) value by the average imputation \( r^2 \) for the variant MAC. We then re-computed association P values from the corrected chi-squared statistics to compare P value distributions in simulated versus empirical data. We plotted the distribution of association P values for variants of different frequency classes in a quantile–quantile (QQ) plot, and compared these curves to the empirical T2D results (Fig. 3). Focusing on low-frequency variants, we also asked how many unique low-frequency signals achieved significant association at a given T2D risk under each simulated model, and compared these quantities to empirical observation (Fig. 3). These analyses demonstrate that the intermediate and rare-variant models produce an excess of association signal among low-frequency variants compared to observation, whereas the common polygenic model is consistent with the genome-wide distribution of association signals observed.
50. Guex, L. T. et al. Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. Genet. Epidemiol. 35, 236–246 (2011).
51. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
52. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498 (2011).
53. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
54. Jun, G. et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotyping data. Am. J. Hum. Genet. 91, 839–848 (2012).
55. Abecasis, G. R. et al. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012).
56. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genomic structural polymorphism by sequencing on a population scale. Nat. Genet. 43, 269–276 (2011).
57. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81, 1084–1097 (2007).
58. Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. R. Low-coverage sequencing: implications for design of complex trait association studies. Genet. Epidemiol. 35, 809–822 (2011).
59. Price, A. L. et al. Long-range LD can confound genome scans in admixed populations. Am. J. Hum. Genet. 83, 132–135, author reply 135–139 (2008).
60. Weale, M. E. Quality control for genome-wide association studies. Methods Mol. Biol. 628, 341–372 (2010).
61. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 475, 803–808 (2011).
62. Price, A. L. et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature 506, 185–190 (2014).
63. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13, 762–775 (2012).
64. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39, 906–913 (2007).
65. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
66. Eyre-Walker, A. Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc. Natl Acad. Sci. USA 107, 1752–1756 (2010).
67. Purcell, S. M. et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature 506, 185–190 (2014).
68. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13, 762–775 (2012).
69. Churchhouse, C. & Marchini, J. Multway admixture deconvolution using phased or unphased ancestral panels. Genet. Epidemiol. 37, 1–12 (2015).
70. Purcell, S. M. et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature 506, 185–190 (2014).
71. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13, 762–775 (2012).
72. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39, 906–913 (2007).
Extended Data Figure 1 | Summary of samples and quality control procedures. This figure summarizes data generation for whole-genome sequencing (GoT2D), exome sequencing (GoT2D and T2D-GENES), exome array genotyping (DIAGRAM), and GWAS imputation (DIAGRAM).
Extended Data Figure 2  |  See next page for caption.
Extended Data Figure 2 | Power for single and aggregate variant association. 

a–g. Power to detect single-variant association ($\alpha = 5 \times 10^{-8}$) at varying minor allele frequencies (x-axis) and allelic ORs (y-axis) for seven effective sample size ($N_{\text{eff}}$) scenarios relevant to the genomes (a–c) and exomes (d–g) components of this project. 

a, Variant observed in 2,657 samples (the effective size of the GoT2D integrated panel). 
b, Variant observed in 28,350 samples (the effective size of the imputed data set). 
c, Variant observed in the GoT2D integrated panel and the imputed data set (effective sample size 31,007). 
d, Ancestry-specific variant in 2,000 samples (the size of each of the non-European exome sequence data sets). 
e, European-specific variant in 5,000 samples (the combined size of the European exome sequence data sets). 
f, Variant observed with shared frequency across all ancestry groups in 12,940 samples (the size of the combined exome sequence data set). 
g, Variant observed in the combined exome array and sequencing data set (effective sample size 82,758). 

h, i, Power for gene-based test of association (SKAT-O) according to liability variance explained. In h, 50% of the variants contribute to disease risk and the remaining 50% have no effect on disease risk; in i, 100% of the variants contribute to disease risk. For each, sample sizes considered are 2,000 (ancestry-specific effects; green) and 12,940 (ancestry-shared effects; blue). Power is shown for two levels of significance ($\alpha = 2.5 \times 10^{-6}$ and $\alpha = 0.001$). From these simulation studies, it is clear that under the optimistic model, where effects are shared across all ethnicities (blue line) and all variants contribute, power is >60% for 1% variance explained and $\alpha = 2.5 \times 10^{-6}$. However, power declines rapidly if either criterion is relaxed.
Extended Data Figure 3 | Single variant analyses. a–c, Manhattan plot of single-variant analyses generated from exome sequence data in 6,504 cases and 6,436 controls of African American, East Asian, European, Hispanic, and South Asian ancestry (a); exome array genotypes in 28,305 cases and 51,549 controls of European ancestry (b); and combined meta-analysis of exome array and exome sequence samples (c). Coding variants are categorized according to their relationships to the previously reported lead variant from GWAS region. Loci achieving genome-wide significance only in the combined analysis are highlighted in bold. The HNF1A variant reaching genome-wide significance in the combined analysis is a synonymous variant (Thr515Thr). The dashed horizontal line in each panel designates the threshold for genome-wide significance ($P < 5 \times 10^{-8}$).
Extended Data Figure 4 | See next page for caption.
Extended Data Figure 4 | Classification of coding variants according to their relationship to reported lead variants for each GWAS region.

The ideogram shows the location of 25 coding variant associations at 16 loci described in the text. The number in each circle corresponds to the number of associated variants at each locus. Variants are grouped into five categories based on inferred relationship with the GWAS lead variant. For some of these categories, the figure includes representative regional association plots based on exome array meta-analysis data from 28,305 cases and 51,549 controls. The locus displayed for each category is designated in bold. The first plot in each panel shows the unconditional association results; the middle plot the association results after conditioning on the non-coding GWAS SNP; and the last plot the results after conditioning on the most significantly associated coding variant.

Each point represents an SNP in the exome array meta-analysis, plotted with its $P$ value (on a $-\log_{10}$ scale) as a function of the genomic position (hg19). In each panel, the lead coding variant is represented by the purple symbol. The colour-coding of all other SNPs indicates LD with the lead SNP (estimated by European $r^2$ from 1000G March 2012 reference panel: red $r^2 \geq 0.8$; gold $0.6 \leq r^2 < 0.8$; green $0.4 \leq r^2 < 0.6$; cyan $0.2 \leq r^2 < 0.4$; blue $r^2 < 0.2$; grey $r^2$ unknown). Gene annotations are taken from the University of California Santa Cruz genome browser. GWS: genome-wide significance. *Seven variants, three at ASCC2, and one each at THADA, TSPAN8, FES and HNF4A did not achieve genome-wide significance themselves, but are included because they fall into genes and/or regions with other significant association signals (see text).
Extended Data Figure 5 | Exclusion of synthetic associations and construction of credible causal variant sets at T2D GWAS loci. Ten T2D GWAS loci were selected for synthetic association testing ($P < 0.001$; see Methods). a, The effect size observed at the GWAS index SNV (sequence data) before (navy blue) and after (light blue, grey) conditioning on candidate rare and low-frequency (MAF < 5%) variants which could produce synthetic association. b, Example of synthetic association exclusion at the TCF7L2 locus. Error bars represent 95% confidence intervals for the index SNP odds ratio as rare variants are greedily added to the model. c, The size of credible sets at T2D GWAS loci when constructed from the GoT2D data, compared to the sizes when restricted to variants in the 1000G or HapMap data.
Extended Data Figure 6 | Genome enrichment analysis in GoT2D whole genome sequence data. n = 2,657. a, Functional annotation categories were defined using transcription, chromatin state and transcription factor binding data from GENCODE, ENCODE and other studies. b, T2D association statistics for variants at each T2D locus were jointly modelled with functional annotation using fgwas. In the resulting model we identified enrichment of coding exons (CDS), mature adipose active enhancers and promoters (hASC-t4 EnhA, TssA), pancreatic islet active and weak enhancers (HI EnhA, EnhWk), pre-adipose active and weak enhancers (hASC-t1 EnhA, EnhWk), embryonic stem cell active promoters (H1-hESC TssA) and 5′ UTRs. Dots represent enrichment estimates and horizontal lines the 95% confidence intervals. c, At the CCND2 locus, three variants not present in HapMap2 have a combined 90% posterior probability of being causal (rs4238013, rs3217801, rs73040004). One of these variants, rs3217801, is a 2-bp indel that overlaps an islet enhancer element.
Extended Data Figure 7 | Low frequency variants in exome array data.
Results from meta-analysis of 43,045 low-frequency and common coding variants on the exome array (assayed in 79,854 European subjects).
a, Observed allelic ORs as a property of allele MAF. Variants missing in more than eight cohorts or polymorphic in only one cohort were excluded. Coloured lines represent contours for liability variance explained. Regions shaded grey denote ranges of OR and MAF consistent with 80% power (in this case, at $\alpha = 5 \times 10^{-7}$) to detect single-variant associations in this data set (given the observed range of missing data). Variants with a black collar are those highlighted by a bounding analysis as having a probability $>0.8$ of having liability-scale variance (LVE) $>0.1\%$. b, Distribution of each variant in the MAF/OR space was computed by assuming T2D prevalence of 8% and a beta and normal distribution for MAF and OR, respectively. Probability is obtained by integrating the joint MAF–OR distributions over ranges of LVE. c, Single variant association, liability and bounding results for the known T2D GWAS variants on the exome array (see Methods).
## Extended Data Table 1 | Summary information for sample sets used in the association analyses

| Ancestry | Study | Countries of Origin | Num. of Cases (% female) | Num. of Controls (% female) | Effective Sample Size |
|----------|-------|---------------------|--------------------------|----------------------------|----------------------|
| Whole Genome Sequencing Studies | European | Finland–United States investigation of NIDDM Genetics (FUSION) Study | Finland | 489 (41.3) | 486 (45.2) | 979 |
| | European | Kooperative Gesundheitsforschung in der Region Augsburg (KORA) | Germany | 139 (44.5) | 150 (66.3) | 205 |
| | European | Malmö-Borås Study | Finland, Sweden | 410 (51.5) | 419 (44.1) | 829 |
| | European | UK Type 2 Diabetes Genetics Consortium (UKT2D) | UK | 322 (46.2) | 322 (82.2) | 644 |
| Total Whole Genome Sequence | | | | | | 1,526 |
| Genome-Wide Array Studies | European | INTERACT | France, Germany, Italy, Netherlands, Spain, Sweden, UK | 4624 (51.8) | 4668 (64.2) | 9292 |
| | European | Welcome Trust Case Control Consortium (WTCCC) | UK | 1568 (49.6) | 2936 (50.6) | 4120 |
| | European | Kooperative Gesundheitsforschung in der Region Augsburg (KORA) | Germany | 993 (45.1) | 2995 (52.2) | 2990 |
| | European | Framingham Heart Study (FHS) | US | 673 (42.5) | 7660 (60.5) | 2475 |
| | European | Finland–United States investigation of NIDDM Genetics (FUSION) Study | Finland | 1060 (43.1) | 1090 (51.3) | 2150 |
| | European | Diabetes Genetics Initiative (DGI) | Finland, Sweden | 899 (46.6) | 1657 (49.8) | 1943 |
| | European | Estonian Genome Center, University of Tartu (EGCUT–OMNI) | Estonia | 399 (58.9) | 6813 (52.4) | 1461 |
| | European | Diabetes Gene Discovery Group (DGDG) | France, Canada | 677 (39.3) | 897 (59.7) | 1374 |
| | European | Mt Sinai Biomarker Platform (BioMe; Illumina) | US | 255 (29.5) | 1647 (51.4) | 883 |
| | European | Uppsala Longitudinal Study of Adult Men (ULSAM) | Sweden | 166 (6) | 950 (53) | 565 |
| | European | Mt Sinai BioMe Biomarker Platform (BioMe) | US | 132 (28.5) | 455 (34.7) | 409 |
| | European | Prospective Investigation of the Vasculature in Uppsala Seniors (PVUS) | Sweden | 111 (41.4) | 838 (51.2) | 392 |
| | European | Estonian Genome Center, University of Tartu (EGCUT–OMNI) | Estonia | 80 (46.8) | 1768 (51) | 306 |
| Total Genome-Wide Array | | | | | | 11,645 |
| Whole Exome Sequencing Studies | African American | Jackson Heart Study | US | 500 (65.6) | 526 (63.3) | 1,026 |
| | African American | Wake Forest School of Medicine Study | US | 518 (59.5) | 530 (56.6) | 1,048 |
| | East Asian | Korea Association Research Project | Korea | 526 (45.9) | 581 (58.5) | 1,066 |
| | East Asian | Singapore Diabetes Cohort Study; Singapore Prospective Study Program | Singapore (Chinese) | 436 (52.1) | 592 (61.3) | 1,028 |
| | European | Ashkenazi | US, Israel | 506 (47.0) | 355 (56.9) | 854 |
| | European | Metabolic Syndrome in Men Study (METSIM) | Finland | 484 (6) | 498 (0) | 982 |
| | European | Finland–United States investigation of NIDDM Genetics (FUSION) Study | Finland | 472 (42.6) | 476 (45.0) | 948 |
| | European | Finnish–United States investigation of NIDDM Genetics (FUSION) Study | Germany | 97 (44.3) | 90 (63.3) | 186 |
| | European | UK Type 2 Diabetes Genetics Consortium (UKT2D) | UK | 322 (45.7) | 329 (82.8) | 642 |
| | European | Malmö-Borås Study | Finland, Sweden | 478 (54.6) | 443 (43.8) | 920 |
| | | San Antonio Family Heart Study, San Antonio Family Diabetes/Gallbladder Study, Vatarans Administration Genetic Epidemiology Study, and the Investigation of Nephropathy and Diabetes Study Family Component | USA | 272 (58.8) | 218 (58.7) | 484 |
| | Hispanic | Starr County, Texas | US | 749 (59.7) | 704 (71.9) | 1,452 |
| | South Asian | London Life Sciences Population Study (LOLIPOP) | UK (Indian Asian) | 531 (14.1) | 538 (15.8) | 1,068 |
| | South Asian | Singapore Indian Eye Study | Singapore (Indian Asian) | 583 (44.4) | 585 (49.2) | 1,148 |
| Total Whole Exome Sequence | | | | | | 6,504 |
| Exome Array Studies | European | ADDITION: Steno Diabetes Centre (SDC); Health2000; Health2000; Vejle Biobank; Inter99 | Denmark | 5613 (40.6) | 7987 (54.4) | 13,458 |
| | European | Wellcome Trust Case Control Consortium (UK Type 2 Diabetes Consortium); Young Diabetics Study (YDS); Genetics of Diabetics and Audit Research Tayside Study (GORDATS); Oxford Biobank; TwinsUK; 1958 Birth Cohort (BCGS) | UK | 3575 (51.7) | 12675 (41.2) | 11,156 |
| | European | Finland–United States investigation of NIDDM Genetics (FUSION) Study; FinnDias; Metabolic Syndrome in Men Study (METSIM); Diet–Responses to Exercise Training (DIREXTRA); EXO2017 | Finland | 3956 (33.4) | 8222 (26.0) | 10,001 |
| | European | Malmö Diabetes Cohort (MDC); All New Diabetics in Skane (ANDIS) | Sweden | 4633 (41.0) | 5494 (59.5) | 9,078 |
| | European | Prevalence, Prediction and Prevention of Diabetes (PPPD); Diabetes Register in Vaasa (DIREVA) | Finland | 2910 (43.7) | 4596 (53.7) | 7,127 |
| | European | Nurses’ Health Study (NHS) | US | 1412 (100) | 1695 (100) | 3,082 |
| | European | Health Professionals Follow-up Study (HPFS) | US | 1104 (0) | 1287 (0) | 2,467 |
| | European | The Ewax Family Study of YDS Health; EFSCOCH | UK | 1448 (39.6) | 1587 (52.0) | 3,008 |
| | European | Kooperative Gesundheitsforschung in der Region Augsburg (KORA) | Germany | 933 (45.3) | 2705 (51.7) | 2,775 |
| | European | Estonian Genome Center at the University of Tartu (EGCUT) | Estonia | 892 (43.7) | 1506 (44.2) | 2,225 |
| | European | Gene–Lifestyle Interactions and Complex Traits Involved in Elevated Disease Risk (GLACIER) | Sweden | 960 (47.6) | 957 (54.5) | 1,917 |
| | European | Finland cohort of the European Prospective Investigation of Cancer (EPIC) | UK | 691 (47.0) | 1157 (54.5) | 1,730 |
| | European | The Prospective Investigation of the Vasculature in Uppsala Seniors (PVUS); Uppsala Longitudinal Study of Adult Men (ULSAM) | Sweden | 271 (16.9) | 1791 (23.9) | 942 |
| Total Exome Array | | | | | | 28,505 |
| Total Whole Exome Sequence + Exome Array | | | | | | 34,809 |

© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
Extended Data Table 2 | Counts and properties of variants identified in sequenced subjects

a

Genomes integrated panel

| Variant Type | SNV | Indel | SV |
|--------------|-----|-------|----|
| N (%/total)  | 25.2M (94%) | 1.50M (5.6%) | 8.876E-03 (0.03%) |

Coding Non-coding

| Function | N (%/total) |
|----------|-------------|
| Rare (MAF<0.5%) | 888K (3.3%) | 25.8M (97%) |
| Low frequency (0.5<MAF<5%) | 2.36M (33%) | 2.16M (16%) |
| Common (MAF>5%) | 4.16M (90%) | 16.3M (61%) |

| Frequency spectrum | N (%/total) |
|-------------------|-------------|
| b137 | Novel |
| 14.6M (55%) | 12.1M (45%) |

b

Exome sequence data

| Samples | All samples | African-American | East-Asian | European | Hispanic | South-Asian |
|---------|-------------|------------------|------------|----------|----------|-------------|
| T2D cases | 13,008 | 2,086 | 2,165 | 4,579 | 1,959 | 2,219 |
| T2D controls | 6,436 | 1,056 | 1,153 | 2,182 | 922 | 1,123 |

Excluded from association analysis | 68 | 12 | 0 | 28 | 16 | 2 |

Coverage:

Coding:

Mean (M) per gene | 81.7 ±23.7 | 83.2 ±24.0 | 84.6 ±23.8 | 78.6 ±23.3 | 83.8 ±24.1 | 78.2 ±23.2 |

| # of genes with M<20 | 368 | 302 | 302 | 351 | 269 | 325 |

Non-coding:

Mean per gene | 59.0 ±21.0 | 60.9 ±21.5 | 62.2 ±21.6 | 57.5 ±20.6 | 59.2 ±21.2 | 56.4 ±20.3 |

| # of genes with M<20 | 1,150 | 738 | 731 | 1,102 | 804 | 945 |

Variant annotations:

Synonymous SNV | 627,630 | 237,430 | 178,392 | 192,582 | 156,231 | 211,218 |

Missense SNV | 1,110,897 | 354,797 | 296,707 | 327,049 | 231,351 | 344,191 |

Start SNV | 2,055 | 593 | 523 | 629 | 384 | 583 |

Nonsense SNV | 26,321 | 7,188 | 6,668 | 8,030 | 4,660 | 7,339 |

Frameshift INDEL | 26,901 | 6,605 | 6,159 | 7,515 | 4,155 | 6,699 |

Inframe INDEL | 11,090 | 3,471 | 2,963 | 3,145 | 2,068 | 3,165 |

3UTR SNV, INDEL | 65,013 | 24,583 | 19,149 | 21,102 | 16,959 | 22,177 |

5UTR SNV, INDEL | 43,965 | 16,920 | 13,520 | 15,562 | 11,634 | 15,595 |

Intron SNV, INDEL | 931,449 | 352,368 | 270,564 | 296,070 | 243,139 | 314,810 |

Essential splicing SNV, INDEL | 14,286 | 3,648 | 3,454 | 4,108 | 2,301 | 3,744 |

Other splicing SNV, INDEL | 120,544 | 45,876 | 35,413 | 38,263 | 30,301 | 41,122 |

Non-coding RNA SNV, INDEL | 18,113 | 7,247 | 5,996 | 6,715 | 5,084 | 8,706 |

Intergenic SNV, INDEL | 37,345 | 14,335 | 11,498 | 13,614 | 10,700 | 12,957 |

All | 3,043,709 | 1,075,091 | 850,846 | 934,094 | 718,967 | 990,196 |

Coding frequency spectrum:

Rare (MAF<0.5%)

private | 77.93% | 53.79% | 65.47% | 51.80% | 37.26% | 61.55% |

| cosmopolitan | 0.85% | 1.80% | 3.02% | 1.88% | 2.24% | 1.73% |

Low frequency (0.5<MAF<5%)

| private | 2.57% | 10.36% | 4.61% | 5.52% | 8.21% | 5.10% |

| cosmopolitan | 0.17% | 1.43% | 1.10% | 0.26% | 0.52% | 1.02% |

Common (MAF>5%)

| private | 1.65% | 6.35% | 5.33% | 5.29% | 7.23% | 5.00% |

| cosmopolitan | 0.09% | 0.00% | 0.00% | 0.00% | 0.01% | 0.00% |

Intron/UTR frequency spectrum:

Rare (MAF<0.5%)

| private | 94.09% | 78.68% | 86.91% | 86.17% | 81.43% | 86.68% |

| cosmopolitan | 74.70% | 49.81% | 61.36% | 45.26% | 31.03% | 56.96% |

Low frequency (0.5<MAF<5%)

| private | 3.52% | 12.57% | 5.65% | 6.51% | 9.43% | 6.32% |

| cosmopolitan | 0.25% | 1.74% | 1.25% | 0.29% | 0.47% | 1.18% |

Common (MAF>5%)

| private | 2.39% | 8.76% | 7.46% | 7.32% | 9.14% | 7.00% |

| cosmopolitan | 0.15% | 0.00% | 0.00% | 0.01% | 0.00% | 0.00% |

a. Variant numbers for the 2,657 individuals with whole-genome sequence data passing QC and included in the association analysis data set. b. Variant numbers are provided for the 13,008 individuals passing initial rounds of QC from which further QC defined the 12,940 subjects included in the association analysis data set. Private refers to variants seen in only a single ancestral group; cosmopolitan to variants seen in all five major ancestry groups.
Extended Data Table 3 | Characterization of variant associations through conditional analysis

| Locus   | Variant | MAF       | Unconditional and conditional association $P$-values | Interpretation |
|---------|---------|-----------|-----------------------------------------------------|---------------|
| **Non-coding variant associations characterized in 38,738 samples in GoT2D genome-wide imputed meta-analysis** |
| IRS1    | rs78124264 | 0.022     | 8.5x10^{-6} 2.5x10^{-6} 2.5x10^{-6} 2.5x10^{-6} | The association signal rs78124264 and the GWAS SNPs at this locus are distinct. Signals are not extinguished in reciprocal conditional analysis. Previous GWAS signals are not mediated through rs78124264, which represents a distinct association signal at this locus. |
|         | rs7578326  | 0.35      | 1.2x10^{-6} 1.1x10^{-6} n.d. n.d. n.d. | |
|         | rs2943640   | 0.35      | 2.5x10^{-6} n.d. 4.5x10^{-6} n.d. | |
|         | rs2943641   | 0.36      | 9.0x10^{-6} n.d. n.d. 1.5x10^{-6} | |
| PPARG   | rs79656023 | 0.022     | 1.2x10^{-6} 9.2x10^{-6} | The association signal rs79656023 and the GWAS SNP at this locus are distinct. Signals are not extinguished in reciprocal conditional analysis. Previous GWAS signal is not mediated through rs79656023, which represents a distinct association signal at this locus. |
|         | rs1801282   | 0.13      | 1.6x10^{-6} 1.2x10^{-6} | |

**Coding variant associations characterized in 28,305 cases and 51,549 controls typed on exome array**

| PAM     | Asp563Gly (PAM) | Ser1207Gly (PPPRS2) | Association signals for PAM Asp563Gly and PPPRS2 Ser1207Gly are indistinguishable in reciprocal conditional analysis. Gene biology, as well as previous reports of additional PAM variants associated with T2D in Icelandic cohorts, highlights PAM as the probable transcript at this locus. |
|---------|-----------------|---------------------|--------------------------------------------------|
|         | Asp563Gly       | 0.054               | 1.7x10^{-6}                                     |               |
|         | Ser1207Gly      | 0.054               | 0.10                                            |               |

**MTMR-ASCC2**

| MTMR-ASCC2 | Asn960Ser (MTMR3) | Val123G1e (ASCC2) | Asp407His (ASCC2) | Pro422Ser (ASCC2) | Association signals for the MTMR3 and ASC2 coding variants are indistinguishable in reciprocal conditional analysis. The MTMR3 Asn960Ser variant has the strongest signal, and highlights MTMR3 as the most likely effector transcript at this locus. |
|-------------|------------------|-------------------|------------------|------------------|--------------------------------------------------|
| IRS1        | 0.083            | 3.2x10^{-6}       | 0.022            | 0.027            | 0.022                                            |
| PPARG       | 0.083            | 0.15              | 2.0x10^{-6}      | 0.056            | 0.78                                             |
| Coding      | 0.083            | 0.18              | 0.99             | 1.9x10^{-6}      | 0.88                                             |
| Coding      | 0.083            | 0.18              | 0.67             | 0.98             | 2.0x10^{-6}                                      |

**KCNJ11-ABC28**

| KCNJ11-ABC28 | Val337lle (KCNJ11) | Lys233Glu (ABC28) | Ala1369Ser (KCNJ11) | Association signals for KCNJ11 Val337lle and Lys233Glu and ABC28 Ala1369Ser are indistinguishable in reciprocal conditional analysis. The relative causal contributions of the two genes, making up the two components of the sulfonylurea-responsive potassium channel, are indistinguishable on statistical grounds. |
|--------------|-------------------|-------------------|--------------------|--------------------------------------------------|
| Coding       | 0.40              | 3.4x10^{-10}      | 0.17               | 0.049                                            |
| Coding       | 0.40              | 0.48              | 5.1x10^{-6}        | 0.082                                            |
| Coding       | 0.40              | 0.68              | 0.84               | 2.3x10^{-6}                                      |

**WFS1**

| WFS1       | Val333lle (WFS1) | Asn500Asn (WFS1) | Arg511His (WFS1) | rs4689388 | Association signals for the WFS1 coding variants are indistinguishable from each other and the previously reported non-coding GWAS SNP at this locus in reciprocal conditional analysis. WFS1 is the likely effector transcript for the non-coding GWAS signal at this locus, although the causal variant in the gene is unclear. |
|------------|------------------|------------------|------------------|----------|--------------------------------------------------|
| Coding      | 0.40             | 9.3x10^{-10}     | 0.024            | 0.00070  | 0.0000                                          |
| Coding      | 0.41             | 0.0070           | 2.0x10^{-12}     | 0.0049   | 0.027                                            |
| Coding      | 0.47             | 0.0200           | 1.3x10^{-10}     | 0.19     | 0.19                                             |
| Coding      | 0.43             | 0.0111           | 0.62             | 0.024    | 2.3x10^{-10}                                     |

**CILP2-TM6SF2**

| CILP2-TM6SF2 | Glu167Lys (TM6SF2) | rs10401969 | Association signals for TM6SF2 Glu167Lys and the previously reported non-coding GWAS SNP at this locus are indistinguishable from each other in reciprocal conditional analysis. TM6SF2 is the probable effector transcript for the non-coding GWAS signal at this locus, with the effect mediated through Glu167Lys. |
|--------------|-------------------|-----------|--------------------------------------------------|
| Coding       | 0.82             | 1.9x10^{-6} | 0.52                                             |
| Coding       | 0.83             | 0.62       | 4.2x10^{-6}                                      |

**GRB14-COB1L1**

| GRB14-COB1L1 | Asn939Gly (GRB14) | rs13389219 | Association signals for COB1L1 Asn939Gly and the previously reported non-coding GWAS SNP at this locus are partially correlated. The association signal for the GWAS signal is not entirely extinguished in reciprocal conditional analysis. COB1L1 is a candidate effector transcript for the GWAS signal at this locus. |
|--------------|------------------|-----------|--------------------------------------------------|
| Coding       | 0.12             | 4.7x10^{-10} | 3.0x10^{-6}                                      |
|Coding        | 0.39             | 7.0x10^{-6} | 1.9x10^{-6}                                      |

**Coding variant associations characterized in 44,414 samples in GoT2D genome-wide imputed meta-analysis**

| THADA       | Cys1605Tyr | rs10030174 | Association signals THADA Cys1605Tyr and the GWAS SNP are partially correlated. The association signal for the GWAS SNP is not entirely extinguished in reciprocal conditional analysis. THADA is a candidate effector transcript for the GWAS signal at this locus. |
|-------------|------------|-----------|--------------------------------------------------|
| Coding      | 0.10       | 0.00035   | 0.92                                             |
| Coding      | 0.10       | 0.0063    | 5.7x10^{-6}                                     |

| RREB1       | Asp1171Asn | rs9502570 | The association signals of RREB1 Asp1171Asn and the GWAS SNP at this locus are distinct. The association signal is not extinguished in reciprocal conditional analysis. Previous GWAS signal is not mediated through RREB1 Asp1171Asn. RREB1 Asp1171Asn represents a distinct association signal at this locus. |
|-------------|------------|-----------|--------------------------------------------------|
| Coding      | 0.11       | 0.0018    | 0.0017                                           |
| Coding      | 0.28       | 0.0037    | 0.0042                                           |

For each locus, significantly associated SNVs are presented. Unconditional $P$-values are given in italics, and conditional $P$-values are shown for each pair of SNVs ($P$-values are for SNVs in the Variant column, with SNVs listed in header included as covariates in association analysis). The IRS1 and PPARG non-coding associations were characterized using exact conditional analysis in 38,738 samples from the GoT2D genome-wide imputed meta-analysis. Conditional analysis for coding variant associations was, for most loci, restricted to the exome array genotypes (28,305 cases, 51,549 controls). At THADA and RREB1, neither the non-coding lead GWAS SNVs nor close proxies were typed on the exome array, so genome-wide imputed conditional analyses were undertaken using GCTA in 44,414 samples from the GoT2D genome-wide imputed meta-analysis (see Methods). For several of these loci, unconditional association $P$-values for these loci do not reach genome-wide significance as sample sizes are smaller. At the GPSM1 locus, the previously reported GWAS SNP was not available on exome array and too poorly imputed in the GoT2D meta-analysis to allow meaningful inference. *Conditional analysis was performed once for rs78124264 with all three previously known GWAS variants included as covariates. **Non-coding GWAS lead variant. n.d., not determined.
Extended Data Table 4 | Testing for synthetic associations across GWAS-identified T2D loci

| Gene     | Index SNV signal before inclusion of missense variants | Synthetic association by missense variants | Synthetic association by all low-frequency and rare variants across 5Mb region |
|----------|-------------------------------------------------------|------------------------------------------|--------------------------------------------------------------------------------|
|          | OR [95% interval] | p-value | OR [95% interval] | p-value | Relative likelihood of LF model | Best LF Variant | MAF [95% interval] | p-value | n<sub>1</sub> | n<sub>2</sub> |
| TCF7L2   | 1.75 [1.54-1.99] | 2.80×10<sup>-10</sup> | 6 | 1.73 [1.52-1.97] | 2.33×10<sup>-17</sup> | 1.8×10<sup>-17</sup> | 10:114787948 1.6% | 1.72 [1.51-1.95] | 1.62×10<sup>-10</sup> | >50 | 35 |
| ADCY5    | 0.69 [0.60-0.79] | 1.12×10<sup>-7</sup> | 13 | 0.70 [0.61-0.81] | 9.00×10<sup>-7</sup> | 9.7×10<sup>-8</sup> | 3:123096056 2.5% | 0.71 [0.61-0.82] | 3.04×10<sup>-4</sup> | 13 | 6 |
| IRS1     | 0.84 [0.75-0.86] | 2.80×10<sup>-4</sup> | 5 | 0.87 [0.69-0.86] | 4.30×10<sup>-9</sup> | 4.5×10<sup>-7</sup> | 2:226993370 1.7% | 0.78 [0.70-0.88] | 2.19×10<sup>-8</sup> | 12 | 6 |
| KCNO1    | 0.78 [0.70-0.87] | 1.22×10<sup>-3</sup> | >50 | 0.84 [0.75-0.86] | 2.07×10<sup>-3</sup> | 1.0×10<sup>-7</sup> | 11:2828279 4.7% | 0.81 [0.71-0.91] | 3.19×10<sup>-4</sup> | 16 | 6 |
| CDC123CAMK1D | 1.33 [1.17-1.52] | 1.19×10<sup>-3</sup> | 4 | 1.30 [1.13-1.50] | 2.06×10<sup>-4</sup> | 7.1×10<sup>-5</sup> | 10:12325477 3.8% | 1.29 [1.12-1.48] | 3.03×10<sup>-4</sup> | 10 | 5 |
| CDKN2A-CDKN2B | 1.28 [1.14-1.45] | 4.52×10<sup>-3</sup> | 4 | 1.27 [1.13-1.45] | 9.28×10<sup>-9</sup> | 4.3×10<sup>-9</sup> | 9:22133773 3.5% | 1.25 [1.10-1.41] | 5.98×10<sup>-4</sup> | 22 | 7 |
| IGF2BP2  | 1.25 [1.11-1.41] | 1.65×10<sup>-4</sup> | 14 | 1.21 [1.07-1.36] | 2.12×10<sup>-3</sup> | 3.0×10<sup>-4</sup> | 3:185550500 4.1% | 1.20 [1.07-1.36] | 2.91×10<sup>-5</sup> | 8 | 3 |
| KLRDC5   | 0.76 [0.66-0.86] | 2.19×10<sup>-4</sup> | 3 | 0.77 [0.66-0.86] | 4.45×10<sup>-4</sup> | 1.2×10<sup>-3</sup> | 12:27883062 2.0% | 0.80 [0.68-0.92] | 3.04×10<sup>-3</sup> | 10 | 4 |
| SLC30A8  | 0.81 [0.72-0.91] | 2.95×10<sup>-4</sup> | 2 | 0.81 [0.72-0.91] | 3.73×10<sup>-4</sup> | 0.02 | 8:117856024 2.2% | 0.83 [0.73-0.93] | 1.23×10<sup>-3</sup> | 17 | 6 |
| CDKAL1   | 1.28 [1.11-1.48] | 6.05×10<sup>-4</sup> | 1 | 1.28 [1.11-1.48] | 7.57×10<sup>-4</sup> | 0.007 | 6:20718780 2.8% | 1.23 [1.06-1.43] | 7.71×10<sup>-3</sup> | 9 | 3 |

Gene names refer to protein-coding transcript(s) closest to the index SNV. Reported index SNVs are the previously reported GWAS variants (in European populations) with the strongest association signal in the GoT2D sequencing data (n = 2,657). Relative likelihoods are based on causal models with only the chosen low-frequency and rare missense variants, relative to models with only the GWAS index SNV, assessed using the Akaike Information content (AIC) of each regression model, calculated as exp[(AICindex – AIClow-frequency or rare)/2]. n<sub>1</sub>, number of low-frequency or rare variants required for the residual odds ratio at the GWAS index SNV, after joint conditioning on the low-frequency and rare variants, to switch direction of effect. n<sub>2</sub>, number of low-frequency or rare variants required for the association P value remaining at the GWAS index SNV, after joint conditioning on the low-frequency and rare variants, to exceed 0.05.