Genetic fine-mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci

DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium[1,2]

To advance understanding of the mechanisms through which variants at genome-wide association signals impact susceptibility to type 2 diabetes (T2D), we performed fine-mapping of 39 established risk loci in 27,206 cases and 57,574 controls of European ancestry. Through high-density imputation and conditional analyses, we identified a total of 49 distinct association signals at these loci, including five mapping in/near *KCNQ1*. We constructed "credible sets" that capture the variants most likely to drive each distinct association signal, and considered their overlap with functional annotation from T2D-relevant tissues. We identified credible set coding alleles that are likely to be driving seven association signals, including non-synonymous variants in *HNF1A* and *HNF4A*. Outside of these signals, credible set variants mapped predominantly to non-coding sequence, implying that T2D association is mediated through gene regulation. Credible set variants were enriched for overlap with FOXA2 chromatin immunoprecipitation binding sites determined in human islet and liver cells. FOXA2 enrichment was observed at 19 distinct signals, including *MTNR1B*, where genetic fine-mapping implicates rs10830963 as the variant driving the association with T2D (99.8% posterior probability). We confirmed that the T2D-risk allele of rs10830963 increased FOXA2-bound enhancer activity in both islet and liver cells. We also observed allele-specific differences in NEUROD1 binding in islets, consistent with evidence that the T2D-risk allele increases islet *MTNR1B* expression. Our study demonstrates how integration of genetic and genomic information can define the molecular mechanisms through which variants underlying genome-wide association signals exert their effects on disease risk.

Type 2 diabetes (T2D) is a heritable, chronic metabolic disorder with multifactorial pathogenesis that is characterised by hyperglycaemia, deficient insulin secretion, and resistance to insulin action[1]. Large-scale genome-wide association studies (GWAS) of common variants, defined here by minor allele frequency (MAF) of at least 5%, have been extremely successful in identifying loci contributing genetic effects to T2D susceptibility[2-6]. It has become standard practice to represent these GWAS loci by a single "lead" SNP, most often the variant with the strongest signal of association in the region. However, the lead SNP may not directly impact disease susceptibility, but instead be a proxy for the causal variant because of linkage disequilibrium (LD). Interpretation may be further complicated by the presence of more than one causal variant at a locus, acting in isolation or through the joint effects of alleles on the same haplotype, potentially resulting in multiple "distinct" association signals in the same region, which can only be delineated, statistically, through conditional analyses.

---

[1] A full list of authors and affiliations appears at the end of the paper.
[2] A list of members and affiliations of the DIAGRAM Consortium appears in the **Supplementary Note**.

With the exception of a handful of loci, including *PPARG*[7], *KCNJ11-ABCC8*[8], *SLC30A8*[9], and *GCKR*[10], where the lead SNPs are protein altering variants, the mechanisms by which associated alleles influence T2D susceptibility are largely unknown. At other loci, direct biological interpretation of the effect of genetic variation on disease susceptibility is more challenging because the association signals mostly map to non-coding sequence. While recent reports have demonstrated a relationship between T2D-associated variants and transcriptional enhancer activity, particularly in human pancreatic islets, liver cells, adipose tissue and muscle[11-15], the DNA-binding proteins through which these effects may be mediated remain obscure. Localisation of non-coding causal variants may highlight the specific regulatory elements they perturb, and potentially the genes through which they operate, providing invaluable insight into the pathophysiological basis of T2D susceptibility at GWAS loci.

To improve the localisation of causal variants for T2D association signals, and characterise the mechanisms through which they alter disease risk, we performed comprehensive fine-mapping of 39 established susceptibility loci through high-density imputation and meta-analysis of 27,206 cases and 57,574 controls from 23 studies of European ancestry, each genotyped with the Metabochip[16] (**Supplementary Tables 1 and 2**). This custom array of 196,725 variants was designed to facilitate cost-effective follow-up of nominal associations for metabolic, cardiovascular and anthropometric phenotypes, and to enhance fine-mapping, through high-density SNP coverage, of established loci for these traits. Within each locus, we aimed to: (i) evaluate the evidence for multiple distinct association signals through conditional analyses; (ii) undertake fine-mapping by defining credible sets of variants that account for ≥99% of the probability of driving each distinct association signal on the basis of statistical evidence from genetic data; and (iii) interrogate these credible sets for functional and regulatory annotation to provide insight into the mechanisms through which likely causal variants influence disease risk.

RESULTS

Coverage and imputation into Metabochip fine-mapping regions

The fine-mapping content of Metabochip includes high-density SNP coverage of 257 loci with previous evidence of genome-wide significant association ($p<5\times10^{-8}$) for at least one of 23 metabolic, cardiovascular, and anthropometric traits[16]. SNPs in fine-mapping regions were selected from reference panels from the International HapMap Consortium[17] and the 1000 Genomes (1000G) Project Consortium pilot data[18]. At design, 27 T2D susceptibility loci were selected for fine-mapping. However, subsequent T2D GWAS discovery efforts have identified additional loci that overlap, at least partially, a further 12 fine-mapping regions on Metabochip that were initially selected for other traits (**Supplementary Table 3**). Loci were ranked according to their priority for fine-mapping to determine the degree to which tagging was used to prune SNPs within each region to create the final design of the array[16] (**Supplementary Table 3**).

A comprehensive approach to locus fine-mapping would directly interrogate all genetic variation throughout the region. The phase 1 integrated reference panel (March 2012 release) of the 1000G Project Consortium[19] provides one of the most complete catalogues of variants with MAF≥0.5% in European ancestry populations. However, across

all fine-mapping regions on Metabochip, only 47.7% and 24.6%, respectively, of common and low-frequency (0.5%≤MAF<5%) variants in the European ancestry haplotypes of the 1000G reference panel are included on the array. To enhance direct interrogation of variation in the fine-mapping regions, we thus undertook imputation into the Metabochip scaffold, in each study, up to the 1000G phase 1 integrated reference panel, including haplotypes from all ancestries to reduce error rates[20] (**Online Methods**).

The quality of imputation of variants in the 1000G reference panel was variable across studies, and highly dependent on the scaffold sample size and MAF in the European ancestry haplotypes (**Supplementary Table 4**). We defined a variant to be "well-imputed" if it attained the widely-used thresholds[21] of IMPUTEv2[22] info≥0.4 or minimac[23] r^2≥0.3 in at least 80% of the total effective sample size (N_{eff}≥59,122) across studies. With this definition, 99.4% and 89.0%, respectively, of common and low-frequency variants in European ancestry haplotypes in the 1000G reference panel were well imputed, and therefore retained for downstream association analyses. Within studies, imputation quality is remarkably consistent across loci, despite the differential priority of fine-mapping regions and their coverage at design (**Supplementary Table 5**). 1000G imputation into the Metabochip scaffold thus provides consistent and near complete coverage of common and low-frequency variation across the 39 T2D susceptibility loci, and can supports direct, high-quality interrogation of the majority of potential causal variants with MAF≥0.5% in European ancestry populations.

Identification of distinct association signals at T2D susceptibility loci

The first step in fine-mapping GWAS loci is to delineate, statistically, distinct association signals arising from multiple causal variants in the same region through conditional analyses. In the context of meta-analysis, this can most efficiently be achieved through approximate conditioning, implemented in GCTA[24]. This approach uses a stepwise selection approach to ascribe "index" variants to represent each distinct association signal for which the corresponding *p*-values in an approximate joint regression model (denoted p_J) achieve a pre-defined significance threshold. GCTA has the advantage that it makes use only of unconditional meta-analysis association summary statistics and a reference study to approximate LD between variants, and hence correlation in parameter estimates in the regression model, without the need for additional cohort-level information.

Within each of the 39 T2D fine-mapping regions, we identified distinct association signals achieving "locus-wide" significance ($p_J<10^{-5}$) by applying GCTA in two stages (**Online Methods**): (i) selection of index variants on the basis of fixed-effects meta-analysis across Metabochip studies; and (ii) *in silico* replication of the index variants in a validation meta-analysis of an additional 19,662 T2D cases and 115,140 controls from 10 GWAS of European ancestry (**Supplementary Tables 1, 2 and 6**). This locus-wide significance threshold reflects a conservative Bonferroni correction allowing for up to 5,000 variants in each fine-mapping region. The allelic odds ratios (OR) and *p*-values of index variants in the GCTA joint regression model were robust to the choice of reference study used to approximate LD across the fine-mapping region (**Supplementary Table 7, Supplementary Figure 1**).

Next, because GCTA is only an approximation, we confirmed the association of each index variant in loci with multiple distinct signals through exact conditional analysis across Metabochip studies (**Online Methods**). Within each study, and for each distinct signal, we tested for T2D association with the corresponding index variant, in turn, after adjustment

for genotypes at all other index variants in the fine-mapping region. We then combined association summary statistics across studies through fixed-effects meta-analysis, and compared the allelic OR and p-value from exact conditioning with that from the GCTA joint regression model (**Supplementary Table 8**).

The most dramatic delineation of distinct association signals was observed for the fine-mapping region flanking *KCNQ1*, where five non-coding index variants attained locus-wide significance across our European ancestry Metabochip and validation studies (**Table 1**). Distinct association signals represented by three of the index variants have been reported in previous GWAS of European[5] and East Asian[25] ancestry: rs74046911 (p_J=3.6x10^{-26}, r^2=0.98 with East Asian lead SNP, rs2237897) and rs2237895 (p_J=2.1x10^{-9}, r^2=0.75 with one European lead SNP, rs163184), both of which map to a <50kb intronic recombination interval; and chr11:2692322:D (p_J=7.2x10^{-16}, r^2=0.59 with second European lead SNP, rs163184), which resides in the *KCNQ1OT1* transcript that controls regional imprinting[26]. The remaining two distinct association signals at this locus are novel, and are indexed by: rs458069 (p_J=3.2x10^{-6}), which maps to the same recombination interval as rs74046911 and rs2237895, but is in only weak LD with both (r^2=0.019 and r^2=0.245, respectively); and rs2283220 (p_J=2.2x10^{-7}), which resides in a neighbouring intron of *KCNQ1*, but outside of the recombination interval (**Supplementary Figure 2**).

At the *HNF1A* locus, we observed three distinct association signals (**Table 1, Supplementary Figure 3**), each represented by index variants that are in only weak LD with the previously reported lead GWAS SNP in this region, rs12427353. They include two non-synonymous variants, rs1169288 (p_J=4.4x10^{-14}, r^2=0.09, *HNF1A* p.I27L) and rs1800574 (p_J=4.2x10^{-10}, r^2=0.01, *HNF1A* p.A98V), and one inter-genic SNP, chr12:121440833:D (p_J=2.9x10^{-10}, r^2=0.19).

We also observed four loci with two distinct association signals (*CDKN2A-B*, *DGKB*, *MC4R* and *GIPR*), each represented by non-coding index variants (**Table 1, Supplementary Figure 4**). The two index variants at the *CDKN2A-B* locus represent the known T2D haplotype association signal mapping to a 12kb inter-genic recombination interval, downstream of the non-coding *CDKN2B-AS1* transcript[27-29]. The index variants at *DGKB* and *MC4R* confirm previously reported distinct association signals at these loci, described in meta-analyses[5] of studies that are partly overlapping with the present investigation. However, the association signal represented by the index variant mapping upstream of *GIPR*, rs4399645 (p_J=4.0x10^{-8}), has not been reported in previous GWAS of T2D susceptibility.

Finally, we observed a novel distinct association signal of at the *HNF4A* locus, represented by the coding index variant rs1800961 (p_J=1.4x10^{-9}, *HNF4A* p.T139I, [p.T130I in some previous studies[30]]). Unfortunately, this fine-mapping region was included on Metabochip because of its association with high-density lipoprotein cholesterol[16,31] (**Supplementary Table 3**), and does not include the previously reported lead T2D GWAS SNP at this locus, rs4812829, thus precluding conditional analyses in these data. However, rs4812829 is not in LD with our novel index variant (r^2=0.02), suggesting that there are at least two distinct T2D association signals at the *HNF4A* locus.

Characteristics of index variants for T2D susceptibility association signals

The "synthetic association" hypothesis posits that common lead SNPs at GWAS loci may reflect unobserved causal variants of lower frequency and greater effect size[32]. Given the

near complete coverage of variation with MAF≥0.5% in Metabochip fine-mapping regions after 1000G imputation, we sought to investigate the support for this model by considering the characteristics of index variants in terms of their allelic OR and risk allele frequency (**Supplementary Table 6, Supplementary Figure 5**). Of the distinct association signals achieving locus-wide significance across the fine-mapping regions, only three index variants are not common: rs1800574 (MAF=2.2%, OR=1.21) at the *HNF1A* locus; rs1800961 (MAF=3.9%, OR=1.16) at the *HNF4A* locus; and rs17066842 (MAF=4.8%, OR=1.12) at the *MC4R* locus. Further, as indicated above, the T2D associations with these three low-frequency alleles are distinct from those for the previously described lead SNPs at these loci, and could not account for the reported common variant GWAS signals. The contribution of even rarer variants (MAF<0.5%) to T2D susceptibility could not be directly investigated with these data because of the low quality imputation for such infrequent alleles. Nevertheless, the substantial effect sizes that would be required for rare causal variants to drive association signals are inconsistent with the observed heritability of the disease at most GWAS loci[33].

Localisation of variants driving T2D susceptibility association signals

High-density imputation and conditional analyses have identified a total of 49 distinct T2D association signals across the 39 established loci represented on Metabochip: five at *KCNQ1*, three at *HNF1A*, two each at *CDKN2A-B*, *DGKB*, *MC4R* and *GIPR*, and one each in the remainder (the previously reported lead SNP at *HNF4A* maps outside of the fine-mapping region). To fine-map these loci, we used statistical evidence of association from the meta-analysis of Metabochip studies to construct a "credible set" of variants[29] that is most likely to drive each distinct signal (**Online Methods, Supplementary Table 9, Supplementary Figure 6**). Assuming that the variant driving the association signal is reported in the meta-analysis, there is 99% probability that it will be contained within the 99% credible set. Smaller credible sets, in terms of the number of variants they contain, or the genomic interval they span, correspond to fine-mapping at higher resolution.

The 99% credible set included no more than ten variants for ten distinct association signals mapping to nine loci (**Table 2**). The greatest refinement was observed at the *MTNR1B* locus, where the credible set included only the index variant, rs10830963, accounting for more than 99.8% of the posterior probability (π_C) of driving the association signal. Small credible sets were also observed for distinct association signals at *TCF7L2* (three variants, indexed by rs7903146, mapping to 4.3kb), and *KCNQ1* (three variants, indexed by rs74046911, mapping to 200bp of the intronic recombination interval). The 99% credible sets for both distinct association signals at *CDKN2A-B* together include just 11 variants mapping to less than 2kb.

We performed functional annotation of credible set variants to search for evidence that association signals are driven by coding alleles that have not previously been interrogated by the most recent GWAS of T2D susceptibility[2-6]. In addition to previously reported non-synonymous T2D-risk alleles at *PPARG*[7], *KCNJ11-ABCC8*[8,34,35], *SLC30A8*[9,36], and *GCKR*[10,37], we identified three additional credible set coding variants that attained substantive posterior probability of driving distinct association signals at loci represented on Metabochip (**Supplementary Table 10**). These include the index variants for the novel association signals mapping to *HNF4A* (p.T139I, rs1800961, π_C=97.4%) and *HNF1A* (p.I27L, rs1169288, π_C=75.5%; p.A98V, rs1800574, π_C=34.0%) reported above. These results provide

compelling evidence that these three coding variants are likely to be driving distinct association signals at these loci. Our findings are supported by earlier studies, which not only reported evidence for association with T2D and defects in insulin secretion *in vivo*, but also demonstrated reduced transcriptional activity in a variety of *in vitro* assays[30,38]. These data provide strong evidence that *HNF4A* and *HNF1A* are effector transcripts for T2D susceptibility at these loci, a view further supported by the known impact of rare, loss of function mutations in these genes on maturity onset diabetes of the young[39,40].

Outside of these seven distinct association signals (including two at *HNF1A*), where genetic fine-mapping and previous functional data have highlighted the likely causal protein-altering alleles and transcripts, no further coding variants attained posterior probability of more than 1%. Given the near complete coverage of common and low-frequency variants in Metabochip fine-mapping regions after 1000G imputation, it is improbable that any additional distinct association signals mapping to established T2D susceptibility loci represented on the array are driven by coding variation with MAF≥0.5%. These data thus confirm previous genome-wide reports that these association signals are most likely to be mediated, instead, through gene regulation[11,14,15,41].

Investigation of regulatory mechanisms through which non-coding credible variants influence T2D susceptibility at GWAS loci

Given the evidence from recent studies that T2D-associated variants are preferentially located in enhancer elements[14,15], we next sought to gain insight into the specific transcription factors they perturb. To do this, we began by intersecting the 99% credible sets for each distinct association signal with experimental chromatin immunoprecipitation sequence (ChIP-seq) data marking binding sites for 165 proteins from ENCODE[11] and other resources[15] (**Online Methods, Supplementary Table 11**). We applied an enrichment procedure that compares the mean posterior probability of driving the association signal for credible set variants directly overlapping ChIP-seq sites, for each protein, with a null distribution obtained from randomly shifted site locations within a 1Mb window.

We first applied this procedure to credible set variants for all 49 distinct association signals, simultaneously, across the 39 established T2D susceptibility loci represented on Metabochip (**Figure 1a**). Variants in FOXA2 ChIP-seq binding sites, assayed in human hepatocellular carcinoma (HepG2)[11] and primary human pancreatic islet[15] cells, had a significantly higher posterior probability of driving the association signal (p<0.00030, Bonferroni correction for 165 proteins) than expected under the null distribution (9.2-fold enrichment, p=0.00025). We also observed nominally significant enrichment for variants in FOXA1 (5.7-fold, p=0.00054) and NKX2.2 (7.0-fold, p=0.0012) sites. The over-representation of FOXA2 ChIP-seq sites was: (i) exclusive to those shared with at least one other factor (9.6-fold, p=0.00017) compared to those that were unique (0.13-fold enrichment, p=0.38); and (ii) stronger for those in pancreatic islets (12.6-fold, p=0.00019) than in HepG2 (6.0-fold, p=0.0071) (**Figure 1b**). Given the previously reported preponderance of T2D-associated variants for islet enhancers, we tested to what extent FOXA2 enrichment is driven by co-localisation with islet enhancer sites[15]. Variants in FOXA2-bound sites remained enriched, at nominal significance (p<0.05), for the posterior probability of driving the association after removing enhancer sites (4.5-fold, p=0.018), implying that they capture additional signal outside of these regions (**Figure 1b**). These results suggest that FOXA2 binding assayed by

ChIP-seq, often at locations shared by other proteins, is a genomic marker of variants with a high posterior probability of driving association signals for T2D from genetic fine-mapping.

Having demonstrated global over-representation for FOXA2 ChIP-seq binding sites by considering all loci simultaneously, we applied the same procedure to the 99% credible set for each distinct association signal separately, to identify those with the strongest evidence for local enrichment (**Figure 1c**). We observed over-representation of credible set variants for FOXA2 binding (i.e. fold-enrichment > 1) for 19 association signals, 12 of which attained nominal significance (p<0.05). A total of 94 credible set variants at these 19 distinct association signals overlap a FOXA2 ChIP-seq site, and thus were posited to be most likely to be causal for T2D susceptibility. Of these, 18 variants (mapping to nine distinct association signals) were predicted to disrupt *de novo* recognition motifs enriched in FOXA2-bound sequence (**Supplementary Table 12**). Furthermore, these 18 variants had a mean posterior probability of driving the association signal of π_C=15.6% on the basis of genetic fine-mapping, which was more than seven times greater than for those in FOXA2 ChIP-seq sites that were not predicted to be motif-disrupting at the same signals (mean π_C=2.2%, p=0.00091) (**Figure 1d**). They include two variants with experimentally-validated differences in regulatory activity, rs7903146 (π_C=77.6%) at *TCF7L2*[42] and rs11257655 (π_C=21.1%) at *CDC123*[43]. They also include rs10830963, the index variant at the *MTNR1B* locus, which accounts for 99.8% of the posterior probability of driving the association signal on the basis of genetic fine-mapping. These results suggest that FOXA2 binding patterns can be used to identify specific variants that are potentially causal for T2D susceptibility through altered regulatory binding.

Altered regulatory activity of the credible variant at the *MTNR1B* locus

To demonstrate how local enrichment of FOXA2 binding can be used to highlight regulatory mechanisms through which credible variants might impact T2D susceptibility, we focussed on the *MTNR1B* locus. Variants mapping to this region have amongst the strongest known effects on both T2D risk[5] and regulation of fasting plasma glucose concentration[44], and physiological data suggests an impact of *MTNR1B* on both insulin secretion and insulin action[45]. The credible variant, rs10830963, overlaps a FOXA2 ChIP-seq binding site, and the risk allele, G, is predicted to create a recognition motif that matches the known consensus sequence of NEUROD1, in addition to several other factors (**Figure 2a, Supplementary Table 13**). We tested *in silico* predictions of protein binding at rs10830963 via electrophoretic mobility shift assay (EMSA) with 25bp probe fragments surrounding each allele in human pancreatic islet beta-cell (EndoC-βH1)[46] or HepG2 cell extracts. We observed allele-specific binding with extracts from both cell lines (**Figure 2b, Supplementary Figure 7**). To determine the specific protein bound, we then repeated the EMSA in the presence of antibodies, initially against NEUROD1, but then also FOXA2 and three other factors (TAL1, PTF1A and YY1), whose known sequence motifs also resemble the recognition motif (**Online Methods**). We observed a shifted risk allele band using the NEUROD1 antibody in EndoC-βH1 extracts, and an excess of the unlabelled NEUROD1 consensus sequence probe competed away the signal (**Figure 2b**). None of the tested antibodies shifted the risk allele band in HepG2 cell extracts (**Supplementary Figure 7**). This suggests that the risk allele of rs10830963 binds NEUROD1 in islets, but binds a protein that could not be identified from known recognition motifs in liver.

To relate allelic differences in protein binding to genomic activity at this site, we cloned the 224bp FOXA2 site surrounding rs10830963 into luciferase reporter vectors containing a minimal promoter, and transfected these constructs into EndoC-βH1 and HepG2 cell lines. Consistent with EMSA results and *in silico* predictions, we observed a significant increase in luciferase expression on risk allele compared to the protective allele constructs in both cell lines (p<0.05) (**Figure 2b,c**). Furthermore, RNA-seq data reported from human islets link the risk allele of rs10830963 to increased expression of *MTNR1B*[47]. These, and other data[48], indicate that T2D risk reflects increased MTNR1B expression, rather than loss of function[49]. Taken together, these results suggest that the G allele of rs10830963 increases T2D risk through increased FOXA2-bound enhancer activity, and in human islets, preferential binding of NEURDO1, which leads to higher expression of *MTNR1B*.

Identification of candidate effector genes for FOXA2 enriched association signals

Finally, we attempted to relate FOXA2 binding at the 19 enriched association signals to target effector genes. We hypothesized that the locus-specific effects of *Foxa2* knockouts in mice would serve as a proxy for identifying causal genes at human T2D association signals with local enrichment of FOXA2 binding. We obtained previously reported pancreatic islet gene expression profiles from wild-type and *Foxa1/2*-null mice[50], and mapped murine genes to human orthologs at each locus (**Online Methods**). Genes mapping within 500kb of the interval covered by the credible set at the 19 FOXA2-enriched signals were significantly down-regulated (0.47x decrease) in *Foxa1/2* knockout mice (**Supplementary Figure 8**) compared to those genome-wide (0.21x increase, p=0.00064), whilst those mapping within 500kb of other T2D association signals were not (0.26x increase, p=0.35). We observed a consistent down-regulation (0.73x decrease) when considering only those genes mapping closest to each FOXA2-enriched signal, compared to those genome-wide (0.21x increase, p=0.031) (**Supplementary Figure 8**). These results implicate candidate effector genes at enriched association signals that are affected by altered FOXA2 activity (**Supplementary Table 14**), such as *REG4* (at the *NOTCH2* locus, 15.2x decreased expression), *HHEX* (at the *HHEX-IDE* locus, 1.82x decrease), and *CAMK1D* (at the *CDC123* locus, 1.81x decrease). These data suggest that loss of *Foxa1/2* leads to preferential down-regulation of genes at FOXA2-enriched association signals at T2D susceptibility loci.

DISCUSSION

We have undertaken comprehensive fine-mapping of 39 established T2D susceptibility loci represented on Metabochip to localise potential causal variants for association signals in 27,206 cases and 57,574 controls of European ancestry. Through 1000G imputation, we achieve near complete coverage of genetic variation with MAF≥0.5% in fine-mapping regions. Conditional analyses have demonstrated that multiple distinct association signals are widespread at established loci for the disease, including five mapping to/near *KCNQ1*. These distinct association signals may reflect multiple causal variants acting in isolation or in tandem through their joint effects, *in cis*, on a haplotype, although further work would be required to distinguish between these possibilities, from both statistical and functional perspectives.

Credible set variants, which together account for 99% of the posterior probability of driving each distinct association signal, are predominantly common. Although we cannot evaluate the impact of rare variation (MAF<0.5%) in established T2D susceptibility loci without large-scale re-sequencing, our data strongly argue against a role for low-frequency variants of large effect on disease risk via synthetic association[31]. Through functional annotation of credible sets, we have demonstrated that seven distinct association signals, mapping to six established T2D susceptibility loci (amongst 39 represented on Metabochip), are likely to be driven by coding variation, including novel index variants mapping to *HNF1A* and *HNF4A*. Outside of these regions, our fine-mapping confirms previous genome-wide reports that T2D association signals are primarily driven by non-coding alleles, with effects that are mediated through gene regulation[11,14,15,41].

We have demonstrated, by genomic annotation and *in vitro* assay, that FOXA2 binding assayed by ChIP-seq can be used to pinpoint candidate causal regulatory elements, providing routes to understanding the biology of specific T2D susceptibility loci, and highlight the variants and effector transcripts through which association signals are mediated. For example, at the *MTNR1B* locus, the risk allele of the credible variant, rs10830963, which drives the T2D association signal, increases FOXA2-bound enhancer activity in human islet and liver cell lines. These data are entirely consistent with previous reports correlating the risk allele with higher *MTNR1B* expression[47,48]. FOXA2 is a pioneer factor that binds native chromatin and bookmarks genomic regions for transcriptional activity[51], and is involved in pancreatic and hepatic development[52,53]. *Foxa2* null mice have impaired insulin secretion[50], and common variants at the *FOXA2* locus are associated with fasting plasma glucose concentrations[44]. Our findings are thus consistent with the involvement of FOXA2 in maintaining normal glucose homeostasis. Common variants in *FOXA2* have also been nominally associated with T2D susceptibility in North Indians[54], but do not achieve genome-wide significance in the largest GWAS for the disease from multiple ancestry groups[2-6].

In conclusion, we have identified likely effector transcripts for distinct T2D association signals that mediate the effects of credible set variants by altering protein function or gene regulation. We have highlighted that FOXA2 binding patterns can be used to inform future hypothesis-driven investigation of the variants, genes and molecular mechanisms underlying T2D association signals mapping to non-coding sequence. Finally, our study demonstrates the utility of fine-mapping through integration of genetic and genomic information from relevant tissues and cellular models to elucidate the pathophysiology of complex human diseases, thus offering a promising avenue for translation of GWAS findings for clinical utility.

ONLINE METHODS

Metabochip imputation and association analysis. We considered a total of 27,206 T2D cases and 57,574 controls from 23 studies from populations of European ancestry (**Supplementary Table 1**), all genotyped with the Metabochip. Sample and variant quality control was performed within each study (**Supplementary Table 2**). To improve the quality of the genotype scaffold in each study, variants were subsequently removed if: (i) allele frequencies differed from those for European ancestry haplotypes from the 1000 Genomes Project Consortium[19] phase 1 integrated reference panel (March 2012 release) 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. Each scaffold was then imputed up to up to the phase 1 integrated reference panel (all ancestries, March 2012 release) from the 1000 Genomes Project Consortium[19], using IMPUTEv2[22] or minimac[23]. Within each study, well-imputed variants (IMPUTEv2[22] info>0.4 or minimac[23] r^2>0.3) were tested for T2D association under an additive model after adjustment for study-specific covariates (**Supplementary Table 2**), including principal components to adjust for population structure. Association summary statistics for each variant for each study were corrected for residual population structure using the genomic control inflation factor[55] obtained from 3,598 independent (r^2<0.05) QT-interval variants, which were not expected to be associated with T2D[5] (**Supplementary Table 2**). We then combined association summary statistics for each variant across studies via fixed-effects inverse-variance weighted meta-analysis. The results of the meta-analysis were subsequently corrected by a second round of QT-interval genomic control (λ_{QT}=1.18) to account for structure between studies. Variants were excluded from downstream analyses if they were reported in less than 80% of the total effective sample size, defined as $N_{eff} = 4 \text{x} N_{cases} \text{x} N_{controls}/(N_{cases}+N_{controls})$, thus removing those that were not well imputed in the majority of studies.

Identification of distinct association signals in established GWAS loci. We used GCTA[24] to select index variants in each of the 39 established loci represented on Metabochip with nominal evidence of association (p_J<0.001) in the approximate joint regression model on the basis of: (i) summary statistics from the fixed-effects meta-analysis Metabochip studies; and (ii) genotype data for 3,298 T2D cases and 3,708 controls of UK ancestry from GoDARTS as a reference for LD across each fine-mapping region. For comparison, we also obtained association summary statistics for the selected index variants from the GCTA joint regression model on the basis of genotype data from an alternative reference consisting of 4,435 T2D cases and 5,757 controls of Scandinavian ancestry from FUSION (**Supplementary Table 7, Supplementary Figure 1**). Selected index variants were then carried forward for *in silico* follow-up in validation meta-analysis.

The validation meta-analysis consisted of 19,662 T2D cases and 115,140 controls from 10 GWAS from populations of European ancestry, genotyped with a range of genome-wide arrays (**Supplementary Table 1**). Sample and variant quality control was performed within each study (**Supplementary Table 2**). Each scaffold was then imputed up to up to the phase 1 integrated reference panel (all ancestries, March 2012 release) from the 1000 Genomes Project Consortium[19], using IMPUTEv2[22] or minimac[23]. Within each study, well-imputed variants (IMPUTEv2[22] info≥0.4 or minimac[23] r^2≥0.3) were tested for T2D association under an additive model after adjustment for study-specific covariates (**Supplementary Table 2**), including principal components to adjust for population structure. Association

summary statistics for each variant for each study were corrected for residual population structure using the genomic control inflation factor[55] (**Supplementary Table 2**). We then combined association summary statistics for each variant across studies via fixed-effects inverse-variance weighted meta-analysis.

Association summary statistics for the selected index variants from the Metabochip and validation meta-analyses were next combined via fixed-effects inverse-variance weighted meta-analysis. In each of the 39 established loci represented on Metabochip, we used GCTA[24] was used to select index variants with locus-wide evidence of association ($p_J<10^{-5}$) in the approximate joint regression model on the basis of: (i) summary statistics from the combined meta-analysis; and (ii) genotype data for 3,298 T2D cases and 3,708 controls from GoDARTS as a reference for LD across each fine-mapping region.

For established loci with multiple index variants selected at locus-wide significance from the GCTA approximate joint regression model in combined meta-analysis, we performed exact conditioning within each Metabochip study (**Supplementary Table 8**). To obtain the association signal attributed to a specific index variant, high-quality variants (IMPUTEv2[22] info>0.4 or minimac[23] r^2>0.3) were tested for T2D association under an additive model after adjustment for study-specific covariates (**Supplementary Table 2**) and genotypes at other selected index variants in the fine-mapping region. Association summary statistics for each study were corrected for residual population structure using the QT interval genomic control inflation factor obtained in the Metabochip meta-analysis. For each association signal, summary statistics for each variant were then combined across discovery studies via fixed-effects inverse-variance meta-analysis, and subsequently corrected by a second round of QT-interval genomic control (λ_{QT}=1.18).

Credible set construction. We calculated the posterior probability that the *j*th variant, π_{Cj}, is driving a distinct association signal by

$$\pi_{\mathrm{C}j} = \frac{\Lambda_j}{\sum_k \Lambda_k},$$

where the summation is over all variants in the fine-mapping region. In this expression, Λ_j is the approximate Bayes' factor[56] for the *j*th variant, given by

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

where β_j and V_j denote the allelic effect (log-OR) and corresponding variance from the meta-analysis for the association signal across Metabochip studies. In loci with multiple distinct signals of association, results are presented from exact conditional meta-analysis after adjusting for all other index variants in the fine-mapping region. In loci with a single association signal, results are presented from unconditional meta-analysis. The parameter ω denotes the prior variance in allelic effects, taken here to be 0.04[56]. A 99% credible set[30] was then constructed by: (i) ranking all variants according to their Bayes' factor, Λ_j; and (ii) including ranked variants until their cumulative posterior probability exceeds 0.99.

Transcription factor binding data and enrichment analyses. We obtained genomic annotation data for genomic sites of protein binding (TFBS) assayed through ChIP experiments from multiple sources. We used data from the ENCODE Project Consortium[9]

for 161 proteins available from the UCSC human genome browser. We also obtained raw ChIP sequence data for additional factors assayed in primary pancreatic islets[15]. We then processed these additional factors using similar protocols to those employed by the ENCODE Project Consortium[11]. First, sequence reads were aligned to the human genome (hg19) using BWA[57] with sex-specific references, and were then converted to BAM files using SAMtools[58]. Binding sites were called from reads of each replicate, as well as reads pooled across all replicates, using MACS2[59] with a *p*-value of 0.01. Sites from each replicate of a protein were compared using an irreproducible discovery rate (IDR) threshold of 0.01. The resulting number of sites passing this IDR threshold was then used to filter the pooled sites of a protein. The set of sites were further filtered for artefacts using a blacklist of genomic regions from the ENCODE Project Consortium. Sites from all sources for each protein, including ENCODE, were then merged. The complete set of 165 proteins employed in these analyses is presented in **Supplementary Table 11**.

For each factor, we tested for overall enrichment of the posterior probability that overlapping variants in 99% credible sets are driving distinct association signals. We first calculated the mean posterior probability of all variants overlapping a factor binding site. We then generated a null distribution of the mean posterior probability by: (i) shifting the genomic locations of binding sites a random distance within 500kb in either direction; (ii) recalculating the mean posterior probability for variants overlapping shifted sites; and (iii) repeating this procedure 100,000 times. We estimated the fold-enrichment of each overlap by calculating the expected null posterior probability, and dividing the observed probability by the expected probability. We calculated a *p*-value for the enrichment by the proportion of permutations for which the expected posterior probability of driving the association signal was greater than or equal to that observed. We consider factors significantly enriched if the *p*-value was less than 0.05/165 = 0.00030 (Bonferroni correction for 165 factors). We next partitioned binding sites into those that are "shared" with another factor (i.e. genomic co-ordinates intersect a site for at least one other factor), and those that are "unique". We also partitioned binding sites based on overlap with islet active enhancer (G3) elements[15]. For each factor with significant enrichment across all credible sets (FOXA2), we applied the same enrichment analysis, but restricted to credible set variants for each distinct association signal, separately.

Motif analysis. We conducted recognition motif enhancement analyses for the set of FOXA2 ChIP-seq binding sites. First, we obtained repeat-masked genomic sequence underlying each site using the UCSC human genome browser. We scanned sequences for enrichment in these motifs using MEME-ChIP[60]. This resulted in 192 enriched motifs with E-value (expected number of hits) less than 0.05 (**Supplementary Table 13**). We compared each motif to those known from JASPAR[61], ENCODE[11] and Homer[62] using Tomtom[63].

We then identified variants in FOXA2 ChIP-seq sites predicted to disrupt an enriched recognition motif by: (i) scanning a 25bp of sequence flanking each variant allele using FIMO[64] (p<0.0001); and (ii) retaining variants in highly conserved positions (entropy less than 0.05). For the 18 variants at FOXA2-enriched association signals disrupting at least one recognition motif (**Supplementary Table 12**), we compared their posterior probabilities of causality with non-disrupting variants in FOXA2 ChIP-seq sites at the same signals using a two-sided *t*-test.

Electrophoretic mobility shift assays. EMSA was performed using nuclear extracts from human HepG2 and EndoC-BH1[47] cells. Nuclear extracts were treated with 32P gamma-ATP end-labeled double-stranded DNA probes (PerkinElmer, MA). The forward strand probe sequences used were:

rs10830963-C	TTCACACCATCTCCTATCCAGAACC
rs10830963-G	TTCACACCATCTGCTATCCAGAACC
NEUROD1	AAATCTGGCCAGCTGCTGATCCAAA

For each lane of the EMSA, 5ug of nuclear extract was incubated with 100 fmol labeled probes in a 10ul binding reaction containing 10mM Tris-HCl pH7.5, 4% glycerol, 1mM MgCl2, 0.5mM EDTA, 0.5mM DTT, 50mM NaCl and 1ug poly(dI-dC). For competition assays unlabeled probe at 100-fold excess was added to the binding reaction before addition of labeled probes. For super-shift assays the nuclear extract was pre-incubated with 1ug antibody for 30 minutes on ice before the probe was added. The following antibodies were used: anti-NEUROD1 (sc-1084X, Santa Cruz Biotechnology, Texas), anti-PTF1A (sc-98612X, Santa Cruz Biotechnology, Texas), anti-HNF3B (FOXA2) (sc-6554X, Santa Cruz Biotechnology, Texas), anti-YY1 (sc281X, Santa Cruz Biotechnology, Texas), anti-TAL1 (sc12984X, Santa Cruz Biotechnology, Texas), normal rabbit Ig (sc-2027, Santa Cruz Biotechnology, Texas), normal goat Ig (sc-2028, Santa Cruz Biotechnology, Texas).

Luciferase activity. We synthesised 224bp nucleotide sequences containing either the risk or protective allele of the *MTNR1B* enhancer sequence rs10830963 in either the forward or reverse orientation by GeneArt (Life Technologies). Complementary single-stranded oligos were then annealed and sub-cloned into the minimal promoter-driven luciferase vector pGL4.23 (Promega) using NheI and XhoI. Isolated clones were verified by sequencing.

For luciferase assays, human liver HepG2 and human beta-cell EndoC-βH1[47] cells were counted and seeded into 24 well trays (Corning) at 1.5×10^5 (HepG2) or 1.4×10^5 (EndoC-βH1) cells/well. Transfections were performed in triplicate with either Lipofectamine 2000 (HepG2) or Fugene 6 (EndoC-βH1) as per manufacturer’s instructions. Cells were transfected with 700ng pGL4.23 DNA containing the protective or risk *MTNR1B* enhancer sequence in either the forward or reverse orientation, or an equivalent amount of empty vector DNA, plus 10ng pRL-SV40 DNA (Promega) as a transfection control, per well. Cells were lysed 48 hours post-transfection and analysed for Firefly and *Renilla* luciferase activities using the Dual Luciferase Assay System (Promega) as per manufacturer's instructions, in half-volume 96 well tray format on an Enspire Multimode Plate Reader (Perkin Elmer). Firefly luciferase activity was normalised to *Renilla* luciferase activity for each well, and the results expressed as a mean normalised activity relative to empty vector-transfected cells. All experiments were performed three times in triplicate. A two-sided unpaired t-test was used to compare luciferase activity between alleles.

Mouse gene expression analysis. We obtained fold-changes in pancreatic islet gene expression in wild type compared to *Foxa1/Foxa2*-null mice[50]. We used ENSEMBL to map mouse genes to human orthologs. We filtered for human genes annotated as protein coding in GENCODE. This filtering resulted in 4,786 human protein coding genes for analysis.

First, we calculated the genomic interval spanned by the variants in each credible set. We expanded this interval for 500kb on either side, and identified the set of genes overlapping this region using BEDtools[65]. Second, for each distinct association signal, we identified the closest gene to the index variant using BEDtools[65]. We then partitioned distinct association signals into those with evidence for enriched FOXA2 binding (fold-enrichment > 1) and those without. For each analysis, we compared the fold-change in expression using a one-sided t-test between genes in each partition and all 4,786 protein coding genes.

1. Stumvoll M, *et al*. (2005). Type 2 diabetes: principles of pathogenesis and therapy. *Lancet* 365: 1333-46.
2. Kooner JS, *et al*. (2011). Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet* 43: 984-9.
3. Cho YS, *et al*. (2012). Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in East Asians. *Nat Genet* 44: 67-72.
4. Voight BF, *et al*. (2010). Twelve type 2 diabetes susceptibility loci identified through large scale association analysis. *Nat Genet* 42: 579-89.
5. Morris AP, *et al*. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* 44: 981-90.
6. Mahajan A, *et al*. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* 46: 234-44.
7. Altshuler D, *et al*. (2000). The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* 26: 76-80.
8. Gloyn AL, *et al*. (2003). Large-scale association studies of variants in genes encoding the pancreatic-cell KATP channel subunits Kir6.2 (*KCNJ11*) and SUR1 (*ABCC8*) confirm that the *KCNJ11* E23K variant is associated with type 2 diabetes. *Diabetes* 52: 568-72.
9. Sladek R, *et al*. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 445: 881-5.
10. Dupuis J, *et al*. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes. *Nat Genet* 42: 105-16.
11. The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57-74.
12. Elbein SC, *et al*. (2012). Genetic risk factors for type 2 diabetes: a trans-regulatory genetic architecture? *Am J Hum Genet* 91: 466-77.
13. Trynka G, *et al*. (2013). Chromatin marks identify critical cell types for fine-mapping complex trait variants. *Nat Genet* 45: 124-30.
14. Parker SCJ, *et al*. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbour human disease risk variants. *Proc Natl Acad Sci USA* 110: 17921-6.
15. Pasquali L, *et al*. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* 46: 136-43.
16. Voight BF, *et al*. (2012). The Metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet* 8: e1002793.
17. International HapMap Project Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851-61.
18. 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* 467: 1061-73.
19. 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56-65.
20. Marchini J, Howie B (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11: 499-511.
21. Winkler TW, *et al*. (2014). Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* 9: 1192-1212.

22. Howie BN, *et al*. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 5: e1000529.
23. Howie B, *et al*. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 44: 955-9.
24. Yang J, *et al*. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 44: 369-75.
25. Unoki H, *et al*. (2008). SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat Genet* 40: 1098-1102.
26. Fitzpatrick GV, *et al*. (2002). Regional loss of imprinting and growth deficiency in mice with a targeted deletion of KvDMR1. *Nat Genet* 32: 426-31.
27. Zeggini E, et al. (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 316: 1336-41.
28. Shea J, *et al*. (2011). Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nat Genet* 43: 801-5.
29. Maller JB, *et al*. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* 44: 1294-301.
30. Jafar-Mohammadi B, *et al*. (2011). A role for coding functional variants in HNF4A in type 2 diabetes susceptibility. *Diabetologia* 54: 111-9.
31. Teslovich TM, *et al*. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466: 707-13.
32. Dickson SP, *et al.* (2010). Rare variants create synthetic genome-wide associations. *PLoS Biol* 8: e1000294.
33. Stahl EA, *et al.* (2012). Bayesian inference of the polygenic architecture of rheumatoid arthritis. *Nat Genet* 44: 483-9.
34. Florez JC, *et al*. (2004). Haplotype structure and genotype-phenotype correlations of the sulfonylurea receptor and the islet ATP-sensitive potassium channel gene. *Diabetes* 53: 1360-8.
35. Hamming KS, *et al*. (2009). Co-expression of the type 2 diabetes susceptibility gene variants KCNJ11 E23K and ABCC8 S1369A alter the ATP and sulfonylurea sensitivities of the ATP-sensitive K+ channel. *Diabetes* 58: 2419-24.
36. Nicolson TJ, *et al*. (2009). Insulin storage and glucose homeostasis in mice null for the granule zinc transporter ZnT8 and studies of the type 2 diabetes-associated variants. *Diabetes* 58: 2070-83.
37. Beer NL, *et al*. (2009). The P446L variant in GCKR associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. *Hum Mol Genet* 18: 4081-8.
38. Holmkvist J, *et al*. (2006). Common variants in HNF-1 alpha and risk of type 2 diabetes. *Diabetologia* 49: 2882-91.
39. Yamagata K, *et al*. (1996). Mutations in the hepatocyte nuclear factor-1 alpha gene in maturity-onset diabetes of the young (MODY3). *Nature* 384: 455-8.
40. Yamagata K, *et al*. (1996). Mutations in the hepatocyte nuclear factor-4 alpha gene in maturity-onset diabetes of the young (MODY1). *Nature* 284: 458-60.
41. Gusev A, *et al*. (2014). Regulatory variants explain much more heritability than coding variants across 11 common diseases. bioRxiv doi: http://dx.doi.org/10.1101/004309

42. Gaulton KJ, *et al*. (2010). A map of open chromatin in human pancreatic islets. *Nat Genet* 42: 255-9.
43. Fogarty M, *et al*. (2014). Identification of a regulatory variant that binds FOXA1 and FOXA2 at the *CDC123/CAMK1D* type 2 diabetes GWAS locus. *PLoS Genet* (in press).
44. Manning AK, *et al*. (2012). A genome-wide approach accounting for body-mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* 44: 659-69.
45. Dimas AS, *et al*. (2014). Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes* 63: 2158-71.
46. Ravassard P, *et al*. (2011). A genetically engineered human pancreatic β cell line exhibiting glucose-inducible insulin secretion. *J Clin Invest* 121: 3589-97.
47. Fadista J, *et al*. (2014). Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc Natl Acad Sci USA* 111: 13924-9.
48. Lyssenko V, *et al*. (2009). Common variant in *MTNR1B* associated with increased risk of type 2 diabetes and impaired early insulin secretion. *Nat Genet* 41: 82-8.
49. Bonneford A, *et al*. (2012). Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat Genet* 44: 297-301.
50. Gao N, *et al*. (2010). Foxa1 and Foxa2 maintain the metabolic and secretory features of the mature beta-cell. *Mol Endocrinol* 24: 1594-1604.
51. Zaret KS, Carroll JS (2011). Pioneer transcription factors: establishing a competence for gene expression. *Genes Dev* 25: 2227-41.
52. Gao N, *et al*. (2008). Dynamic regulation of Pdx1 enhancers by Foxa1 and Foxa2 is essential for pancreas development. *Genes Dev* 22: 3435-48.
53. Lee CS, *et al*. (2005). The initiation of liver development is dependent on Foxa transcription factors. *Nature* 435: 944-7.
54. Tabassum R, *et al*. (2008). Genetic variants of FOXA2: risk of type 2 diabetes and effect on metabolic traits in North Indians. *J Hum Genet* 53: 957-65.
55. Devlin B, Roeder K (1999). Genomic control for association studies. *Biometrics* 55: 997-1004.
56. Wakefield JA (2007). Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet* 81: 208-27.
57. Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinf* 25: 1754-60.
58. Li H, *et al*. (2009). The sequence alignment/map format and SAMtools. *Bioinf* 25: 2078-9.
59. Feng J, *et al*. (2012). Identifying ChIP-seq enrichment using MACS. *Nat Protoc* 7: 1728-40.
60. Machanick P, Bailey TL (2011). MEME-ChiP. Motif analysis of large DNA datasets. *Bioinf* 27: 1696-7.
61. Mathelier A, *et al*. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* 42: D142-7.
62. Heinz S, *et al*. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b-cell identifies. *Mol Cell* 38: 576-89.

63. Bailey TL, *et al*. (2009). MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res* 37: W202-8.
64. Grant CE, *et al*. (2011). FIMO: scanning for occurrences of a given motif. *Bioinf* 27: 1017-8.
65. Quinlan AR, Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinf* 26: 841-2.

Kyle J Gaulton[1,2,154], Teresa Ferreira[1,154], Yeji Lee[3,154], Anne Raimondo[4,154], Reedik Mägi[5,154], Michael E Reschen[6,154], Anubha Mahajan[1], Adam Locke[3], N William Rayner[1,4,7], Neil Robertson[1,4], Robert A Scott[8], Inga Prokopenko[9], Laura J Scott[3], Todd Green[10], Thomas Sparso[11], Dorothee Thuillier[12], Loic Yengo[12], Harald Grallert[13,14,15], Simone Wahl[13,14,15], Mattias Frånberg[16,17,18], Rona J Strawbridge[16], Hans Kestler[19,20], Himanshu Chheda[21], Lewin Eisele[22], Stefan Gustafsson[23], Valgerdur Steinthorsdottir[24], Gudmar Thorleifsson[24], Lu Qi[25,26,27], Lennart C Karssen[28], Elisa van Leeuwen[28], Sara M Willems[8], Man Li[29], Han Chen[30,31], Christian Fuchsberger[3], Phoenix Kwan[3], Clement Ma[3], Michael Linderman[32], Yingchang Lu[33], Soren K Thomsen[4], Jana K Rundle[4], Goncalo R Abecasis[3], Peter Almgren[34], Damiano Baldassarre[35,36], Beverley Balkau[37,38], Rafn Benediktsson[39,40], Matthias Blüher[41,42], Heiner Boeing[43], Lori L Bonnycastle[44], Erwin P Borringer[45], Noël P Burtt[10], Jason Carey[10], Guillaume Charpentier[46], Peter S Chines[44], Marilyn C Cornelis[47], David J Couper[48], Andrew T Crenshaw[10], Rob M van Dam[26,49], Alex SF Doney[50,51], Mozhgan Dorkhan[52], Sarah Edkins[7], Johan G Eriksson[53,54,55,56], Tonu Esko[5,57,58], Elodie Eury[59], João Fadista[34], Jason Flannick[10], Pierre Fontanillas[10], Caroline Fox[60,61], Paul W Franks[26,34,62,63], Karl Gertow[16], Christian Gieger[13,14], Bruna Gigante[64], Omri Gottesman[45], George B Grant[10], Niels Grarup[11], Christopher J Groves[4], Maija Hassinen[65], Christian T Have[11], Christian Herder[66,67], Oddgeir L Holmen[68], Astradur B Hreidarsson[40], Steve E Humphries[69], David J Hunter[25,26,70,71], Anne U Jackson[3], Anna Jonsson[34], Marit E Jørgensen[72], Torben Jørgensen[73,74,75], Wen-Hong L Kao[29,153], Nicola D Kerrison[8], Norman Klopp[13,76], Augustine Kong[24], Peter Kovacs[41,42], Peter Kraft[25,31,71], Jasmina Kravic[34], Cordelia Langford[7], Karin Leander[64], Liming Liang[25,31], Peter Lichtner[77], Cecilia M Lindgren[1,10], Eero Lindholm[34], Allan Linneberg[73,78,79], Ching-Ti Liu[30], Stéphane Lobbens[59], Jian'an Luan[8], Valeriya Lyssenko[34,72], Satu Männistö[53], Olga McLeod[16], Julia Meyer[80], Evelin Mihailov[5], Ghazala Mirza[81], Thomas W Mühleisen[82,83], Martina Müller-Nurasyid[80,84,85,86], Carmen Navarro[87,88,89], Markus M Nöthen[82,83], Nikolay N Oskolkov[34], Katharine R Owen[4,90], Domenico Palli[91], Sonali Pechlivanis[22], Leena Peltonen[7,10,21,53,153], John RB Perry[8], Carl GP Platou[68,92], Michael Roden[66,67,93], Douglas Ruderfer[94], Denis Rybin[95], Yvonne T van der Schouw[96], Bengt Sennblad[16,17], Gunnar Sigurðsson[40,97], Alena Stančáková[98], Gerald Steinbach[99], Petter Storm[34], Konstantin Strauch[65,80], Heather M Stringham[3], Qi Sun[26,70], Barbara Thorand[14], Emmi Tikkanen[21,100], Anke Tonjes[41,42], Joseph Trakalo[1], Elena Tremoli[35,36], Tiinamaija Tuomi[56,101], Benjamin F Voight[102], Roman Wennauer[103], Steven Wiltshire[1,153], Andrew R Wood[104], Eleftheria Zeggini[7], Ian Dunham[105], Ewan Birney[105], Lorenzo Pasquali[106], Jorge Ferrer[107,108], Christopher A O'Callaghan[6], Ruth JF Loos[8,33,45,109], Josée Dupuis[30,60], Jose C Florez[58,110,111,112], Eric Boerwinkle[113,114], James S Pankow[115], Cornelia van Duijn[28,116], Eric Sijbrands[103], James B Meigs[110,117], Frank B Hu[25,26,27], Unnur Thorsteinsdottir[24,39], Kari Stefansson[24,39], Timo A Lakka[65,118,119], Rainer Rauramaa[65,119], Michael Stumvoll[41,42], Nancy L Pedersen[120], Lars Lind[121], Sirkka M Keinanen-Kiukaanniemi[122,123], Eeva Korpi-Hyövälti[124], Timo E Saaristo[125,126], Juha Saltevo[127], Johanna Kuusisto[98], Markku Laakso[98], Andres Metspalu[5,128], Raimund Erbel[129], Karl-Heinz Jöckel[22], Susanne Moebus[22], Samuli Ripatti[7,21,100,130], Veikko Salomaa[53], Erik Ingelsson[1,23], Bernhard O Boehm[131,132], Richard N Bergman[133], Francis S Collins[44], Karen L Mohlke[134], Jaakko Tuomilehto[135,136,137,138], Kristian Hveem[68], Inger Njølstad[139], Panagiotis Deloukas[7,140], Peter J Donnelly[1,141], Timothy M Frayling[104], Andrew T Hattersley[142], Ulf de Faire[64], Anders Hamsten[16], Thomas Illig[13,76], Annette Peters[14,15,86], Stephane Cauchi[12], Rob Sladek[143,144], Philippe Froguel[9,12,59], Oluf B Pedersen[11,74,145], Torben Hansen[11,146], Andrew D Morris[50,51], Collin NA Palmer[50,51], Sekar Kathiresan[10, 111,147], Olle Melander[34], Peter M Nilsson[34], Leif C Groop[34], Inês Barroso[7,148], Claudia Langenberg[8], Nicholas J Wareham[8], Anna L Gloyn[1,4,90,155],

David Altshuler[10,110,111,112,149,150,155], Michael Boehnke[3,155], Tanya M Teslovich[3,155], Mark I McCarthy[1,4,90,155], Andrew P Morris[1,5,151,152,155]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [2]Department of Genetics, Stanford University, Stanford, CA, USA. [3]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA. [4]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK. [5]Estonian Genome Center, University of Tartu, Tartu, Estonia. [6]Centre for Cellular and Molecular Physiology, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [7]Wellcome Trust Sanger Institute, Hinxton, UK. [8]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK. [9]Genomics of Common Disease, Imperial College London, London, UK. [10]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA, USA. [11]The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [12]Lille Institute of Biology, European Genomics Institute of Diabetes, Lille, France. [13]Research Unit of Molecular Epidemiology, Helmholtz Zentrum Muenchen, German Research Centre for Environmental Health, Neuherberg, Germany. [14]Institute of Epidemiology II, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany. [15]German Center for Diabetes Research, Neuherberg, Germany. [16]Atherosclerosis Research Unit, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden. [17]Science for Life Laboratory, Stockholm, Sweden. [18]Department for Numerical Analysis and Computer Science, Stockholm University, Stockholm, Sweden. [19]Friedrich-Schiller University Jena and Leibniz Institute for Age Research, Fritz Lipmann Institute, Jena, Germany. [20]Research Group Bioinformatics and Systems Biology, Institute of Neural Information Processing, University of Ulm, Ulm, Germany. [21]Institute for Molecular Medicine Finland (FIMM), Helsinki, Finland. [22]Institute for Medical Informatics, Biometry and Epidemiology, University Hospital of Essen, Essen, Germany. [23]Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [24]deCODE Genetics, Amgen inc., Reykjavik, Iceland. [25]Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA. [26]Department of Nutrition, Harvard School of Public Health, Boston, MA, USA. [27]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [28]Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands. [29]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [30]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [31]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA. [32]Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [33]The Genetics of Obesity and Related Metabolic Traits Program, The Icahn School of Medicine at Mount Sinai, New York, NY, USA. [34]Lund University Diabetes Centre, Department of Clinical Science Malmo, Scania University Hospital, Lund University, Malmo, Sweden. [35]Centro Cardiologico Monzino, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS), Milan, Italy. [36]Dipartimento di Scienze Farmacologiche e Biomolecolari, Università di Milano, Milan, Italy. [37]INSERM CESP U1018, Villejuif, France. [38]University Paris Sud 11, UMRS 1018, Villejuif, France. [39]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [40]Landspitali University Hospital, Reykjavik, Iceland. [41]IFB Adiposity Diseases, University of Leipzig, Leipzig, Germany. [42]Department of Medicine, University of Leipzig, Leipzig, Germany. [43]German Institute of Human Nutrition, Potsdam-Rehbruecke, Germany. [44]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [45]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [46]Endocrinology-Diabetology Unit, Corbeil-Essonnes Hospital, Corbeil-Essonnes, France. [47]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [48]Collaborative Studies Coordinating Center, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [49]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore.

[50]Diabetes Research Centre, Biomedical Research Institute, University of Dundee, Ninewells Hospital, Dundee, UK. [51]Pharmacogenomics Centre, Biomedical Research Institute, University of Dundee, Ninewells Hospital, Dundee, UK. [52]Lund University Diabetes Centre, Department of Clinical Science Malmo, NovoNordisk Scandinavia AB, Malmo,Sweden. [53]Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland. [54]Department of General Practice and Primary Health Care, University of Helsinki, Helsinki, Finland. [55]Unit of General Practice, Helsinki University General Hospital, Helsinki, Finland. [56]Folkhalsan Research Center, Helsinki, Finland. [57]Division of Genetics and Endocrinology, Children's Hospital, Boston, MA, USA. [58]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA. [59]CNRS-UMR-8199, Institute of Biology and Lille 2 University, Pasteur Institute, Lille, France. [60]National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA, USA. [61]Division of Endocrinology and Metabolism, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [62]Department of Clinical Sciences, Lund University, Malmo, Sweden. [63]Department of Public Health and Clinical Medicine, Umea University, Umea, Sweden. [64]Division of Cardiovascular Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [65]Kuopio Research Institute of Exercise Medicine, Kuopio, Finland. [66]Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Dusseldorf, Dusseldorf, Germany. [67]German Center for Diabetes Research, Partner Site Dusseldorf, Germany. [68]HUNT Research Center, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger, Norway. [69]Cardiovascular Genetics, BHF Laboratories, Institute Cardiovascular Sciences, UCL, London, UK. [70]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [71]Program in Genetic Epidemiology and Statistical Genetics, Harvard School of Public Health, Boston, MA, USA. [72]Steno Diabetes Center, Gentofte, Denmark. [73]Research Centre for Prevention and Health, Capital Region of Denmark, Copenhagen, Denmark. [74]Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [75]Faculty of Medicine, University of Aalborg, Aalborg, Denmark. [76]Hannover Unified Biobank, Hannover Medical School, Hannover, Germany. [77]Institute of Human Genetics, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany. [78]Department of Clinical Experimental Research, Glostrup University Hospital, Glostrup, Denmark. [79]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark. [80]Institute of Genetic Epidemiology, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany. [81]BRC Genomics Core Facility, GSTT Foundation Trust, Guy's & St Thomas' Hospital, London. [82]Institute of Human Genetics, University of Bonn, Bonn, Germany. [83]Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany. [84]Department of Medicine I, University Hospital Grosshadern, Ludwig Maximilians Universitat, Munich, Germany. [85]Chair of Genetic Epidemiology, Institute of Medical Informatics, Biometry and Epidemiology, Ludwig Maximilians Universitat, Neuherberg, Germany. [86]Deutsches Forschungszentrum fur Herz-Kreislauferkrankungen, partner site Munich Heart Alliance, Munich, Germany. [87]Department of Epidemiology, Murcia Regional Health Council, IMIB-Arrixaca, Murcia, Spain. [88]CIBER Epidemiología y Salud Pública (CIBERESP), Spain. [89]Department of Health and Social Sciences, Universidad de Murcia, Murcia, Spain. [90]Oxford National Institute for Health Research Biomedical Research Centre, Churchill Hospital, Oxford, UK. [91]Cancer Research and Prevention Institute (ISPO), Florence, Italy. [92]Department of Internal Medicine, Levanger Hospital, Nord-Trondelag Health Trust, Levanger, Norway. [93]Department of Endocrinology and Diabetology, University Hospital Dusseldorf, Dusseldorf, Germany. [94]Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [95]Boston University Data Coordinating Center, Boston, MA, USA. [96]University Medical Center Utrecht, Utrecht, The Netherlands. [97]Icelandic Heart Association, Kopavogur, Iceland. [98]Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland. [99]Department of Clinical Chemistry and Central Laboratory, University of Ulm, Ulm, Germany. [100]Department of Public Health, Hjelt Institute, University of Helsinki, Helsinki, Finland.

[101]Department of Medicine, Helsinki University Hospital, University of Helsinki, Helsinki, Finland. [102]Perelman School of Medicine, Department of Pharmacology, University of Pennsylvania, Philadelphia, PA, USA. [103]Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands. [104]Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK. [105]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK. [106]Division of Endocrinology, Germans Trias i Pujol University Hospital and Research Institute and Josep Carreras Leukaemia Research Institute, Badalona, Spain. [107]Department of Medicine, Imperial College London, London, UK. [108]Institut d'Investigacions Biomediques August Pi i Sunyer, Centre Esther Koplowitz, Barcelona, Spain. [109]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [110]Department of Medicine, Harvard Medical School, Boston, MA, USA. [111]Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA. [112]Diabetes Research Center, Diabetes Unit, Massachusetts General Hospital, Boston, MA, USA. [113]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA. [114]Human Genome Sequencing Center at Baylor College of Medicine, Houston, TX, USA. [115]Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, MN, USA. [116]Netherlands Genomics Initiative, Netherlands Consortium for Healthy Ageing and Center for Medical Systems Biology, Rotterdam, The Netherlands. [117]General Medicine Division, Massachusetts General Hospital, MA, USA. [118]Institute of Biomedicine/Physiology, University of Eastern Finland, Kuopio, Finland. [119]Department of Clinical Physiology and Nuclear Medicine, Kuopio University Hospital, Kuopio, Finland. [120]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [121]Department of Medical Sciences, Uppsala University Hospital, Uppsala, Sweden. [122]Faculty of Medicine, Institute of Health Sciences, University of Oulu, Oulu, Finland. [123]Unit of General Practice, Oulu University Hospital, Oulu, Finland. [124]South Ostrobothnia Central Hospital, Seinajoki, Finland. [125]Finnish Diabetes Association, Tampere, Finland. [126]Pirkanmaa District Hospital, Tampere, Finland. [127]Department of Medicine, Central Finland Central Hospital, Jyvasklya, Finland. [128]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia. [129]Clinic of Cardiology, West German Heart Centre, University Hospital of Essen, University Duisdurg-Essen, Essen, Germany. [130]Public Health Genomics Unit, National Institute for Health and Welfare, Helsinki, Finland. [131]Division of Endocrinology and Diabetes, Department of Internal Medicine, University Medical Centre Ulm, Ulm, Germany. [132]Lee Kong Chian School of Medicine, Imperial College London and Nanyang Technological University, Singapore, Singapore. [133]Diabetes and Obesity Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [134]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA. [135]Diabetes Prevention Unit, National Institute for Health and Welfare, Helsinki, Finland. [136]Instituto de Investigacion Sanitaria del Hospital Universario LaPaz, Madrid, Spain. [137]Centre for Vascular Prevention, Danube-University Krems, Krems, Austria. [138]Diabetes Research Group, King Abdulaziz University, Jeddah, Saudi Arabia. [139]Department of Community Medicine, Faculty of Health Sciences, University of Tromso, Tromso, Norway. [140]William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University London, London, UK. [141]Department of Statistics, University of Oxford, Oxford, UK. [142]Institute of Biomedical and Clinical Science, University of Exeter Medical School, Exeter, UK. [143]Montreal Diabetes Research Center, Centre de Recherche du Centre Hospitalier de l'Universite de Montreal, Montreal, Canada. [144]McGill University and Centre d'Innovation Génome Quebec, Montreal, Canada. [145]Faculty of Health Sciences, Aarhus University, Aarhus, Denmark. [146]Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark. [147]Cardiovascular Research Center, Massachusettes General Hospital, Boston, MA, USA. [148]University of Cambridge Metabolic Research Laboratories and NIHR Cambridge Biomedical Research Centre, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK. [149]Department of Genetics, Harvard Medical School, Boston, MA, USA. [150]Department of Molecular Biology, Harvard Medical School, Boston, MA, USA. [151]Department of Biostatistics, University of Liverpool, Liverpool, UK. [152]Department of

Molecular and Clinical Pharmacology, University of Liverpool, Liverpool, UK. [153]Deceased. [154]These authors contributed equally to this work. [155]These authors jointly directed this research.

Correspondence should be addressed to:

K.J.G. (kgaulton@gmail.com)
M.I.M. (mark.mccarthy@drl.ox.ac.uk)
A.P.M. (apmorris@liverpool.ac.uk)

AUTHOR CONTRIBUTIONS

Writing group. K.J.G., T.F., Y.Lee, A.R., R.M., M.Reschen, A.L.G., D.A., M.Boehnke, T.M.T., M.I.M., A.P.M.

Central analysis group. K.J.G., T.F., Y.Lee, R.M., A.Mahajan, A.Locke, N.W.R., N.R., T.M.T., M.I.M., A.P.M.

Annotation and functional analysis group. K.J.G., A.R., M.Reschen, S.K.T., J.K.R., I.D., E.Birney, L.Pasquali, J.Ferrer, C.A.O'C., A.L.G., M.I.M.

Validation meta-analysis group. R.M., R.A.S., I.P., L.J.S., A.P.M.

Metabochip cohort-level primary analysis. Y.Lee, T.G., T.S., D.T., L.Y., H.G., S.Wahl, M.F., R.J.S., H.K., H.Chheda, L.E., S.G., T.M.T., A.P.M.

Validation cohort-level primary analysis. V.Steinthorsdottir, G.T., L.Q., L.K., E.v.L., S.M.W., M.Li, H.Chen, C.Fuchsberger, P.Kwan, C.M., M.Linderman, Y.Lu

Cohort sample collection, genotyping, phenotyping, or additional analysis. G.R.A., P.A., D.B., B.B., R.B., M.Blüher, H.B., L.L.B., E.P.B., N.B., J.C., G.C., P.S.C., M.C.C., D.J.C., A.T.C., R.M.v.D., A.S.F.D., M.D., S.E., J.G.E., T.E., E.E., J.Fadista, J.Flannick, P.Fontanillas, C.Fox, P.W.F., K.G., C.G., B.G., O.G., G.B.G., N.G., C.J.G., M.H., C.T.H., C.H., O.L.H., A.B.H., S.E.H., D.J.H., A.U.J., A.J., M.E.J., T.J., W.H.L.K, N.D.K., N.K., A.K., P.Kovacs, P.Kraft, J.Kravic, C.Langford, K.L., L.Liang, P.L., C.M.L., E.L., A.Linneberg, C.-T.L., S.L., J.L., V.L., S. Männistö, O.McLeod, J.M., E.M., G.M., T.W.M., M.M.-N., C.N., M.M.N., N.N.O., K.R.O., D.P., S.P., L.Peltonen, J.R.B.P., C.G.P.P., M.Roden, D.Ruderfer, D.Rybin., Y.T.v.d.S., B.S., G.Sigurðsson, A.S. , G.Steinbach, P.S., K.Strauch, H.M.S., Q.S., B.T., E.Tikkanen, A.T. , J.Trakalo, E.Tremoli, T.T., B.F.V., R.W., S.Wiltshire, A.R.W., E.Z.

Validation cohort principal investigators. R.L., J.D., J.C.F., E.Boerwinkle, J.S.P., C.v.D., E.S., J.B.M., F.B.H., U.T., K.Stefansson, P.D., P.J.D., T.M.F., A.T.H., I.B., C.Langenberg, N.J.W., M.Boehnke, M.I.M.

Metabochip cohort principal investigators. T.A.L., R.R., M.S., N.L.P., L.Lind, S.K.-K., E.K.-H., T.E.S., J.S., J.Kuusisto, M.Laakso, A.Metspalu, R.E., K.-H.J., S.Moebus, S.R., V.Salomaa, E.I., B.O.B., R.N.B., F.S.C., K.L.M., J.Tuomilehto, K.H., I.N., P.D., P.J.D., T.M.F., A.T.H., U.d.F., A.H.,

T.I., A.P., S.C., R.S., P.Froguel, O.B.P., T.H., A.D.M., C.N.A.P., S.K., O.Melander, P.M.N., L.C.G., I.B., C.Langenberg, N.J.W., D.A., M.Boehnke, M.I.M.

Project management. K.J.G., T.M.T., M.I.M., A.P.M.

DIAGRAM Consortium management. D.A., M.B., M.I.M.

ACKNOWLEDGEMENTS

Funding for the research undertaken in this study has been received from: Academy of Finland (including grant numbers 77299, 102318, 118065, 123885, 124243, 129293, 129680, 136895, 139635, 213506, 251217, and 263836); Agence National de la Recherche; Association de Langue Francaise pour l'Etude du Diabete et des Maladies Metaboliques; Association Diabete Risque Vasculaire; Association Francaise des Diabetiques; British Diabetes Association (BDA) Research; Central Norway Health Authority; Central Finland Hospital District; Center for Inherited Disease Research (CIDR); City of Kuopio; City of Leutkirch; Danish Council for Independent Research; Deutsche Forschungsgemeinschaft (including project ER 155/6-2); Diabetes UK; Doris Duke Charitable Foundation; Estonian Government (SF0180142s08); European Commission (including ENGAGE HEALTH-F4-2007-201413, FP7-201413, FP7-245536, EXGENESIS LSHM-CT-2004-005272, LSHM-CT-2007-037273, Directorate C-Public Health 2004310); European Regional Development Fund; Federal Ministry of Education and Research, Germany; Federal Ministry of Health, Germany; Finnish Diabetes Association; Finnish Diabetes Research Foundation; Finnish Foundation for Cardiovascular Research; Finnish Heart Association; Finnish Medical Society; Folkhalsan Research Foundation; Foundation for Life and Health in Finland; Foundation for Old Servants; French Region of Nord Pas de Calais (Contrat de Projets Etat-Region); German Center for Diabetes Research; German Research Council (including grant number GRK1041); German National Genome Research Network; Groupe d'Etude des Maladies Metaboliques et Systemiques; Heinz Nixdorf Foundation; Helmholtz Zentrum Munchen; Helsinki University Central Hospital Research Foundation; IngaBritt and Arne Lundberg's Research Foundation (including grant number 359); Karolinska Institutet; Knut and Alice Wallenberg Foundation (including grant number KAW 2009.0243); Kuopio University Hospital; Lundbeck Foundation; Magnus Bergvall Foundation; Medical Research Council, UK (including grant numbers G0000649 and G0601261); Ministry of Education and Culture, Finland (including grant number 627;2004-2011); Ministry of Social Affairs and Health, Finland; Ministry of Innovation, Science, Research and Technology of North Rhine-Westphalia, Germany; Munich Center of Health Sciences; Narpes Health Care Foundation; National Health Screening Service of Norway; National Heart, Lung, and Blood Institute, USA (including grant numbers/contracts HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, HHSN268201100012C, N01HC25195, N02HL64278, R01HL087641, R01HL59367, and R01HL086694); National Human Genome Research Initiative, USA (including grant numbers/contracts U01HG004402 and N01HG65403); National Institute for Diabetes and Digestive and Kidney Diseases, USA (including grant numbers R01DK078616, K24DK080140, and R01DK073490); National Institute for Health and Welfare, Finland; National Institutes of Health, USA (including grant numbers/contracts HHSN268200625226C, UL1RR025005,

R01DK062370, R01DK072193, 1Z01HG000024, AG028555, AG08724, AG04563, AG10175, AG08861, U01HG004399, DK58845, CA055075, DK085545, and DK098032); Nord-Trondelag County Council; Nordic Center of Excellence in Disease Genetics; Norwegian Institute of Public Health; Norwegian Research Council; Novo Nordisk Foundation; Ollquist Foundation; Oxford NIHR Biomedical Research Centre; Paavo Nurmi Foundation; Paivikki and Sakari Sohlberg Foundation; Perklen Foundation; Pirkanmaa Hospital District, Finland; Programme Hospitalier de Recherche Clinique; Programme National de Recherche sur la Diabete; Robert Dawson Evans Endowment, Department of Medicine, Boston University School of Medicine and Boston Medical Center; Royal Swedish Academy of Sciences; Sarstedt AG & Co., Germany; Signe and Ane Gyllenberg Foundation; Sigrid Juselius Foundation; Slottery Machine Association, Finland; Social Insurance Institution of Finland; South OstroBothnia Hospital District; State of Baden-Wurttemberg, Germany; Stockholm County Council (including grant number 560183); Swedish Cultural Foundation, Finland; Swedish Diabetes Foundation; Swedish e-science Research Center ; Swedish Foundation for Strategic Research; Swedish Heart-Lung Foundation; Swedish Research Council (including grant numbers 2009-1039, 521-2010-3490, 521-2007-4037, 521-2008-2974, 825-2010-5983, 349-2008-6589, and 8691); Swedish Society of Medicine; Thuring Foundation; Tore Nilsson Foundation; Torsten and Ragnar Soderbergs Stiftelser (including grant number MT33/09); University Hospital Essen; University of Tromso; UK NIHR Cambridge Biomedical Research Centre; Uppsala University; Uppsala University Hospital; Wellcome Trust (including the Biomedical Collections Grant GR072960, and grant numbers 076113, 083948, 090367, 090532, 083270, 086596, 098017, 095101, 098051, 098381). The authors are grateful to Raphael Scharfmann (INSERM U1016, Cochin Institute Paris) for the gift of EndoC BH-1 cells and for providing technical support with their maintenance. Detailed acknowledgements are provided in the **Supplementary Note**.

COMPETING FINANCIAL INTERESTS

V.Steinthorsdottir, G.T., A.K., U.T., and K.Stefansson are employed by deCODE Genetics/Amgen inc. I.B. and spouse own stock in GlaxoSmithKline and Incyte.

FIGURE LEGENDS

Figure 1. FOX2A bound sites are a genomic marker of T2D risk variants. (A) Variants in ChIP-seq binding sites for 165 proteins were tested for enrichment of posterior probabilities compared to variants in shifted sites. Variants in FOXA2 ChIP-seq sites were significantly enriched (*p*<0.00030). (B) FOXA2 ChIP-seq sites were partitioned based on overlap with other genomic features. Sites overlapping a ChIP-seq site from at least one other factor showed stronger enrichment compared to unique sites. There was at least nominal evidence for enrichment among sites identified in islets or HepG2 cells, and sites overlapping islet enhancers or not (*p*<0.00030; **p*<0.05). (C) Variants at each signal were tested for FOXA2 enrichment. Nineteen signals had greater enrichment than expected compared to shifted sites, and at twelve signals this estimate was nominally-significant (*p*<0.05). (D) FOXA2-bound variants disrupting recognition motifs have an increased probability of being causal.**

Figure 2. Credible variant at *MTNR1B* affects FOXA2-bound enhancer activity. (A) The intronic variant, rs10830963, has a 99.8% probability of driving the association signal at the *MTRN1B* locus. This variant overlaps a FOXA2 binding site, and the risk allele G is predicted to create a binding site for a de novo recognition motif, which closely matches the NEUROD1 consensus. (B) Electrophoretic mobility shift assay of a 25bp fragment surrounding both alleles in EndoC-BH1 cell extracts. Proteins were bound to both alleles. In the presence of an antibody against NEUROD1, the risk allele bands shifted, and in the presence of a cold NEUROD1 consensus probe, the signal was competed away. (C, D) The 224bp sequence surrounding each allele was cloned into a reporter construct containing a minimal promoter and tested for luciferase activity in (C) EndoC-BH1 and (D) HepG2 cells. The risk allele had significantly increased enhancer activity over the non-risk allele in both orientations in both cell types.

Table 1. Established T2D susceptibility loci with multiple distinct signals of association at locus-wide significance in the GCTA joint regression model (p_J<10^{-5}).

Locus	Index variant	Chr	Position (b37)	Risk allele	Other allele	Metabochip GCTA joint model 27,206 cases and 57,574 controls			Validation GCTA joint model 19,662 cases and 115,140 controls			Combined GCTA joint model 46,868 cases and 172,714 controls	
						RAF	OR (95% CI)	p_J	RAF	OR (95% CI)	p_J	OR (95% CI)	p_J
DGKB	rs10276674	7	14,922,007	C	T	0.183	1.08 (1.04-1.11)	4.5×10^{-6}	0.216	1.09 (1.05-1.12)	1.3×10^{-6}	1.08 (1.06-1.11)	2.8×10^{-11}
	rs1974620	7	15,065,467	T	C	0.519	1.06 (1.04-1.09)	1.6×10^{-6}	0.515	1.05 (1.03-1.08)	0.00014	1.06 (1.04-1.08)	1.0×10^{-9}
CDKN2B	rs10811660	9	22,134,068	G	A	0.830	1.32 (1.27-1.38)	2.4×10^{-44}	0.817	1.21 (1.17-1.26)	2.6×10^{-21}	1.27 (1.23-1.30)	1.1×10^{-61}
	rs10757283	9	22,134,172	T	C	0.437	1.14 (1.10-1.17)	7.4×10^{-18}	0.436	1.11 (1.07-1.14)	1.3×10^{-10}	1.12 (1.10-1.14)	3.6×10^{-26}
KCNQ1	chr11:2692322:D	11	2,692,322	D	R	0.374	1.08 (1.05-1.10)	3.5×10^{-8}	0.413	1.09 (1.06-1.12)	1.2×10^{-8}	1.08 (1.06-1.10)	2.3×10^{-15}
	rs2283220	11	2,755,548	A	G	0.661	1.06 (1.03-1.09)	0.000016	0.710	1.05 (1.02-1.08)	0.0031	1.06 (1.03-1.08)	2.4×10^{-7}
	rs2237895	11	2,857,194	C	A	0.428	1.08 (1.05-1.11)	6.6×10^{-7}	0.433	1.07 (1.03-1.10)	2.8×10^{-4}	1.07 (1.05-1.10)	5.3×10^{-10}
	rs74046911	11	2,858,636	C	T	0.951	1.32 (1.24-1.40)	1.7×10^{-17}	0.943	1.25 (1.17-1.34)	4.8×10^{-10}	1.29 (1.23-1.35)	9.6×10^{-26}
	rs458069	11	2,858,800	G	C	0.707	1.06 (1.03-1.10)	0.00026	0.707	1.07 (1.03-1.11)	0.00085	1.06 (1.04-1.09)	1.0×10^{-6}
HNF1A	rs1169288	12	121,416,650	C	A	0.334	1.10 (1.07-1.13)	5.4×10^{-10}	0.316	1.08 (1.05-1.12)	2.8×10^{-6}	1.09 (1.07-1.12)	8.1×10^{-15}
	rs1800574	12	121,416,864	T	C	0.027	1.21 (1.11-1.31)	5.2×10^{-6}	0.020	1.23 (1.12-1.35)	0.000026	1.22 (1.14-1.29)	5.1×10^{-10}
	chr12:121440833:D	12	121,440,833	R	D	0.416	1.06 (1.03-1.09)	0.000028	0.382	1.08 (1.04-1.11)	2.5×10^{-6}	1.07 (1.05-1.09)	2.9×10^{-10}
MC4R	chr18:57739289:D	18	57,739,289	D	R	0.234	1.05 (1.02-1.09)	0.00079	0.254	1.07 (1.03-1.10)	0.000059	1.06 (1.04-1.08)	1.9×10^{-7}
	rs17066842	18	58,040,624	G	A	0.961	1.13 (1.06-1.21)	0.00033	0.948	1.11 (1.04-1.19)	0.0012	1.12 (1.07-1.17)	1.4×10^{-6}
GIPR	rs4399645	19	46,166,073	T	C	0.395	1.07 (1.04-1.10)	4.4×10^{-7}	0.441	1.05 (1.01-1.08)	0.0046	1.06 (1.04-1.08)	1.4×10^{-8}
	rs2238689	19	46,178,661	C	T	0.425	1.09 (1.07-1.12)	9.7×10^{-12}	0.424	1.07 (1.04-1.10)	9.0×10^{-6}	1.08 (1.06-1.11)	8.3×10^{-16}
HNF4A[a]	rs1800961	20	43,042,364	T	C	0.034	1.16 (1.09-1.24)	0.000011	0.041	1.16 (1.08-1.25)	0.000051	1.16 (1.10-1.22)	2.3×10^{-9}

Each distinct association signal was represented by an index variant in the GCTA joint regression model on the basis of: (i) summary statistics from a combined meta-analysis of 46,868 cases and 172,714 controls of European ancestry; and (ii) reference genotype data from GoDARTS (3,298 cases and 3,708 controls of European ancestry from the UK) to approximate LD across fine-mapping regions.

Chr: chromosome. RAF: risk allele frequency. OR: odds-ratio for risk allele. CI: confidence interval.

[a]The previously reported T2D GWAS SNP at the *HNF4A* locus (rs4812829) is not included in the fine-mapping region. However, the reported index variant, rs1800961, is independent of the GWAS SNP, and thus represents a novel distinct association signal at this locus.

Table 2. Distinct association signals at established T2D susceptibility for which the 99% credible set contains no more than ten variants.

Locus	Index variant	Chr	Position (b37)	Risk allele	Other allele	RAF	*p*-value	OR (95% CI)	99% credible set			
									SNPs	Interval (bp)	Interval start (bp)	Interval stop (bp)
MTNR1B	rs10830963	11	92,708,710	G	C	0.283	2.9×10^{-12}	1.10 (1.07-1.13)	1	1	92,708,710	92,708,710
TCF7L2	rs7903146	10	114,758,349	T	C	0.260	5.8×10^{-120}	1.39 (1.35-1.43)	3	4,279	114,754,071	114,758,349
KCNQ1	rs74046911	11	2,858,636	C	T	0.951	5.9×10^{-18}	1.33 (1.25-1.42)	3	197	2,858,440	2,858,636
ZBED3	rs7732130	5	76,435,004	G	A	0.278	6.4×10^{-10}	1.09 (1.06-1.12)	5	10,056	76,424,949	76,435,004
CDKN2A-B	rs10757283	9	22,134,172	T	C	0.437	2.8×10^{-19}	1.14 (1.11-1.18)	5	1,007	22,133,645	22,134,651
SLC30A8	rs13266634	8	118,184,783	C	T	0.676	1.3×10^{-18}	1.13 (1.10-1.16)	6	33,133	118,184,783	118,217,915
CDKN2A-B	rs10811660	9	22,134,068	G	A	0.830	7.0×10^{-43}	1.32 (1.27-1.37)	6	1,397	22,132,698	22,134,094
HNF1B	rs4430796	17	36,098,040	G	A	0.455	6.3×10^{-12}	1.09 (1.07-1.12)	7	5,791	36,097,775	36,103,565
CDKAL1	rs35261542	6	20,675,792	A	C	0.280	9.6×10^{-23}	1.15 (1.12-1.18)	8	30,073	20,673,880	20,703,952
GLIS3	chr9:4294707:I	9	4,294,707	I	R	0.360	6.5×10^{-8}	1.07 (1.05-1.10)	10	15,453	4,283,137	4,298,589

Association summary statistics and credible set construction are based on the meta-analysis of Metabochip studies in 27,206 cases and 57,574 controls of European ancestry. In loci with multiple distinct signals of association, results are presented from exact conditional analysis after adjusting for all other index variants in the fine-mapping region. In loci with a single signal of association, results are presented from unconditional analysis.
Chr: chromosome. RAF: risk allele frequency. OR: odds-ratio for risk allele. CI: confidence interval.