nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0241-6

Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps

We expanded GWAS discovery for type 2 diabetes (T2D) by combining data from 898,130 European-descent individuals (9% cases), after imputation to high-density reference panels. With these data, we (i) extend the inventory of T2D-risk variants (243 loci, 135 newly implicated in T2D predisposition, comprising 403 distinct association signals); (ii) enrich discovery of lower-frequency risk alleles (80 index variants with minor allele frequency <5%, 14 with estimated allelic odds ratio >2); (iii) substantially improve fine-mapping of causal variants (at 51 signals, one variant accounted for >80% posterior probability of association (PPA)); (iv) extend fine-mapping through integration of tissue-specific epigenomic information (islet regulatory annotations extend the number of variants with PPA >80% to 73); (v) highlight validated therapeutic targets (18 genes with associations attributable to coding variants); and (vi) demonstrate enhanced potential for clinical translation (genome-wide chip heritability explains 18% of T2D risk; individuals in the extremes of a T2D polygenic risk score differ more than ninefold in prevalence).

Array-based genome-wide association studies (GWAS) have identified ~140 loci influencing the risk of T2D[1–3]. Follow-up of these genetic discoveries has been compromised by the incomplete coverage of the most frequently used genotyping arrays, the imperfect performance of the reference panels available for imputation, extensive local linkage disequilibrium (LD), and inadequate sample sizes. These factors together have limited the power to detect low-frequency alleles with population-scale effects, to deliver clinically relevant risk prediction, and to define molecular mechanisms involved in disease predisposition. Here, we address the limitations of previous studies by combining GWAS from ~900,000 Europeans with dense, high-quality imputation, producing the most comprehensive view to date of the genetic contribution to T2D with respect to locus discovery, causal-variant resolution, and mechanistic insight.

Results

Study overview. We combined data from 32 GWAS, including 74,124 T2D cases and 824,006 controls of European ancestry. The effective sample size (N_{eff}) of 231,436 represents a 3.2-fold increase in N_{eff} relative to the largest previous genome-wide study of T2D risk in Europeans[1]. After harmonized quality control, 31 of the 32 GWAS were imputed using 64,976 whole-genome-sequenced haplotypes from the Haplotype Reference Consortium (HRC)[4]: the exception was the deCODE GWAS, which was imputed with a population-specific reference panel of 30,440 Icelandic haplotypes[5] (Methods and Supplementary Table 1). We conducted T2D-association analyses with and without adjustment for body-mass index (BMI).

Discovery of novel loci for T2D susceptibility. We tested for T2D association with ~27 million variants passing quality-control filters, ~21 million of which had a minor allele frequency (MAF) <5%. Our meta-analysis identified variants at 231 loci reaching genome-wide significance ($P < 5 \times 10^{-8}$) in the BMI-unadjusted analysis (N_{eff} 231,436) and 152 in the smaller (N_{eff} 157,401) BMI-adjusted analysis. Of the 243 loci identified across these two analyses, 135 mapped outside regions previously implicated in T2D risk (Methods, Fig. 1 and Supplementary Table 2).

Among samples not included in previous discovery efforts (42,734 cases and 497,261 controls), we replicated associations (directionally consistent, $P < 0.05$) at 126 of 140 previously reported T2D loci, including all 106 regions first discovered in European-only or transancestry efforts[3,6–8] and 20 initially reported in studies of non-European individuals[9,10]. The 14 loci not replicated were all first identified in non-European-ancestry samples: at five, the reported lead variant had MAF <1% in Europeans.

Multiple association signals at T2D-susceptibility loci. Across the 243 associated loci, we identified 160 additional signals at 'locus-wide' significance ($P < 10^{-5}$; Methods), 110 of which were within previously reported T2D loci. Overall, we observed one signal at 151 loci, and two to ten signals at the remaining 92 loci (Supplementary Table 2), for a total of 403 distinct T2D-association signals.

We observed the first evidence for multiple signals at the *TCF7L2* locus. In addition to rs7903146, the largest-effect common-variant signal for T2D in Europeans, we detected seven secondary signals, each represented by noncoding index variants (0.5% < MAF < 47.6%, 1.05 < odds ratio (OR) < 1.36).

In the ~1-Mb telomeric region of chromosome 11 that encompasses the (previously annotated) *INS–IGF2* and *KCNQ1* loci, we detected 15 distinct signals (0.15% < MAF < 42.8%, 1.03 < OR < 1.68). This multiplicity of signals in a region notable for complex imprinting effects, and several strong biological candidates (*INS*, *IGF2*, *KCNQ1*, and *CDKN1C*), illustrates a previously unrecognized degree of complexity in the risk-variant architecture at this locus.

The effects of BMI and sex. At most T2D loci, there were only minimal differences in the estimated T2D effect sizes between BMI-adjusted and unadjusted models (Methods and Fig. 2). However, at index SNPs for 41 signals (mapping to 21 known and 16 novel loci), we observed significant differences in effect sizes between BMI-adjusted and unadjusted analyses ($P_{diff} < 0.00012$, corrected for 403

Full lists of authors and affiliations appear at the end of the paper.

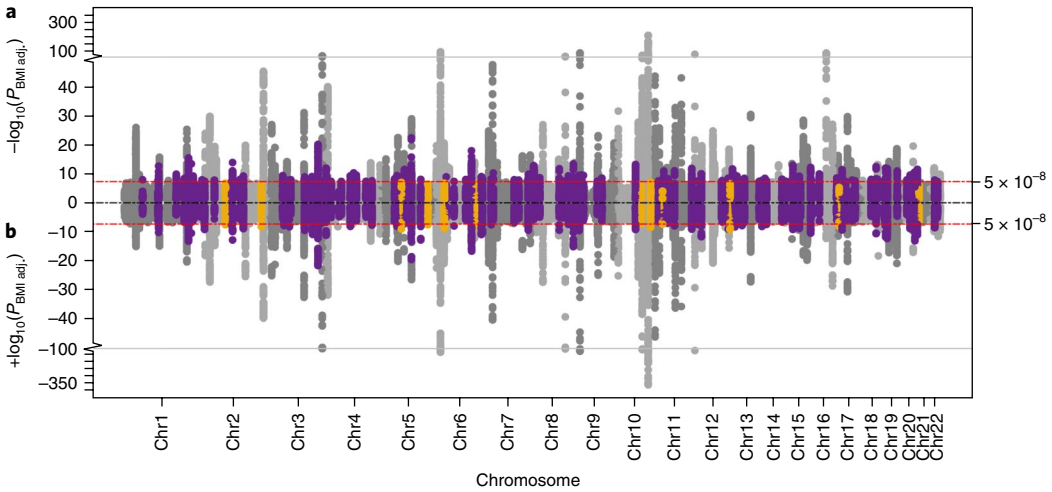

Fig. 1 | Manhattan plots of the sex-combined BMI-unadjusted and BMI-adjusted meta-analysis for T2D. a, Manhattan plot (top) of genome-wide-association results for T2D without BMI adjustment (BMI adj.) from meta-analysis of up to 71,124 cases and 824,006 controls. The $-\log_{10}$-transformed two-tailed *P* value for each SNP obtained from inverse-variance-weighted fixed-effects meta-analysis (*y* axis) is plotted against the genomic position (NCBI Build 37; *x* axis). Novel association signals that reached genome-wide significance ($P<5\times10^{-8}$) are shown in purple. **b**, Manhattan plot (bottom) of genome-wide-association results for T2D with BMI adjustment from meta-analysis of up to 50,409 cases and 523,897 controls. Novel association signals that reached genome-wide significance ($P<5\times10^{-8}$) only in the BMI-unadjusted analysis are shown in orange.

variants; Methods, Supplementary Table 3 and Fig. 2). This effect-size heterogeneity followed two distinct patterns. At 26 signals, including index variants for signals at the *FTO*, *MC4R*, *TMEM18*, *SEC16B*, and *GNPDA2* loci, BMI adjustment produced marked attenuation of associations detected in unadjusted analysis. These signals displayed positive correlations between BMI and T2D effect sizes and represented T2D-risk effects driven primarily by adiposity. The other 15 signals were more strongly associated in the BMI-adjusted analysis and reflected a mixture of associations, some with a marked effect on insulin secretion (for example, *TCF7L2*, *ARAP1*, and *JAZF1*), and others likely to influence T2D risk through a decreased capacity for fat storage in peripheral adipose tissue[11] (for example, *GRB14*, *PPARG*, *HMGA1*, and *ZNF664*).

In a comparative analysis of T2D effects in males (41,846 cases and 383,767 controls) and females (30,053 T2D cases and 434,336 controls; Methods)[12], only one of the 403 T2D signals showed significant ($P_{diff}<0.00012$) differences in effect size (rs2925979 near *CMIP*, female OR = 1.09, male OR = 1.03, $P_{diff}=8.3\times10^{-6}$; Supplementary Fig. 1 and Supplementary Table 4). We observed nominally significant differences at several other loci, including *KLF14* (rs1562396, female OR = 1.09, male OR = 1.04, $P_{diff}=0.00048$), at which there is additional corroboration for sex-specific effects[13,14], thus indicating that additional examples of sex-differentiated signals are likely to be found in larger samples.

Fine-mapping variants driving T2D-association signals. Previous efforts to fine-map causal variants within T2D loci have been hampered by both biological (extensive LD) and technical (diverse genotyping scaffolds or incomplete reference panels) factors. We sought to establish the extent to which the combination of increased sample size, an enlarged reference panel, and harmonized variant quality control would enhance fine-mapping resolution. We were able to undertake fine-mapping for 380 of the 403 distinct T2D-association signals, after conditional decomposition of loci with multiple signals (Methods). For each, we constructed credible sets that collectively accounted for ≥99% of the PPA (Methods)[15]. These credible sets included a median of 42 variants (range 1–3,997; Supplementary Fig. 2) and spanned a median of 116 kb (range 1 bp–995 kb). At 51 signals, involving 44 loci (18 novel), the most strongly associated variant accounted for >80% PPA (Fig. 3 and Supplementary Table 5). At 18 signals, the credible set included a single variant (PPA >99%).

We explored the fine-mapping resolution at 83 distinct signals for which detection in both studies allowed us to compare 99%-credible sets from the HRC-based analysis with those constructed in a subset of these T2D GWAS imputed by the 1000 Genomes Project (1000G) multiancestral reference panel[1] (26,676 T2D cases; 132,532 controls of European ancestry, N_{eff} 72,143). Although the former includes 2.3-fold more variants genome wide than the latter, the HRC-imputed analysis resulted in smaller credible sets. The median number of variants at these 83 signals decreased from 59 to 10, and the interval length decreased from 60.3 kb to 19.2 kb. At 79 of 83 signals, HRC-based credible sets were either smaller than those generated from 1000G or unchanged (Fig. 4 and Supplementary Table 6).

This improved resolution probably reflects the combination of (i) increased N_{eff}; (ii) improved imputation quality, especially for lower-frequency variants[4]; and (iii) more effective, harmonized, quality control across contributing studies (Methods). To estimate the contribution to fine-mapping resolution attributable to the increase in N_{eff} (the other factors are more difficult to tease apart), we constructed 99%-credible sets by downscaling the HRC imputation to a subset of 19 studies (31,387 cases and 326,742 controls, N_{eff} 92,960) that contributed to both 1000G and HRC-based analyses. Among 41 single signal loci with $P<1\times10^{-5}$ in this downscaled meta-analysis, estimates of the credible-set size (median 66) and interval (median 196 kb) indicated that the improvements in causal-variant resolution derived mostly from increased sample size.

The HRC panel provides excellent coverage of all but very rare single-nucleotide variants. However, one HRC limitation is the absence of indels, which constitute 4% of total variants in the phase 3 1000G reference panel[16]. We considered the 245,207 indels from the European subset of the 1000G panel that mapped within 500 kb of the index variants at the 380 fine-mapped signals: these accounted for 2.8% of variants across the 380 Mb of sequence. Only 1% of these were in even moderate LD ($r^2>0.5$) with index variants for each T2D-association signal, thus indicating that indel omission probably had a limited effect on our estimates of credible set size.

The contribution of lower-frequency variants. The limited yield of low-frequency and rare-variant signals in previous T2D GWAS

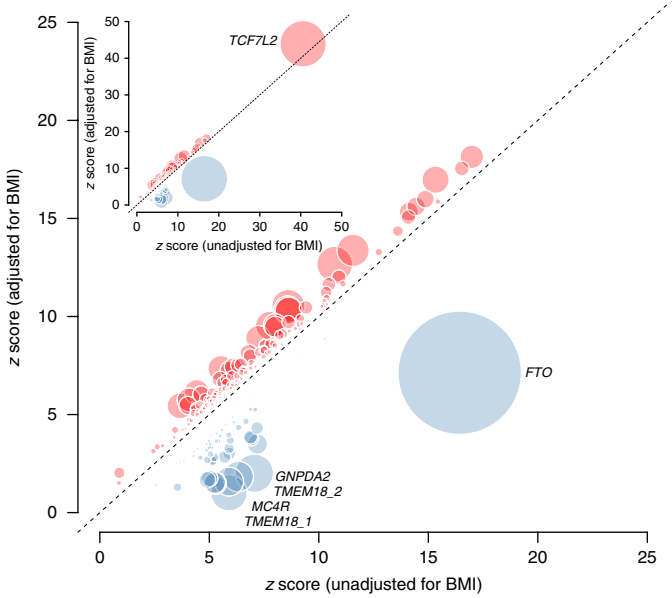

Fig. 2 | Comparison of estimated T2D effect sizes between BMI-adjusted and unadjusted models. *z* scores for each of the 403 distinct signals from BMI-unadjusted analysis (50,791 cases and 526,121 controls; *x* axis) are plotted against the *z* scores from the BMI-adjusted analysis (50,402 cases and 523,888 controls; *y* axis). Variants displaying higher T2D effect sizes in BMI-adjusted analysis are shown in red, and variants with higher T2D effect sizes in BMI-unadjusted analysis are shown in blue. Circle diameter is proportional to $-\log_{10}$ heterogeneity *P* value (two-tailed test). Inset presents the same plot with *TCF7L2* variant included.

placed an upper bound on their individual and collective contributions to disease risk[17]. The present analysis, with a larger sample size and improved imputation, provided greater power in this regard, identifying 56 low-frequency and 24 rare T2D-associated variants across 60 loci (Fig. 5). Six of these 80 signals mapped within known T2D loci, and five reconfirmed earlier observations (Supplementary Table 2).

The allelic OR for low-frequency and rare variants ranged from 1.08 to 8.05 (including 14 with estimated allelic OR >2; at each, the minor allele conferred T2D risk), compared with 1.03–1.37 for common variants (Fig. 5). The 80 lower-frequency risk variants cumulatively explained 1.1% of phenotypic variance in T2D, compared with 16.3% attributable to the 323 common-variant signals (Methods). Extrapolation beyond these discovered signals to estimate the full contribution of lower-frequency variants to T2D risk is intrinsically difficult, given the combination of effect-size overestimation and limited power to capture lower-frequency variants of lesser effect. Nonetheless, these data are consistent with recently proposed models for the genetic architecture of T2D based on GWAS and sequencing data[17]. Notwithstanding, the identification of lower-frequency variants with modest-to-large effects can provide valuable biological inference. Below, we briefly describe some of these signals.

We observed a mix of common-variant and low-frequency-variant signals around *NEUROG3*, including T2D risk attributable to the minor alleles at rs41277236 (p.Gly167Arg, MAF = 4.3%, OR = 1.09, $P = 1.5 \times 10^{-6}$) and rs549498088 (noncoding, MAF = 0.60%, OR = 1.56, $P = 4.7 \times 10^{-7}$). *NEUROG3* encodes the neurogenin-3 transcription factor, which has been implicated in pancreatic islet and enteroendocrine cell development[18]. Rare homozygous, hypomorphic missense mutations in *NEUROG3* (nonoverlapping with those that we detected) are a cause of childhood-onset diabetes associated with severe malabsorptive diarrhea[19]. The age of T2D diagnosis among carriers of these low-frequency T2D-risk alleles was, in the UK Biobank, similar to that among noncarriers (rs41277236: 52.3 versus 52.7 years, $P = 0.21$; rs549498088: 51.1 versus 52.7 years; $P = 0.49$), indicating a spectrum of phenotypes associated with *NEUROG3* variants that extends to typical T2D. In the UK Biobank, T2D-risk alleles at *NEUROG3* were associated with phenotypes recapitulating the gastrointestinal component of the neonatal syndrome (including 'obstruction of bile duct' (OR = 1.29; $P = 0.023$), 'gastrointestinal complications' (OR = 1.79; $P = 0.024$), and 'functional digestive disorders' (OR = 1.06; $P = 0.027$)).

We detected two previously unreported rare alleles with large ORs. The first was intronic to *DENND2C* (rs184660829, MAF = 0.020%, OR = 8.1, $P = 2.5 \times 10^{-8}$). In exploratory analyses within UK Biobank, the T2D-risk allele was associated with 'lower gastrointestinal congenital anomalies' (OR = 17.3 $P = 0.00047$). The second allele mapped near *KIF2B* (rs569511541, MAF = 0.020%, OR = 7.6, $P = 1.5 \times 10^{-8}$) and was also associated with 'congenital anomalies of endocrine gland' (OR = 30.8; $P = 0.00015$), 'disease of pancreas' (OR = 5.9; $P = 0.0017$), and 'hypokalemia' (OR = 6.9; $P = 0.0046$). Both sites are present in the Genome Aggregation Database[20] and met quality-control criteria in our data (average imputation quality >0.7; association signal visible in multiple studies), but their precise contribution to T2D risk requires further validation.

Causal coding variants. We next considered the 51 signals (of 380) for which fine-mapping strongly implicated (PPA >80%) a single causal variant. Eight of these were missense coding variants, six of which were within established T2D-associated regions (Supplementary Table 7). With the exception of p.Cys130Arg at *APOE* (MAF = 15.4%), all have been implicated as causal for T2D: p.Ser539Trp in *PAM* (MAF = 0.83%); p.Thr139Ile in *HNF4A* (MAF = 3.5%); p.Asp1171Asn in *RREB1* (MAF = 11.3%); p.Ala146Val in *HNF1A* (MAF = 2.9%); and p.Pro446Leu in *GCKR* (MAF = 39.3%)[3]. Coding-variant associations at *PATJ* (p.Gly157Val; 9.5% MAF) and *CDKN1B* (p.Val109Gly; 23.5% MAF) are novel and highlight these genes as playing direct roles in T2D risk. *PATJ* is highly expressed in the brain[21] and encodes Pals1-associated tight junction component, a protein with multiple PDZ domains, which mediate protein–protein interactions. Associations for this variant indicated a central mechanism of action: the T2D-risk allele was associated with obesity in the UK Biobank (OR = 1.11; $P = 3.8 \times 10^{-5}$), and the T2D-association signal was attenuated in BMI-adjusted analysis ($P_{diff} = 9.3 \times 10^{-10}$). *CDKN1B* encodes a cyclin-dependent-kinase inhibitor, and deletion of this gene in mice ameliorates hyperglycemia by increasing islet mass and maintaining compensatory hyperinsulinemia[22]. There were four further signals (at *ANKH*, *POC5*, *NEUROG3*, and *ZNF771*) at which a single missense variant accounted for most (>50%) of the PPA (Supplementary Table 7).

Integration of regulatory annotations to support fine-mapping. Of the 51 variants with PPA >80%, 43 mapped to regulatory sequence: 12 of these were low frequency or rare, including variants near *ANKH*, *CCND2*, and *WDR72*. To characterize the regulatory effects of these 51 variants, we overlaid them onto chromatin-state maps from T2D-relevant tissues (islets, liver, adipose, and skeletal muscle[23–25]) and transcription-factor-binding sites[23,24]. Twenty-eight mapped to islet enhancer or promoter elements; for 14, these chromatin states were islet specific (Supplementary Table 8 and Supplementary Fig. 3). These data recapitulate previous findings implicating islet regulatory mechanisms at the *CDC123–CAMKD1* (rs11257655) and *MTNRB1* (rs10830963)[25–27] loci, and indicate that similar molecular mechanisms operate at signals for several other known T2D loci, including *IGF2BP2*, *ANK1*, *GLIS3*, *CDKN2B*, *KCNQ1*, *CCND2*, and *BCL2A*. Novel T2D signals near *ABCB10*, *FAM49A*, *LRFN2*, *CRHR2*, and *CASC11*

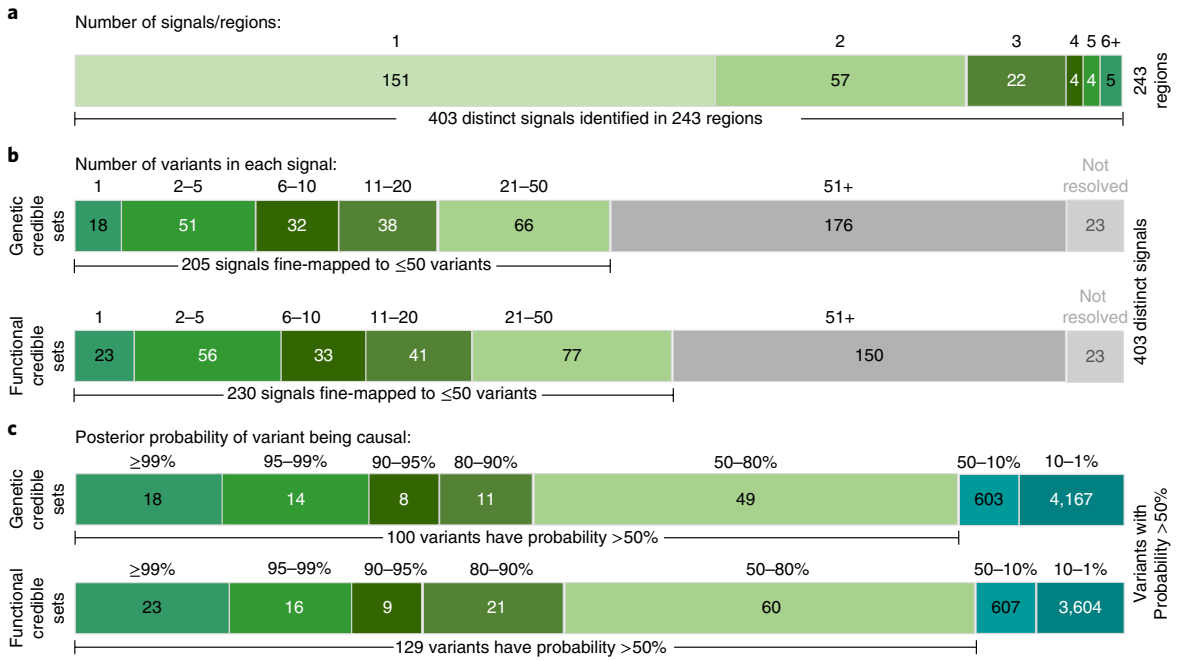

Fig. 3 | Summary of fine-mapped associations. a, Distinct association signals. A single signal at 151 loci, and two to ten signals at 92. **b**, Number of variants in genetic and functional 99%-credible sets. Eighteen and 23 signals were mapped to a single variant in genetic and functional credible sets, respectively. **c**, Distribution of the PPA of the variants in credible sets. Four of the 51 variants with PPA >80% in the genetic credible sets have lower PPAs in the functional credible set, thus giving a total of 73 variants with PPA >80% in either.

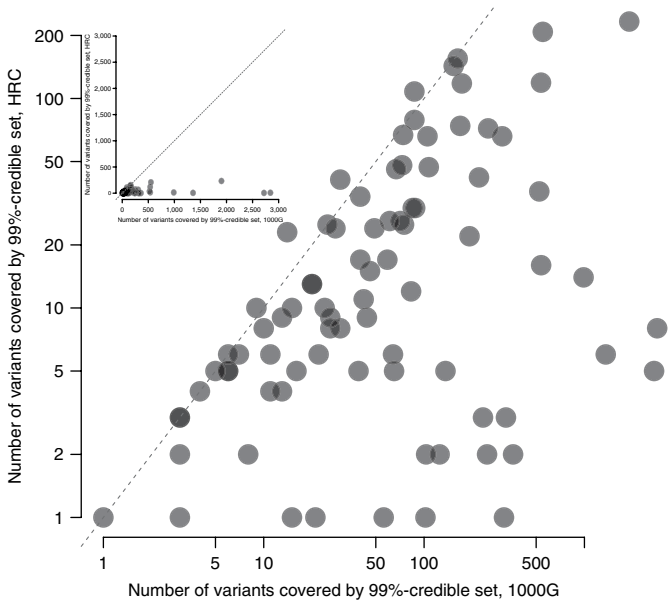

Fig. 4 | Comparison of fine-mapping resolution at 83 distinct signals. The number of variants included in the 99%-credible set for each of the 83 distinct signals constructed through meta-analysis of GWAS data imputed from the 1000G multiancestry reference panel (26,676 T2D cases and 132,532 controls) (*x* axis; logarithmic scale) is plotted against those (*y* axis; logarithmic scale) derived by HRC-based imputation (74,124 T2D cases and 824,006 controls). The inset presents the same plot but with linear scales.

also overlapped islet-specific enhancers or promoters. High-PPA variants (i.e., those with PPA >80%) at 13, 10, and 7 signals overlapped enhancers or promoters in adipose, skeletal muscle, and liver tissues, respectively. All but four of these were also enhancers or promoters in islets: one signal (near *GLI2*) mapped to an adipose-specific enhancer, another (near *WDR72*) mapped to a liver-specific enhancer, and two (near *PTGFRN* and *TSC22D2*) mapped to enhancers in both adipose tissue and skeletal muscle.

We next evaluated whether the integration of genome-wide-regulatory annotation data could refine the mapping resolution at those loci where genetic fine-mapping was less precise[25]. We focused on regulatory annotations from human islets because (i) most established T2D-risk variants are considered, given observed patterns of association with continuous metabolic traits, to act through primary effects on beta-cell function[3,28,29]; (ii) the strongest signal for regulatory enrichment at T2D-association signals involves islet-specific regulatory elements[23,26], a view supported by the annotation overlaps of the high-PPA variants described above and by enrichment analyses that we performed using epigenomic annotations from islets, fat, muscle, and liver[24] (Supplementary Fig. 4); and (iii) we had access to high-resolution epigenomic and chromatin-state annotation maps for human islets combining available data on histone modifications, transcription-factor binding, chromatin accessibility, and whole-genome methylation[25].

Using the hierarchical modeling approach fGWAS[30], we observed strong (1.9- to 8.2-fold), significant (95% confidence not overlapping one), genome-wide enrichment of T2D-associated variation with respect to multiple islet enhancer and promoter states, as well as coding sequence (with concomitant depletion of heterochromatin states; Methods and Supplementary Fig. 5). We used the parameter estimates from the joint annotation model (which retained islet enhancers, promoters, and coding sequence, among other annotations; Methods and Supplementary Fig. 5) as priors to redefine 99%-credible sets for the 380 distinct T2D-association signals amenable to fine-mapping. We circumvented the default assumption in fGWAS of a single casual variant per locus by conducting these analyses on conditionally decomposed data (noting that this procedure still allowed for the possibility that the association at each conditional signal might be distributed across multiple variants on a risk haplotype; Methods).

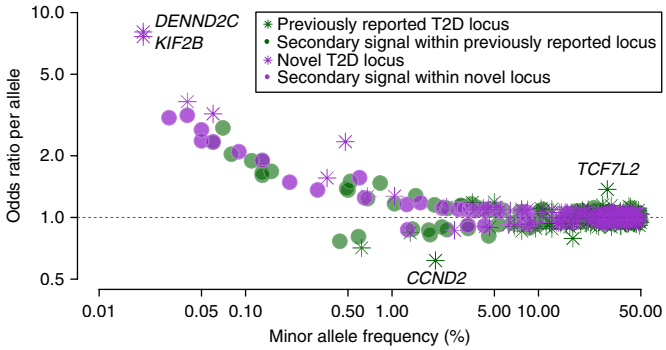

Fig. 5 | The relationship between effect size and MAF. Conditional- and joint-analysis effect size (*y* axis) and MAF (*x* axis) for 403 conditionally independent SNPs. Previously reported T2D-associated variants are shown in green, and novel variants are shown in purple. Stars and circles represent the 'strongest regional lead at a locus' and 'lead variants for secondary signals', respectively.

As expected, this integrated fine-mapping analysis boosted PPA for variants overlapping enriched annotations (Fig. 6). The median 99%-credible-set size declined from 42 to 32, the credible intervals declined from 116 kb to 100 kb, and the maximum variant PPA per signal increased by a median of 21%. The number of signals at which the lead-variant PPA exceeded 80% increased from 51 to 73, and there were dramatic improvements at some (for example, at *GNG4*, for which the PPA for rs291367 rose from 24.0% to 84.2%; Fig. 3).

These annotation-supported analyses highlighted seven additional loci (beyond the 12 determined from genetic evidence alone) where most (>50%) of the PPA was based on a coding variant (Supplementary Table 7). Four were novel: *QSER1* (p.Arg1101Cys; MAF = 4.3%), *SCD5* (p.Glu197Gln, MAF = 33.8%), *IRS2* (p.Gly1057Asp, MAF = 34.0%), and *MRPS30* (p.Glu128Gln = MAF 2.8%).

In our recent study of exome-array genotypes, we demonstrated that, for one-third of loci with coding-variant associations, a causal role could be excluded after information on local LD and annotation enrichment was incorporated[3]. For all 19 coding-variant signals (at 18 loci) described in this study, the results of the present analyses (based on genome-wide data for both discovery and fine-mapping) were consistent with a causal role. These analyses therefore provide additional examples of human validated targets[31]. The value of these targets as leads for therapeutic development will ultimately depend not only on their effects on T2D phenotypes but also on the consequences of perturbation on other traits, including coronary artery disease (CAD). Among the 19 T2D-associated coding variants, nine were also nominally associated ($P < 0.05$) with CAD[32]: for three variants (*APOE*, *GCKR*, and *RREB1*), opposing effects on T2D and CAD predisposition render them less attractive targets (Supplementary Table 7).

Next, we concentrated on noncoding-variant signals. In the annotation-informed analysis, we identified 15 additional signals (beyond the 43 noncoding signals described above) for which the lead-variant PPA exceeded 80% (Supplementary Table 8). These signals overlapped active islet regulatory sites including strong enhancers (for example, at *TCF7L2*, *HNF4A*, *ANKH*, *RNF6*, and *ZBED3*), active promoters (*EYA2*), weak enhancers (*ADSCL2*, *ADCY5*, *CDKN2B*, and *TBCE*), and weak promoters (*DGKB*). For many signals, the data (for example, associations with continuous metabolic traits[3,28,29] and cis expression quantitative trait locus (eQTL) data[33]; Supplementary Table 8) were consistent with a role in islet function. In contrast, for six signals, including three that are likely, on physiological grounds, to act at least partly through effects on islets, we observed decreases (10% to 76%) in the lead-variant PPA after islet-annotation-informed fGWAS (Supplementary Table 8). This decrease occurred when lead variants from the genetic fine-mapping overlapped with annotations depleted in the genome-wide model. Examples included variants at primary *CDKAL1* and secondary *KCNQ1* and *INS–IGF2* signals, where the index-variant PPA decreased by 76% (rs7756992), 34% (rs2283164), and 22% (rs555759341), respectively. One possible explanation for these results is that for these T2D-association signals, the phenotypic effect on insulin secretion may be mediated through long-term consequences of regulatory effects during islet development, which are no longer reflected in the regulatory annotations seen in mature islets.

At many of these fine-mapped regulatory loci, the integrated data provided novel insights into disease mechanisms, three of which are highlighted below. At *ST6GAL1*, rs3887925 achieved PPA = 98.5% through genetic fine-mapping alone (99.3% in fGWAS), and overlapped with enhancers active in islet, as well as liver, adipose, and skeletal muscle tissues (Supplementary Fig. 6). However, the T2D-risk allele at rs3887925 was associated with an increase in *ST6GAL1* cis expression specific to islets[33] (Methods and Supplementary Table 8), in agreement with evidence of decreased insulin secretion in risk-allele carriers during provocative testing[34]. The candidate effector transcript *ST6GAL1* encodes β-galactoside α2,6-sialyltransferase-1, a key enzyme responsible for the biosynthesis of α 2,6-linked sialic acid in N-linked glycans. Altered glycosylation has the potential to affect multiple processes, and global perturbation of *ST6GAL1* has broad effects including, in *St6gal1*-knockout mice, increased body weight and visceral fat accumulation[35]. However, no equivalent association between rs3887925 and anthropometric and lipid phenotypes has been seen in human GWAS[14,36,37]. These results are consistent with the T2D predisposition attributable to rs3887925 being mediated through regulatory mechanisms restricted to the modulation of *ST6GAL1* expression in islets.

At *ANK1*, we observed three distinct association signals. The strongest causal-variant attribution was for the primary signal at rs13262861 (PPA = 97.3% on the basis of genetic data alone; 98.8% with fGWAS). This variant overlaps an islet promoter located 3′ to *ANK1* and 5′ to the transcription-factor-encoding *NKX6-3* (Supplementary Fig. 7). The T2D-risk allele at rs13262861 showed a directionally consistent association with in vivo measures of decreased insulin secretion[3,29,34] and a cis-eQTL for decreased *NKX6-3* expression in human islets (Supplementary Table 8). Members of the NKX6 family (including NKX6.3) have been implicated in islet development and function[38]. A recent study has highlighted the relationship between variants including rs515071 and rs508419 and the expression and splicing of *ANK1* in skeletal muscle[39]. However, in our meta-analysis, variants influencing *ANK1* splicing had a minimal effect on T2D risk (PPA <1% in all three conditionally decomposed signals (genetic fine-mapping only)). Collectively, these data indicate that the mechanism of T2D predisposition at this locus is probably mediated through decreased islet expression of *NKX6-3* rather than altered muscle expression of *ANK1*.

At *TCF7L2*, patterns of overlap with epigenomic annotations across the eight distinct T2D-association signals offered explanations for the diverse metabolic consequences of *TCF7L2* perturbation in humans and animal models[40] (Supplementary Table 9). The primary signal at rs7903146, long established as the largest common-variant effect for T2D in Europeans, overlapped an islet enhancer (boosting PPA from 59.2% to 97.1% in fGWAS), multiple islet-relevant transcription-factor-binding sites, and islet open chromatin[41], features consistent with the islet phenotype (deficiency in insulin secretion) evident in nondiabetic individuals[7] (Supplementary Fig. 8). However, among the seven secondary signals, the picture was more mixed. Of the four secondary signals mapped to fewer than ten credible-set variants, only rs144155527 rose to moderate PPA (68%) after islet-annotation-enriched fGWAS analysis. Other

credible-set variants mapped to adipose and liver enhancers, thus suggesting that their T2D-risk effects are mediated via modulation of *TCF7L2* expression in tissues relevant to insulin action.

Heritability estimates and polygenic-risk-score prediction. Using LD-score regression[42], and empirical estimates of population- and sample-level T2D prevalence, we estimated the chip heritability (on the liability scale) for T2D at 18% (23% in females and 17% in males; Supplementary Fig. 9), accounting for approximately half the median estimates of heritability derived from twin and family studies[43].

Identification of individuals at increased genetic risk for T2D may enhance screening strategies and allow for targeted prevention. Previous attempts to deploy genetic data for disease prediction have shown limited utility[44,45]. We used a revised BMI-unadjusted meta-analysis, generated from all samples other than the UK Biobank samples, to develop genome-wide polygenic risk scores (PRSs)[46], which we then applied to predict T2D status in the 18,197 cases and 423,697 controls from the UK Biobank (Europeans only; Methods)[46]. Maximal discrimination (area-under-the-curve *C* statistic of 66%, equivalent to that derived from BMI, age, and sex in the same sample) was obtained from a PRS of 136,795 variants ($r^2 > 0.6$, $P < 0.076$; Supplementary Fig. 10). Individuals in the top 2.5% of the PRS distribution were at 3.4-fold-increased risk (prevalence = 11.2%) compared with the median (prevalence = 3.3%), and at 9.4-fold-increased risk compared with the bottom 2.5% (prevalence = 1.2%). Low T2D prevalence in the UK Biobank reflected the age distribution of the cohort and preferential ascertainment of healthy individuals; however, similar prevalence ratios were observed in the subset of individuals >55 years of age at recruitment (14.2% versus 1.6%). If applied to the general UK population, an equivalent performance would equate to lifetime T2D risks of ~59.7% and ~6.7% for individuals from those extremes, on the basis of current UK general-population prevalence rates for individuals >55 years of age[47].

Defining relationships with other traits. To characterize genetic relationships with other biomedically relevant traits, we used LD-score regression[42] implemented in LDHub[48]. We tested 182 unique phenotypes after excluding those with low heritability estimates and repeated measures. Eighty-five traits demonstrated a significant (Bonferroni-corrected threshold $P < 0.00027$) genetic correlation with T2D (Supplementary Table 10 and Supplementary Fig. 11).

These results highlighted several interesting genetic correlations, linking increased T2D risk to sleeping behaviors (insomnia and excessive daytime sleeping), smoking (cigarettes smoked per day, and having ever versus never smoked), metabolites (glycoprotein acetyls, isoleucine, and valine), depressive symptoms, urinary albumin-to-creatinine ratio, and urate. T2D risk was negatively correlated with anorexia nervosa, intelligence, parents' ages at death, lung-function measures, education status/duration, age at menarche, and age of mother at first childbirth. Many of these relationships (including those related to intelligence, smoking behavior, age at menarche, and education status) were primarily mediated by the shared effects of BMI on both T2D and the correlated phenotype (Supplementary Fig. 12).

Discussion

This study demonstrates how substantial increases in sample size coupled to more accurate and comprehensive imputation can expand characterization of the genetic contribution to T2D risk. The number of significantly associated genomic regions doubled, and the harvesting of lower-frequency risk alleles, some with relatively large effects, increased. At many of these signals, fine-mapping resolution was substantially improved: we mapped 51 of 380 signals to single-variant resolution on the basis of genetic evidence alone and demonstrated that the integration of genomic annotations (here with a focus on the human islet epigenome) provided further specification of plausible causal variants. We highlight 18 genes as human-validated targets based on causal coding variants and provide novel insights into the biological mechanisms operating at several fine-mapped regulatory signals. These findings suggest mechanistic hypotheses that can now be targeted for large-scale empirical validation at the level of both variants (for example, through massively parallel reporter assays) and candidate effector genes (for example, through CRISPR screens in appropriate cellular models and manipulation in in vivo models). The present study was limited to individuals of European ancestry: integration of these data with large-scale GWAS data from other major ancestral groups (as is being pursued by the DIAMANTE consortium) should provide an additional boost to locus discovery and support further increases in causal-variant resolution, most obviously at loci where extensive LD within Europeans limits the resolution of fine-mapping.

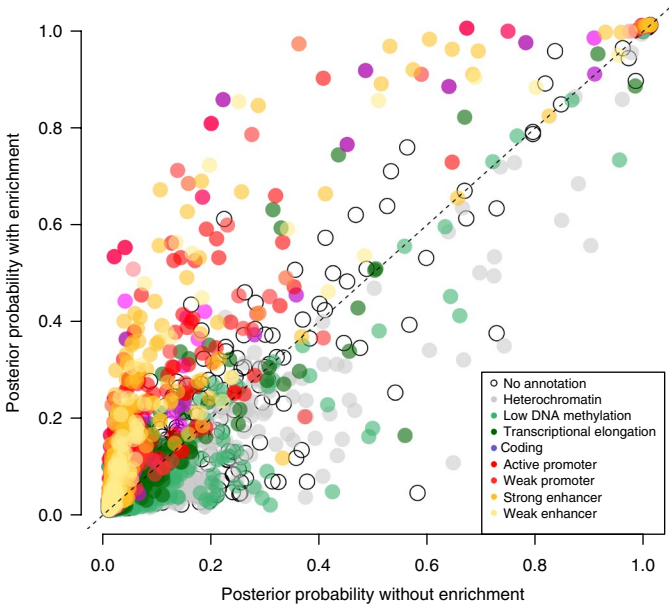

Fig. 6 | Comparison of PPA for each variant with and without incorporation enrichment information. PPA from genetic credible sets (*y* axis) and fGWAS analysis (*x* axis) for each variant included in the 99%-credible sets.

URLs

UK Biobank, http://www.ukbiobank.ac.uk/; MACH, http://csg.sph.umich.edu//abecasis/MaCH/; SHAPEIT, https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html; GTEx, http://www.gtexportal.org/home/; LocusZoom, http://locuszoom.sph.umich.edu/locuszoom/; 1000 Genomes Project, http://www.1000genomes.org/; HapMap project, http://hapmap.ncbi.nlm.nih.gov/; HRC, http://www.haplotype-reference-consortium.org/; GCTA, http://cnsgenomics.com/software/gcta/; LDSC, https://github.com/bulik/ldsc/; LD Hub, http://ldsc.broadinstitute.org/; Bedtools, http://bedtools.readthedocs.io/en/latest/; DIAGRAM Consortium, http://diagram-consortium.org/; fGWAS, https://github.com/joepickrell/fgwas/.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0241-6.

Received: 26 December 2017; Accepted: 10 August 2018;
Published online: 08 October 2018

References

1. Scott, R. A. et al. An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes* **66**, 2888–2902 (2017).
2. Zhao, W. et al. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat. Genet.* **49**, 1450–1457 (2017).
3. Mahajan, A. et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **50**, 559–571 (2018).
4. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
5. Jónsson, H. et al. Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci. Data* **4**, 170115 (2017).
6. Flannick, J. & Florez, J. C. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat. Rev. Genet.* **17**, 535–549 (2016).
7. Voight, B. F. et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
8. 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).
9. Kooner, J. S. et al. Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat. Genet.* **43**, 984–989 (2011).
10. Cho, Y. S. et al. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat. Genet.* **44**, 67–72 (2011).
11. Lotta, L. A. et al. Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance. *Nat. Genet.* **49**, 17–26 (2017).
12. Magi, R., Lindgren, C. M. & Morris, A. P. Meta-analysis of sex-specific genome-wide association studies. *Genet. Epidemiol.* **34**, 846–853 (2010).
13. Small, K. S. et al. Identification of an imprinted master trans regulator at the *KLF14* locus related to multiple metabolic phenotypes. *Nat. Genet.* **43**, 561–564 (2011).
14. Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
15. Maller, J. B. et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
16. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
17. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
18. Gradwohl, G., Dierich, A., LeMeur, M. & Guillemot, F. Neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. *Proc. Natl. Acad. Sci. USA* **97**, 1607–1611 (2000).
19. Rubio-Cabezas, O. et al. Permanent neonatal diabetes and enteric anendocrinosis associated with biallelic mutations in NEUROG3. *Diabetes* **60**, 1349–1353 (2011).
20. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
21. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
22. Uchida, T. et al. Deletion of *Cdkn1b* ameliorates hyperglycemia by maintaining compensatory hyperinsulinemia in diabetic mice. *Nat. Med.* **11**, 175–182 (2005).
23. Pasquali, L. et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143 (2014).
24. Varshney, A. et al. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc. Natl. Acad. Sci. USA* **114**, 2301–2306 (2017).
25. Thurner, M. et al. Integration of human pancreatic islet genomic data refines regulatory mechanisms at Type 2 diabetes susceptibility loci. *eLife* **7**, e31977 (2018).
26. Gaulton, K. J. et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).
27. Fogarty, M. P., Cannon, M. E., Vadlamudi, S., Gaulton, K. J. & Mohlke, K. L. Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMK1D type 2 diabetes GWAS locus. *PLoS. Genet.* **10**, e1004633 (2014).
28. Dimas, A. S. et al. Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes* **63**, 2158–2171 (2014).
29. Wood, A. R. et al. A genome-wide association study of IVGTT-based measures of first-phase insulin secretion refines the underlying physiology of type 2 diabetes variants. *Diabetes* **66**, 2296–2309 (2017).
30. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
31. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug. Discov.* **12**, 581–594 (2013).
32. van der Harst, P. & Verweij, N. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ. Res.* **122**, 433–443 (2018).
33. van de Bunt, M. et al. Transcript expression data from human islets links regulatory signals from genome-wide association studies for type 2 diabetes and glycemic traits to their downstream effectors. *PLoS. Genet.* **11**, e1005694 (2015).
34. Prokopenko, I. et al. A central role for GRB10 in regulation of islet function in man. *PLoS. Genet.* **10**, e1004235 (2014).
35. Kaburagi, T., Kizuka, Y., Kitazume, S. & Taniguchi, N. The inhibitory role of α2,6-sialylation in adipogenesis. *J. Biol. Chem.* **292**, 2278–2286 (2017).
36. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
37. Shungin, D. et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
38. Lizio, M. et al. Mapping mammalian cell-type-specific transcriptional regulatory networks using KD-CAGE and ChIP-seq data in the TC-YIK cell line. *Front. Genet.* **6**, 331 (2015).
39. Scott, L. J. et al. The genetic regulatory signature of type 2 diabetes in human skeletal muscle. *Nat. Commun.* **7**, 11764 (2016).
40. McCarthy, M. I., Rorsman, P. & Gloyn, A. L. TCF7L2 and diabetes: a tale of two tissues, and of two species. *Cell. Metab.* **17**, 157–159 (2013).
41. Gaulton, K. J. et al. A map of open chromatin in human pancreatic islets. *Nat. Genet.* **42**, 255–259 (2010).
42. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
43. Meigs, J. B., Cupples, L. A. & Wilson, P. W. Parental transmission of type 2 diabetes: the Framingham Offspring Study. *Diabetes* **49**, 2201–2207 (2000).
44. Meigs, J. B. et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N. Engl. J. Med.* **359**, 2208–2219 (2008).
45. Weedon, M. N. et al. Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. *PLoS. Med.* **3**, e374 (2006).
46. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2015).
47. Gatineau, M. et al. Adult obesity and type 2 diabetes (Public Health England, London, 2014). https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/338934/Adult_obesity_and_type_2_diabetes_.pdf.
48. Zheng, J. et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).

Acknowledgements

This work was supported primarily by the NIDDK as part of the Accelerating Medicines Partnership-T2D, funded by U01DK105535 (M.I.M.), U01DK062370 (M.B.), and U01DK078616 (J.M.) grants. Part of this work was conducted using the UK Biobank resource under application number 9161. A full list of acknowledgements appears in the Supplementary Note.

Author contributions

Project coordination: A. Mahajan, A.P.M., M.B., and M.I.M. Writing: A. Mahajan, D.T., A.P.M., M.B., and M.I.M. Core analyses: A. Mahajan, D.T., M.T., J.M.T., A.J.P., A.P.M., M.B., and M.I.M. DIAMANTE analysis group: A. Mahajan, J.E.B., D.W.B., J.C.C., Y.J.K., M.C.Y.N., L.E.P., X.S., W.Z., A.P.M., M.B., and M.I.M. Statistical analysis in individual studies: A. Mahajan, D.T., N.R.R., N.W.R., V.S., R.A.S., N.G., J.P.C., E.M.S., M.W., C. Sarnowski, J.N., S.T., C. Lecoeur, M.H.P., B.P.P., X.G., L.F.B., J.B.-J., M.C., K.L., C.-T.L., A.E.L., J'a.L., C. Schurmann, L.Y., G.T., and A.P.M. Genotyping and phenotyping: A. Mahajan, R.A.S., R.M., C.G., S.T., K.-U.E., K.F., S.L.R.K., F.K., I.N., C.M.B., C. Schurmann, E.P.B., I.B., C.C., G.D., I.F., V.G., M.I., M.E.J., S.L., A.L., V.L., V.M., A.D.M., G.N., N.S., A.S., D.R.W., S.S., E.P.B., S.H., C.H., J. Kriebel, T.M., A.P., B.T., A.D., A.K., G.R.A., C. Langenberg, N.J.W., A.P.M., M.B., and M.I.M. Islet annotations: M.T., J.M.T., A.J.B., V.N., A.L.G., and M.I.M. Individual study design and principal investigators: E.P.B., J.C.F., O.H.F., T.M.F., A.T.H., M.A.I., T.J., J. Kuusisto, C.M.L., K.L.M., J.S.P., K. Strauch, K.D.T., U.T., J.T., J.D., P.A.P., E.Z., R.J.F.L., P.F., E.I., L.L., L.G., M.L., F.S.C., J.W.J., C.N.A.P., H.G., A. Metspalu, A.D., A.K., G.R.A., J.B.M., J.I.R., J.M., O.P., T.H., C. Langenberg, N.J.W., K. Stefansson, A.P.M., M.B., and M.I.M.

Competing interests

J.C.F. has received consulting honoraria from Merck and from Boehringer-Ingelheim. O.H.F. works at ErasmusAGE, a center for aging research across the course of life, funded by Nestlé Nutrition (Nestec Ltd.), Metagenics Inc., and AXA. E.I. is a scientific advisor for Precision Wellness and Olink Proteomics for work unrelated to the present project. A.D. has received consultancy fees and research support from Metagenics Inc. (outside the scope of the present work). T.M.F. has consulted for Boeringer Ingelheim and

Sanofi-Aventis on the genetics of diabetes and has an MRC CASE studentship with GSK. G.R.A. is a consultant for 23andMe, Regeneron, Merck, and Helix. R.A.S. is an employee of and shareholder in GlaxoSmithKline. N.S. is working with Boehringer-Ingelheim on a genetics project but has received no remuneration. M.I.M. has served on advisory panels for NovoNordisk and Pfizer, and has received honoraria from NovoNordisk, Pfizer, Sanofi-Aventis, and Eli Lilly. The companies named above had no role in the design or conduct of this study; collection, management, analysis, and interpretation of the data, or in the preparation, review, or approval of the manuscript. Authors affiliated with deCODE (V.S., G.T., U.T. and K.S.) are employed by deCODE Genetics/Amgen, Inc.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0241-6.

Reprints and permissions information is available at www.nature.com/reprints.

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

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Anubha Mahajan[1,2*], Daniel Taliun[3], Matthias Thurner[1,2], Neil R. Robertson[1,2], Jason M. Torres[1], N. William Rayner[1,2,4], Anthony J. Payne[1], Valgerdur Steinthorsdottir[5], Robert A. Scott[6], Niels Grarup[7], James P. Cook[8], Ellen M. Schmidt[3], Matthias Wuttke[9], Chloé Sarnowski[10], Reedik Mägi[11], Jana Nano[12], Christian Gieger[13,14], Stella Trompet[15,16], Cécile Lecoeur[17], Michael H. Preuss[18], Bram Peter Prins[4], Xiuqing Guo[19], Lawrence F. Bielak[20], Jennifer E. Below[21], Donald W. Bowden[22,23,24], John Campbell Chambers[25,26,27,28,29], Young Jin Kim[30], Maggie C. Y. Ng[22,23,24], Lauren E. Petty[21], Xueling Sim[31], Weihua Zhang[25,26], Amanda J. Bennett[2], Jette Bork-Jensen[7], Chad M. Brummett[32], Mickaël Canouil[17], Kai-Uwe Ec kardt[33], Krista Fischer[11], Sharon L. R. Kardia[20], Florian Kronenberg[34], Kristi Läll[11,35], Ching-Ti Liu[10], Adam E. Locke[36,37], Jian'an Luan[6], Ioanna Ntalla[38], Vibe Nylander[2], Sebastian Schönherr[34], Claudia Schurmann[18], Loïc Yengo[17], Erwin P. Bottinger[18], Ivan Brandslund[39,40], Cramer Christensen[41], George Dedoussis[42], Jose C. Florez[43,44,45,46], Ian Ford[47], Oscar H. Franco[12], Timothy M. Frayling[48], Vilmantas Giedraitis[49], Sophie Hackinger[4], Andrew T. Hattersley[50], Christian Herder[14,51], M. Arfan Ikram[12], Martin Ingelsson[49], Marit E. Jørgensen[52,53], Torben Jørgensen[54,55,56], Jennifer Kriebel[13,14], Johanna Kuusisto[57], Symen Ligthart[12], Cecilia M. Lindgren[1,58,59], Allan Linneberg[54,60,61], Valeriya Lyssenko[62,63], Vasiliki Mamakou[64], Thomas Meitinger[65,66,67], Karen L. Mohlke[68], Andrew D. Morris[69,70], Girish Nadkarni[71], James S. Pankow[72], Annette Peters[14,67,73], Naveed Sattar[74], Alena Stančáková[57], Konstantin Strauch[75,76], Kent D. Taylor[19], Barbara Thorand[14,73], Gudmar Thorleifsson[5], Unnur Thorsteinsdottir[5,77], Jaakko Tuomilehto[78,79,80,81], Daniel R. Witte[82,83], Josée Dupuis[10,84], Patricia A. Peyser[20], Eleftheria Zeggini[4], Ruth J. F. Loos[18,85], Philippe Froguel[17,86], Erik Ingelsson[87,88], Lars Lind[89], Leif Groop[62,90], Markku Laakso[57], Francis S. Collins[91], J. Wouter Jukema[16], Colin N. A. Palmer[92], Harald Grallert[13,14,93,94], Andres Metspalu[11], Abbas Dehghan[12,25,29], Anna Köttgen[9], Goncalo R. Abecasis[3], James B. Meigs[43,46,95], Jerome I. Rotter[19,96], Jonathan Marchini[1,97], Oluf Pedersen[7], Torben Hansen[7,98], Claudia Langenberg[6], Nicholas J. Wareham[6], Kari Stefansson[5,77], Anna L. Gloyn[1,2,99], Andrew P. Morris[1,8,11,100], Michael Boehnke[3,100] and Mark I. McCarthy[1,2,99,100*]

[1]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [2]Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [3]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [4]Department of Human Genetics, Wellcome Trust Sanger Institute, Hinxton, UK. [5]deCODE Genetics, Amgen Inc., Reykjavik, Iceland. [6]MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, UK. [7]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [8]Department of Biostatistics, University of Liverpool, Liverpool, UK. [9]Institute of Genetic Epidemiology, Department of Biometry, Epidemiology, and Medical Bioinformatics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany. [10]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [11]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia. [12]Department of Epidemiology, Erasmus University Medical Center, Rotterdam, the Netherlands. [13]Research Unit of Molecular Epidemiology, Institute of Epidemiology, Helmholtz Zentrum München Research Center for Environmental Health, Neuherberg, Germany. [14]German Center for Diabetes Research (DZD), Neuherberg, Germany. [15]Section of Gerontology and Geriatrics, Department of Internal Medicine, Leiden University Medical Center, Leiden, the Netherlands. [16]Department of Cardiology, Leiden University Medical Center, Leiden, the Netherlands. [17]CNRS-UMR8199, Lille University, Lille Pasteur Institute, Lille, France. [18]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [19]Department of Pediatrics, Institute for Translational Genomics and Population Sciences, LABioMed at Harbor–UCLA Medical Center, Torrance, CA, USA. [20]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA. [21]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA. [22]Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [23]Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [24]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. [25]Department of Epidemiology and Biostatistics, Imperial College London, London, UK. [26]Department of Cardiology, Ealing Hospital, London North West Healthcare NHS Trust, Middlesex, UK. [27]Imperial College Healthcare NHS Trust, Imperial College London, London, UK. [28]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore. [29]MRC–PHE Centre for Environment and Health, Imperial College London, London, UK. [30]Division of Genome Research, Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do, Republic of Korea. [31]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore. [32]Department of Anesthesiology, University of Michigan Medical School, Ann Arbor, MI, USA. [33]Department of Nephrology and Medical Intensive Care and German Chronic Kidney Disease Study, Charité, Universitätsmedizin Berlin, Berlin, Germany. [34]Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Medical University of Innsbruck, Innsbruck, Austria. [35]Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia. [36]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA. [37]Division of Genomics & Bioinformatics, Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA. [38]William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. [39]Institute of Regional Health Research, University of Southern Denmark, Odense, Denmark. [40]Department of Clinical Biochemistry, Vejle Hospital, Vejle, Denmark. [41]Medical Department, Lillebælt Hospital Vejle, Vejle, Denmark. [42]Department of Nutrition and Dietetics, Harokopio University of Athens, Athens, Greece. [43]Department of Medicine, Harvard Medical School, Boston, MA, USA. [44]Diabetes Research Center (Diabetes Unit), Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [45]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [46]Programs in Metabolism and Medical & Population Genetics, Broad Institute, Cambridge, MA, USA. [47]Robertson Centre for Biostatistics, University of Glasgow, Glasgow, UK. [48]Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK. [49]Department of Public Health and Caring Sciences, Geriatrics, Uppsala University, Uppsala, Sweden. [50]University of Exeter Medical School, University of Exeter, Exeter, UK. [51]Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [52]Steno Diabetes Center Copenhagen, Gentofte, Denmark. [53]National Institute of Public Health, Southern Denmark University, Copenhagen, Denmark. [54]Research Centre for Prevention and Health, Capital Region of Denmark, Glostrup, Denmark. [55]Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [56]Faculty of Medicine, Aalborg University, Aalborg, Denmark. [57]Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland. [58]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA. [59]Big Data Institute, Li Ka Shing Centre For Health Information and Discovery, University of Oxford, Oxford, UK. [60]Center for Clinical Research and Prevention, Bispebjerg and Frederiksberg Hospital, Frederiksberg, Denmark. [61]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [62]Department of Clinical Sciences, Diabetes and Endocrinology, Lund University Diabetes Centre, Malmö, Sweden. [63]Department of Clinical Science, KG Jebsen Center for Diabetes Research, University of Bergen, Bergen, Norway. [64]Dromokaiteio Psychiatric Hospital, National and Kapodistrian University of Athens, Athens, Greece. [65]Institute of Human Genetics, Technische Universität München, Munich, Germany. [66]Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [67]DZHK (German Centre for Cardiovascular Research), Munich Heart Alliance partner site, Munich, Germany. [68]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA. [69]Clinical Research Centre, Centre for Molecular Medicine, Ninewells Hospital and Medical School, Dundee, UK. [70]Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, UK. [71]Division of Nephrology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [72]Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN, USA. [73]Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [74]Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK. [75]Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [76]Institute of Medical Informatics, Biometry, and Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany. [77]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [78]Department of Health, National Institute for Health and Welfare, Helsinki, Finland. [79]Dasman Diabetes Institute, Dasman, Kuwait. [80]Department of Neuroscience and Preventive Medicine, Danube-University Krems, Krems, Austria. [81]Diabetes Research Group, King Abdulaziz University, Jeddah, Saudi Arabia. [82]Department of Public Health, Aarhus University, Aarhus, Denmark. [83]Danish Diabetes Academy, Odense, Denmark. [84]National Heart, Lung, and Blood Institute Framingham Heart Study, Framingham, MA, USA. [85]Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [86]Department of Genomics of Common Disease, School of Public Health, Imperial College London, London, UK. [87]Division of Cardiovascular Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [88]Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [89]Department of Medical Sciences, Uppsala University, Uppsala, Sweden. [90]Finnish Institute for Molecular Medicine (FIMM), University of Helsinki, Helsinki, Finland. [91]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [92]Pat Macpherson Centre for Pharmacogenetics and Pharmacogenomics, Ninewells Hospital and Medical School, University of Dundee, Dundee, UK. [93]Clinical Cooparation Group Type 2 Diabetes, Helmholtz Zentrum München, Ludwig-Maximilians-Universität, Munich, Germany. [94]Clinical Cooparation Group Nutrigenomics and Type 2 Diabetes, Helmholtz Zentrum München, Technical University, Munich, Germany. [95]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA. [96]Departments of Medicine, Institute for Translational Genomics and Population Sciences, LABioMed at Harbor–UCLA Medical Center, Torrance, CA, USA. [97]Department of Statistics, University of Oxford, Oxford, UK. [98]Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark. [99]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals Trust, Oxford, UK. [100]These authors contributed equally: Andrew P. Morris, Michael Boehnke, Mark I. McCarthy. *e-mail: anubha@well.ox.ac.uk; mark.mccarthy@drl.ox.ac.uk

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.

Study-level analyses. We considered a total of 74,124 T2D cases and 824,006 controls from 32 GWAS undertaken in individuals of European ancestry (Supplementary Table 1), genotyped with a variety of genome-wide SNP arrays. Sample and variant quality control were performed within each study (Supplementary Table 1). To improve the quality of the genotype scaffold in each study, we developed a harmonized protocol in which variants were subsequently removed if (i) allele frequencies differed from those for European-ancestry haplotypes from the HRC reference panel[4] by more than 20%; AT/GC variants had MAF >40% because of potential undetected errors in strand alignment; or (iii) MAF <1% because of difficulties in calling rare variants (with exception of BioMe, MGI, and UPCH, for which no MAF exclusion was implemented, because genotyping was performed with newer GWAS arrays; Supplementary Table 1). Each scaffold, with the exception of the deCODE GWAS, was then imputed up to the HRC reference panel[4]. The GWAS from deCODE was imputed up to a reference panel based on 30,440 Icelandic whole-genome sequences[5], and only variants that were present in the HRC panel were considered for downstream analyses. Within each study, all variants were tested for association with T2D in a regression framework, with and without adjustment for BMI, in sex-combined and sex-specific analyses, under an additive model of the effect of the minor allele, with additional adjustment for study-specific covariates (Supplementary Table 1). To account for population structure and relatedness, association analyses were either adjusted for principal components (after exclusion of related individuals) or implemented in a mixed model with random effects for kinship from a genetic-relationship matrix. For studies analyzed with linear mixed models, implemented in EMMAX[49] or BOLT-LMM[50] (Supplementary Table 1), allelic effects and standard errors were converted to the log-odds scale to correct for case–control imbalance[51]. For each analysis, in each study, variants were removed from a study if (i) minor allele count <5 (in cases and controls combined); (ii) imputation quality $\hat{r}^2$ <0.3 (miniMAC) or proper info <0.4 (IMPUTE4); or (iii) standard error of the allelic log OR >10. The association summary statistics for each analysis within each study were then corrected for residual structure by means of a genomic-control inflation factor[52], calculated after exclusion of variants mapping to established T2D-susceptibility loci (Supplementary Table 1).

Sex-combined meta-analysis. We aggregated association summary statistics from sex-combined analyses for each variant across studies, with and without adjustment for BMI, using fixed-effects meta-analysis with inverse-variance weighting of log ORs, as implemented in METAL[53]. The BMI-unadjusted meta-analysis was subsequently corrected for residual inflation (to account for structure between studies) by means of genomic control ($\lambda = 1.013$) (ref. [52]), calculated after the exclusion of variants mapping to established T2D-susceptibility loci. No adjustment was required for the BMI-adjusted meta-analysis ($\lambda = 0.992$). From the meta-analysis, variants were extracted that passed quality control in at least two studies. Heterogeneity in allelic effect sizes between studies contributing to the meta-analysis was assessed with Cochran's Q statistic[54]. We defined novel loci as those mapping > 500 kb and conditionally independent from a previously reported lead GWAS SNP.

In the present study, we maintained the conventional genome-wide-significance threshold of 5×10^{-8}, for compatibility with previous reports. We recognize that more comprehensive capture of lower-frequency variants in particular increases the effective number of tests and consequently increases the false-positive rate for signals just below this threshold. 162 of the 243 primary signals were significant at a more stringent threshold (5×10^{-9}) recently advocated for whole-genome-sequence data[55], and the major conclusions of the manuscript remained unchanged when we selected this more stringent (and, given that our data lacked the full coverage of WGS data, overconservative) threshold. All summary-level data results are available so that readers can interpret the results themselves.

With our sample size (N_{eff} 231,436), assuming accurate imputation (imputation quality score >0.8), we had >80% power to detect T2D association (at $\alpha = 5 \times 10^{-8}$) with variants of MAF ≥5% and OR ≥1.10, or MAF ≥0.1% and OR ≥1.60.

Sex-differentiated meta-analysis. The meta-analyses described above were repeated for males and females separately, and correction was performed for population structure by genomic control as necessary: (i) male-specific BMI-unadjusted $\lambda = 1.029$; (ii) male-specific BMI-adjusted $\lambda = 1.001$; (iii) female-specific BMI-unadjusted $\lambda = 0.955$; and (iv) female-specific BMI-adjusted $\lambda = 0.932$. The male-specific meta-analysis consisted of up to 41,846 cases and 383,767 controls, whereas the female-specific meta-analysis consisted of up to 30,053 cases and 434,336 controls. The sex-specific meta-analyses were then combined to conduct a sex-differentiated test of association and a test of heterogeneity in allelic effects between males and females[12].

Assessment of effects of BMI adjustment. We compared the genetic effect sizes (beta coefficients) estimated from models with and without BMI adjustments, using a matched meta-analysis conducted on the same subset of 28 studies:

$$\frac{\beta_{noBMI} - \beta_{BMI}}{\sqrt{\mathrm{SE}(\beta_{noBMI})^2 + \mathrm{SE}(\beta_{BMI})^2 - 2\rho \times \mathrm{SE}(\beta_{noBMI}) \times \mathrm{SE}(\beta_{BMI})}}$$

Where β_{BMI} and β_{noBMI} are the estimated genetic effects from models with and without BMI adjustment, SE(β) is the estimated standard error of the estimates, and $\rho = 0.89$, is the estimated correlation between β_{BMI} and β_{noBMI} across all variants[1].

Detection of distinct association signals. We used GCTA[56] to perform approximate conditional analyses to detect distinct association signals at each of the genome-wide-significant risk loci for T2D (newly identified or confirmed, except at the major histocompatibility complex (MHC) region). GCTA performs conditional analysis using association summary statistics from GWAS meta-analysis and estimated LD from a sufficiently large reference study used in the meta-analysis. We used a reference sample of 6,000 (nearly) unrelated (pairwise relatedness <0.025) individuals of white British origin, randomly selected from the UK Biobank, to model patterns of LD between variants. The reference panel of genotypes consisted of the same 39 million variants from the HRC reference panel assessed in our GWAS, but with an additional quality-control step to exclude SNPs with low imputation quality (proper info <0.4) or deviation from Hardy–Weinberg equilibrium ($P < 1 \times 10^{-6}$). For each locus, we first searched ±500 kb surrounding the lead SNP (using summary statistics from BMI-unadjusted or adjusted analysis, as appropriate) to ensure that potential long-range genetic influences were assessed. Within a region, conditionally independent variants that reached locus-wide significance ($P < 10^{-5}$) were considered as index SNPs for distinct association signals. If the minimum distance between any distinct signals from two separate loci was less than 500 kb, we performed additional conditional analysis including both regions (encompassing ±500 kb from both ends) and reassessed the independence of each signal.

Fine-mapping of distinct association signals with T2D susceptibility. We considered 380 of the 403 identified distinct signals, excluding 23 that were not amenable to fine-mapping: (i) 19 signals with MAF <0.25%; (ii) three signals for which the index variant was rare and analyzed in <50% of the total effective sample size, defined as $N_e = 4 \times N_{cases} \times N_{controls}/(N_{cases} + N_{controls})$; and (iii) the one signal in the major histocompatibility complex because of the extended and complex structure of LD across the region, which complicates fine-mapping.

For each of the remaining distinct signals, we first defined a genomic region 500 kb on either side of the index variant, considering only variants with MAF >0.25% that were reported in at least 50% of the total effective sample size, thus removing those that were not well imputed in most samples. We then adopted two approaches to compute 99%-credible sets with a 99% posterior probability of containing the causal variant: (i) using a (functionally unweighted) Bayesian approach, with the strength of evidence for association measured with the Bayes' factor in favor of association for each variant[15,57]; and (ii) using (functionally weighted) fGWAS[30] that reweights the association measures using information from functional genomics data.

Genetic credible sets. For each distinct association signal, we first calculated an approximate Bayes factor[57] in favor of association on the basis of allelic effect sizes and standard errors from the meta-analysis (using BMI-unadjusted or BMI-adjusted meta-analysis, as appropriate). For loci with a single association signal, effect sizes and standard errors from unconditional meta-analysis were used. For loci with multiple distinct association signals, these parameters were derived from the approximate conditional analysis, with adjustment for all other index variants in the region. Specifically, for the jth variant,

$$\Lambda_j = \sqrt{\frac{V_j}{V_j + \omega}} \exp\left[\frac{\omega \beta_j^2}{2V_j(V_j + \omega)}\right]$$

where β_j and V_j denote the estimated allelic effect (log OR) and corresponding variance from the meta-analysis. The parameter ω denotes the prior variance in allelic effects, taken here to be 0.04 (ref. [57]).

We then calculated the posterior probability that the jth variant drives the association signal (PPA), given by

$$\pi_j = \frac{\Lambda_j}{\sum_k \Lambda_k}.$$

The 99%-credible set[15] for each locus was then constructed by (i) ordering all variants in descending order of their PPA; and (ii) including ordered variants until the cumulative PPA reached 0.99. The number of variants and length of the genomic region covered by each 99%-credible set was then calculated.

Functionally weighted credible sets. We first tested each of the 15 chromatin states in human islets and coding DNA sequence separately for enrichment using genome-

wide data with the program fGWAS[30]. Details on generation of the 15 chromatin states have been described elsewhere[25]. The annotation with the most significant enrichment was retained and tested jointly with each remaining annotation. If the most significant two-annotation model improved the model likelihood then the two annotations in the model were retained, and the process continued until the model likelihood did not exceed the previous iteration. The resulting 'full' model was iteratively pruned by dropping each annotation and assessing the cross-validated likelihood of the reduced model (i.e., an annotation was removed from the full model if dropping it increased the cross-validated likelihood). This process resulted in the 'best joint model'.

By default, fGWAS partitions the genome into 'blocks' of 5,000 SNPs and assumes no more than one causal variant per block. However, for direct comparison with the 'genetic' credible sets and to account for multiple distinct association signals within a locus, we used a modified approach. For T2D-associated regions with no evidence of more than one distinct signal, we delineated 1-Mb windows comprising all SNPs within 500 kb of the index variant and partitioned the intervening regions into ~1-Mb windows. These windows were manually input into fGWAS with the --bed command, and a separate fGWAS analysis was performed with only the set of annotations remaining in the best joint model. The genome-wide enrichments were used as priors in a Bayesian fine-mapping analysis implemented in fGWAS to calculate posterior probabilities for each SNP in the designated windows. For the remaining regions with evidence of two or more distinct association signals, we used the results from the approximate conditional analyses described above and similarly performed a manually partitioned fGWAS analysis. We then constructed 99% credible sets as described above.

Association analyses with UK Biobank phenotypes. We performed targeted association analyses using genotype and phenotype data from electronic health records (EHRs) from the UK Biobank. Hierarchical phenotype codes from EHRs were curated by grouping International Classification of Disease, Ninth Revision (ICD-9) clinical/billing codes as previously described[58]. Only phenotype codes with 20 or more cases and with a minor-allele count ≥5 in cases and controls were considered eligible for analysis. Logistic-regression analyses were performed in individuals of European ancestry for relevant phenotype–genotype combinations by adjusting for six genetic-ancestry principal components, array, and sex.

For *NEUROG3*, we tested 12 specific phenotypes that capture gastrointestinal components of the syndrome from more severe mutations in the gene, and we report nominal association without any correction for multiple testing (but we note that the various diagnoses have a complex nested, correlation structure). For the two novel rare variants, we interrogated 52 endocrine/digestive phenotypes and again found a nominal association without any correction for multiple testing.

Estimating phenotypic variance explained by SNPs. We used UK Biobank samples (19,119 T2D cases and 423,698 controls) to calculate the variance explained by genome-wide-significant variants. We ran a model regressing T2D status on all independently associated rare and low-frequency variants, assuming an additive model (and adjusting for sex, age, array, and six principal components). A separate model was run to determine the variance captured by the independently associated common variants.

Colocalization analysis. We used publicly available eQTL results from GTEx version 7 for adipose, liver, and skeletal tissues. Islet eQTLs were called using published imputed genotypes and aligned RNA-seq data (.vcf and .bam files) from human pancreatic islets of 118 individuals, downloaded from the European Genome-phenome Archive (accession number EGAS00001001265). RNA extraction, sequencing, and mapping, as well as DNA extraction, genotyping, imputation, and variant filtering were performed as previously described[33]. Gene-level reads were quantified with featureCounts version 1.5.0-p2 (ref. [59]), on the basis of a patched version of GENCODE 19 published by the GTEx Consortium. Quantified gene-level read counts for pancreatic islets were filtered in line with protocols used for GTEx version 7: only genes with at least six raw counts in 20% of the samples and TPM >0.1 in at least 20% of the samples were used for analysis. Gene-level counts for remaining genes were converted to counts per million, library sizes were normalized in edgeR version 3.16.5 (ref. [60]), and the resulting expression values were rank inverse normalized per gene. Fifteen PEER factors[61] were calculated, and cis-eQTLs were called with FastQTL version 2.0 (ref. [62]) using a cis distance of 1 Mb and PEER factors as covariates.

We performed colocalization analysis in eCAVIAR version 2.0 (ref. [63]). Colocalization was performed for each locus–tissue pair using genetic-credible-set variants from the locus that had (i) PPA >0.01, (ii) correlation data from 1000G, and (iii) available eQTL results from that tissue. Pairwise variant correlations between credible-set SNPs were calculated with PLINK version 1.9 (ref. [64]) using the 1000 Genomes Project genotypes (phase 3, October 2014 release)[16]. Final colocalization results were filtered to include only variant–gene pairs with significant eQTL effects, which were defined as associations with FDR <0.05 for islets or published significant associations based on permuted *P* values for GTEx. For a credible-set variant, an eGene with colocalization posterior probability >0.20 was considered a target gene.

Estimation of genetic variance explained. We used LD-score regression[42] to estimate the proportion of variance explained by common genetic variants for T2D on the liability scale. As advised by the developers, we based these estimates on summary statistics (without any genomic control correction) of variants restricted to the subset of HapMap[65] variants after exclusion of the MHC region. Estimations were performed for both sex-combined and sex-specific (BMI-unadjusted) analyses, by assuming a population prevalence of 10%.

Polygenic-risk-score analyses. PRSs were created for UK Biobank samples using raw genotype data in the software PRsice[46], on the basis of GWAS summary statistics of 4.6 million common variants from the sex-combined BMI-unadjusted T2D meta-analysis excluding UK Biobank samples. PRSs were created using *P*-value thresholds ranging from 5×10^{-8} to 0.5 with LD pruning parameters of $r^2=0.2$–0.8 over 250-kb windows. We then tested each PRS for classification performance in UK Biobank.

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

Data availability

Summary-level data are available at the DIAGRAM consortium website http://diagram-consortium.org/ and Accelerating Medicines Partnership T2D portal http://www.type2diabetesgenetics.org/.

References

49. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
50. Loh, P. R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
51. Cook, J. P., Mahajan, A. & Morris, A. P. Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. *Eur. J. Hum. Genet.* **25**, 240–245 (2017).
52. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
53. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
54. Ioannidis, J. P., Patsopoulos, N. A. & Evangelou, E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS One* **2**, e841 (2007).
55. Pulit, S. L., de With, S. A. & de Bakker, P. I. Resetting the bar: statistical significance in whole-genome sequencing-based association studies of global populations. *Genet. Epidemiol.* **41**, 145–151 (2017).
56. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
57. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81**, 208–227 (2007).
58. Denny, J. C. et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).
59. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
60. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
61. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
62. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
63. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
64. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
65. Frazer, K. A. et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).

nature research

Corresponding author(s): Mark I McCarthy

Reporting Summary

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

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	The software used have been described in details in Online Methods section and Supplementary Table 1. Softwares included: GenCall,Beadstudio, BRLMM, Affymetrix Power Tools, Minimac3, IMPUTE4, EPACTS, SNPTEST, METAL, PLINK, SHAPEITv2, eCAVIAR version 2.0, GCTA v1.26.0, fGWAS v0.3.6, LDSC v1.0.0, PRsice.

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

Data

Policy information about availability of data

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

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

Summary level data is available at the DIAGRAM consortium website http://diagram-consortium.org/ and Accelerating Medicines Partnership T2D portal http://www.type2diabetesgenetics.org/.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	We aimed to bring together the largest possible sample size (N>74,000 T2D cases and >824,00 controls of European ancestry) with GWAS imputed up to Haplotype Refernce Panel to study the role of genetic variants in T2D. Our sample size is adequate to recover known T2D associated regions, and identify 135 novel T2D associated regions. Also, analytical power calculation showed that our dataset has >80% power to identify variant with >5% allele frequency and 1.10 OR or variant with 0.1% allele frequency and OR 1.60.
Data exclusions	Established protocols were used to conduct rigorous data quality control for each GWAS at the study level: variants were first excluded for the following reasons: (i) monomorphic; (ii) call rate <95%; or (iii) exact p<10-4 for deviation from Hardy-Weinberg equilibrium (autosomes only) (details in Supplementary Tables 1 and Online methods). In addition, to improve the quality of the genotype scaffold in each study, we developed a harmonised protocol in which variants were subsequently removed if: (i) allele frequencies differed from those for European ancestry haplotypes from the HRC reference panel by more than 20%; AT/GC variants had MAF>40% because of potential undetected errors in strand alignment; or (iii) MAF<1% because of difficulties in calling rare variants (with exception of BioMe, MGI, and UPCH, where no MAF exclusion was implemented as they were genotyped using newer GWAS arrays; Supplementary Table 1).
Replication	We used the 42,734 T2D cases and 497,261 controls that were not included in previous discovery efforts to test the 140 lead SNPs from previously-reported T2D loci and replicated associations at 126 of these.
Randomization	Not relevant because the study is not experimental.
Blinding	Not relevant because the study is not experimental.

Reporting for specific materials, systems and methods

Materials & experimental systems

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

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Analyses were conducted on GWAS summary statistics of 74,124 T2D cases and 824,006 controls of European ancestry. Full

Population characteristics	description of sample characteristics of each study are provided in Supplementary Table 1.
Recruitment	Described for each study in Supplementary Table 1.