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

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1038/ng.3437</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:27320459">http://nrs.harvard.edu/urn-3:HUL.InstRepos:27320459</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Genetic fine-mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci

Abstract

We performed fine-mapping of 39 established type 2 diabetes (T2D) loci in 27,206 cases and 57,574 controls of European ancestry. We identified 49 distinct association signals at these loci, including five mapping in/near KCNQ1. “Credible sets” of variants most likely to drive each distinct signal 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 in human islet and liver cells, including at MTNR1B.
where fine-mapping implicated rs10830963 as driving T2D association. We confirmed that this T2D-risk allele increases FOXA2-bound enhancer activity in islet- and liver-derived cells. We observed allele-specific differences in NEUROD1 binding in islet-derived cells, consistent with evidence that the T2D-risk allele increases islet MTNR1B expression. Our study demonstrates how integration of genetic and genomic information can define molecular mechanisms through which variants underlying association signals exert their effects on disease.

INTRODUCTION

Genome-wide association studies (GWAS) of common variants, defined by minor allele frequency (MAF) ≥5%, have been successful in identifying loci contributing to type 2 diabetes (T2D) susceptibility. GWAS loci are typically represented by a “lead” SNP with the strongest signal of association in the region. However, lead SNPs may not directly impact disease susceptibility, but instead be proxies for causal variants because of linkage disequilibrium (LD). Interpretation may be further complicated by the presence of more than one causal variant at a locus, possibly acting through the joint effects of alleles on the same haplotype. This complex genetic architecture would result in multiple “distinct” association signals at the same locus, which can only be delineated, statistically, through conditional analyses.

With the exception loci where the lead SNPs are protein altering variants, including PPARδ, KCNJ11-ABCC8, SLC30A8, and GCKR, 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 T2D 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, the DNA-binding proteins through which these effects are 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 valuable insights into the pathophysiological basis of T2D susceptibility at GWAS loci.

To improve the localisation of potential causal variants for T2D, and characterise the mechanisms through which they alter disease risk, we performed comprehensive fine-mapping of 39 established loci through high-density imputation into 27,206 cases and 57,574 controls from 23 studies of European ancestry, genotyped with the Metabochip (Supplementary Tables 1 and 2). 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; and (iii) interrogate credible sets for functional and regulatory annotation to provide insight into the mechanisms through which variants driving association signals influence disease risk.
RESULTS

Imputation into Metabochip fine-mapping regions

The Metabochip includes high-density coverage of 257 “fine-mapping regions” that have been previously associated with 23 metabolic, cardiovascular, and anthropometric traits. SNPs in these regions were selected using reference data from the HapMap and the 1000 Genomes (1000G) Project. At design, 27 T2D susceptibility loci were selected for fine-mapping. However, subsequent T2D GWAS efforts have identified additional loci that overlap 12 further fine-mapping regions that were initially selected for other traits (Supplementary Table 3). To enhance coverage of variation in the fine-mapping regions, we undertook imputation into the Metabochip scaffold up to the 1000G phase 1 integrated reference panel (March 2012 release), including multi-ethnic haplotypes to reduce error rates (Online Methods).

The quality of imputation was variable across studies, particularly for MAF<5% variants, and dependent on the scaffold sample size (Supplementary Table 4). We defined variants to be “well-imputed” at widely-used thresholds of IMPUTEv2 info ≥0.4 or minimac $r^2$ ≥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 (0.5%≤MAF<5%) variants in 1000G European ancestry haplotypes were well imputed, and therefore retained for downstream association analyses. Within studies, imputation quality was consistent across loci, despite the differential priority of fine-mapping regions and their coverage of variation at design (Supplementary Table 5). 1000G imputation into the Metabochip scaffold thus provides near complete coverage of common and low-frequency variation across the 39 T2D susceptibility loci, and supports direct interrogation of the majority of variants with MAF ≥0.5% in European ancestry populations.

Distinct association signals at T2D susceptibility loci

The first step in fine-mapping GWAS loci is to delineate distinct association signals arising from multiple causal variants in the same region, which can efficiently be achieved through approximate conditioning with GCTA. Within each T2D fine-mapping region, we identified distinct signals attaining “locus-wide” significance (represented by an index variant with $p_J<10^{-5}$ in the joint association model) by applying GCTA in two stages (Online Methods). First, we selected index variants on the basis of fixed-effects meta-analysis across Metabochip studies. Second, we performed 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). Finally, because GCTA is only an approximation, we confirmed the association of each index variant through exact conditional analysis across Metabochip studies (Online Methods, Supplementary Table 7).

The most dramatic delineation of distinct association signals was observed for the region flanking KCNQ1, where five non-coding index variants attained locus-wide significance (Table 1, Supplementary Figure 1). Distinct association signals represented by three of the index variants have been reported in previous GWAS of European and East Asian populations. 4
ancestry: rs74046911 \( (p=3.6 \times 10^{-26}, r^2=0.98\) with East Asian lead SNP, rs2237897) and rs2237895 \( (p=2.1 \times 10^{-9}, r^2=0.75\) with one European lead SNP, rs163184), both of which map to a <50kb intronic recombination interval of KCNQ1; and chr11:2692322:D \( (p=7.2 \times 10^{-16}, r^2=0.59\) with second European lead SNP, rs231361), which resides in the KCNQ1OT1 transcript that controls regional imprinting. The remaining two distinct association signals at this locus are novel. The first, indexed by rs458069 \( (p=3.2 \times 10^{-6}\) maps to the same <50kb recombination interval as rs74046911 and rs2237895, but is in only weak LD with both \( r^2=0.02 \) and \( r^2=0.25\), respectively. The second, indexed by rs2283220 \( (p=2.2 \times 10^{-7}\) resides in a neighbouring intron of KCNQ1, outside of the <50kb recombination interval (Supplementary Figure 1).

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

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 3). The index variants at the CDKN2A-B locus represent the known T2D haplotype association signal mapping to a 12kb inter-genic recombination interval. Previous European ancestry GWAS meta-analyses have highlighted a potential distinct association signal, located upstream of the recombination interval in the non-coding CDKN2B-AS1 (ANRIL) transcript. However, our conditional analyses indicate that the association in this region can be fully explained by the two index SNPs in the recombination interval, which when considered together, fully extinguish the CDKN2B-AS1 signal (Supplementary Figure 4). The index variants at DGKB and MC4R also confirm previously reported distinct association signals at these loci in European ancestry GWAS meta-analyses. At the GIPR locus, the two index variants (rs2238689, \( p=8.3 \times 10^{-16}\); rs4399645, \( p=1.4 \times 10^{-8}\)) are not in strong LD with the previously reported lead SNP (rs8108269; \( r^2=0.43\) with rs2238689, \( r^2=0.00\) with rs4399645), but together can better explain the T2D association signal in this region.

Finally, we observed a novel distinct association signal at the HNF4A locus, represented by the coding index variant rs1800961 \( (p=1.4 \times 10^{-9}, HNF4A\ p.T139I,\ referred\ to\ as\ p.T130I\ in\ some\ previous\ studies)\). Unfortunately, this fine-mapping region was included on Metabochip for high-density lipoprotein cholesterol \( r^2=0.02\), suggesting that there are at least two distinct T2D association signals at the HNF4A locus.

Of the 49 distinct association signals achieving locus-wide significance across T2D loci represented on Metabochip (five at KCNQ1, three at HNF1A, two each at CDKN2A-B, DGKB, MC4R and GIPR, and one each at the remainder), only three index variants are not common (Supplementary Table 6, Supplementary Figure 5): rs1800574 (MAF=2.2%).

Nat Genet. Author manuscript; available in PMC 2016 May 18.
Localising variants driving T2D association signals

We used statistical evidence of association from the meta-analysis of Metabochip studies to construct 99% “credible sets” of variants that are most likely to drive the 49 distinct signals (Online Methods, Supplementary Table 8, Supplementary Figure 6). For ten distinct association signals, mapping to nine loci, the 99% credible set included no more than ten variants (Table 2, Supplementary Table 9). 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 of driving the association signal ($\pi_C$). Small credible sets were also observed for the association at TCF7L2 (three variants, indexed by rs7903146, mapping to 4.3kb), and one signal at KCNQ1 (three variants, indexed by rs74046911, mapping to just 200bp). The 99% credible sets for both distinct association signals at CDKN2A-B together included just 11 variants in total, and map to less than 2kb.

We performed functional annotation of credible variants to search for evidence that association signals are driven by coding alleles. Across the 49 signals, only nine coding variants attained $\pi_C>1$% (Supplementary Table 10), including six previously reported non-synonymous T2D-risk alleles at PPARG, KCNJ11-ABCC8, SLC30A8, and GCKR. The remaining three coding alleles were the index variants for association signals mapping to $\text{HNF4A}$ (p.T139I, rs1800961, $\pi_C=97.4$%) and $\text{HNF1A}$ (p.I27L, rs1169288, $\pi_C=75.5$%; p.A98V, rs1800574, $\pi_C=34.0$%). Our findings are supported by earlier studies, which reported nominal evidence for association of these three coding variants with T2D and defects in insulin secretion in vivo, and demonstrated reduced transcriptional activity of HNF1A target genes using in vitro assays. These data provide strong evidence that $\text{HNF4A}$ and $\text{HNF1A}$ are T2D effector transcripts 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. Given the near complete coverage of common and low-frequency variants in fine-mapping regions after 1000G imputation, it is unlikely that additional distinct signals in established T2D susceptibility loci represented on the Metabochip are driven by coding variation with MAF ≥0.5%, confirming reports that these associations are most likely to be mediated by effects on gene regulation.

Regulatory mechanisms underlying T2D association signals

We sought to understand the regulatory mechanisms through which variants at the 39 established T2D susceptibility loci influence disease by intersecting the 99% credible sets for each distinct association signal with chromatin immunoprecipitation sequence (ChIP-seq) data for 165 transcription factors, chromatin state maps from 12 cell types, and long non-coding RNA transcripts from 25 cell types (Online Methods, Supplementary Table 11). We applied an enrichment procedure that compared the mean posterior probability of driving the association signal for credible set variants directly overlapping sites for each regulatory annotation with a null distribution obtained from randomly shifted site locations within 100kb in either direction.
We first applied this procedure to chromatin state and non-coding RNA elements using the 19,266 credible set variants for all 49 distinct association signals (Supplementary Figure 7). Using a Bonferroni correction for the 37 tested cell type annotations \( (p<0.0014) \), variants in pancreatic islet enhancer elements \(^{14} \) had significantly higher posterior probability of driving association signals than that expected from the null distribution \( (1.97\text{-fold}, p=0.00022) \). We also observed nominal evidence for enrichment of the posterior probability of driving association signals among variants in human islet and hepatocellular carcinoma (HepG2) promoters \(^{10,14} \) \( (p=0.0052\text{ and } p=0.0064, \text{ respectively}) \). However, there was no corresponding enrichment of variants in regulatory elements for other cell types or in non-coding transcripts. These results are consistent with previous studies supporting a contribution of regulatory enhancer and promoter variants to T2D susceptibility in specific cell types \(^{11-14} \).

We next sought to gain insight into the transcription factors these regulatory variants perturb, and applied the same procedure to ChIP-seq binding data for 165 proteins (Figure 1, Supplementary Figure 8). Using a Bonferroni correction for the 165 tested proteins \( (p<0.00030) \), the 89 credible set variants overlapping 57 FOXA2 ChIP-seq binding sites, assayed in human HepG2 \(^{10} \) and islet \(^{14} \) cells, had significantly higher posterior probability of driving association signals than expected from the null distribution \( (8.24\text{-fold}, p=0.00028) \). The enrichment of FOXA2 ChIP-seq sites was exclusive to those shared with at least one other factor \( (9.18\text{-fold}, p=0.00028) \) compared to those that were not \( (1.12\text{-fold}, p=0.11) \). FOXA2 enrichment was also more pronounced among sites identified in pancreatic islets \( (15.43\text{-fold}, p=0.00045) \) than in HepG2 cells \( (4.55\text{-fold}, p=0.011) \). To exclude the possibility that this enrichment in HepG2 cells was driven by artefacts caused by a cultured cell line, we compared FOXA2 HepG2 sites to those previously assayed in primary liver \(^{39} \). We observed significant intersection of the HepG2 and liver FOXA2 sites that overlapped credible set variants \( (p=1.5\times10^{-9}) \). Consequently, we detected similar FOXA2 enrichment among sites detected in liver \( (3.63\text{-fold}, p=0.061) \) to that observed in HepG2 cells. We also compared FOXA2 ChIP-seq sites, genome-wide, from liver, HepG2 and islet cells (Supplementary Figure 9). The number of sites varied across cell types \( (8,023 \text{ for liver, } 40,866 \text{ for HepG2, and } 27,291 \text{ for islets}) \), which is likely due, in part, to technical differences including sequencing platform, depth and read length. However, the intersection of FOXA2 sites between each pair of cell types was highly significant \( (p<2.2\times10^{-16}) \).

Given the preponderance of T2D-associated variants for islet enhancers, we next tested to what extent FOXA2 enrichment is driven by co-localisation with these genomic features \(^{14} \). Variants in FOXA2-bound sites were not enriched for posterior probability of driving association signals after removing enhancer sites \( (0.36\text{-fold}, p=0.69) \). Conversely, variants in islet enhancers remained nominally enriched when removing FOXA2 sites \( (1.65\text{-fold}, p=0.014) \). These results suggest that FOXA2 binding assayed by ChIP-seq, at a subset of enhancer element locations that are often shared by other proteins, is a genomic marker of variants with an increased posterior probability of driving T2D association signals.

Having demonstrated global over-representation for FOXA2 ChIP-seq binding by considering all loci simultaneously, we applied the same procedure to the 99% credible sets of each distinct association signal, separately, to identify those with the strongest evidence...
We observed over-representation of credible set variants in islet or HepG2 FOXA2 sites for 19 association signals, 15 of which attained nominal significance ($p<0.05$). A total of 41 credible set variants at these 19 distinct association signals overlap a FOXA2 ChIP-seq site in at least one of the two cell types (Supplementary Table 12). Of these, 12 variants were predicted to disrupt de novo recognition motifs (for FOXA2 and other factors) that were enriched in FOXA2-bound sequence (Table 3, Supplementary Table 13). The mean posterior probability of driving the association ($\pi_C$) for these 12 variants was 22.0% on the basis of genetic fine-mapping (Figure 1), more than four times greater than for those in FOXA2 ChIP-seq sites that were not motif-disrupting at the same signals (mean $\pi_C$ of 5.2%, $p=0.024$). Furthermore, 11 of these 12 variants also overlapped an enhancer element in islets (9 variants) or HepG2 cells (6 variants), indicating that they are in transcriptionally active regions (Table 3). They include two variants with experimentally validated differences in regulatory activity: rs7903146 ($\pi_C=77.6\%$) at TCF7L2$^{40}$ and rs11257655 ($\pi_C=21.1\%$) at CDC123$^{31}$. 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 highlight specific variants that are potentially causal for T2D susceptibility through altered regulatory binding.

Altered regulatory activity of the MTNR1B credible variant

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$^6$ and fasting plasma glucose concentration$^42$, and physiological data indicate an impact of MTNR1B on both insulin secretion and insulin action$^{43}$. The lone credible variant at MTNR1B, rs10830963, overlaps a FOXA2 ChIP-seq binding site, and the risk allele, G, is predicted to create a recognition motif that matches the consensus sequence of NEUROD1 and several other factors (Figure 2, 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)$^{44}$ or human liver HepG2 cell extracts. We observed allele-specific binding with extracts from both cell lines (Figure 2, Supplementary Figure 10).

To determine the specific protein(s) bound at each allele, we then performed supershift experiments using antibodies directed against NEUROD1, FOXA2, and three other factors (TAL1, PTF1A, and YY1), whose consensus binding sequences resemble the recognition motif (Online Methods). We observed a shift in the presence of the NEUROD1 antibody on the risk allele in EndoC-βH1 extracts, which could be competed away by an excess of unlabelled NEUROD1 consensus sequence probe (Figure 2). None of the tested antibodies (including NEUROD1) shifted the risk allele band in HepG2 cell extracts (Supplementary Figure 10). These results demonstrate that, in vitro, the risk allele of rs10830963 preferentially binds NEUROD1 in islet-derived cells, and binds a protein not identified from known recognition motifs in liver-derived cells.
To relate allelic differences in protein binding to genomic activity at this site, we cloned a 224bp region surrounding rs10830963 into a luciferase reporter vector containing a minimal promoter, and tested its enhancer activity in EndoC-βH1 and HepG2 cell lines. Consistent with *in silico* predictions, we observed a significant (*p*<0.05) increase in luciferase expression on the risk allele compared to the protective allele in both cell lines (Figure 2). Furthermore, RNA-seq data reported from human islets have linked the T2D risk allele of rs10830963 to increased expression of *MTNR1B*\(^{45,46}\). Taken together, these results suggest that the G allele of rs10830963 increases T2D risk through increased FOXA2-bound enhancer activity, potentially mediated through NEUROD1 binding in islets, and consequently higher expression of *MTNR1B*.

### Candidate effector genes at FOXA2 enriched T2D signals

We hypothesised that the locus-specific effects of murine transcription factor knockout models would mimic patterns of binding enrichment at human disease loci. We thus attempted to relate FOXA2 binding at the 19 FOXA2-enriched association signals (Figure 1) to target effector genes using previously reported pancreatic islet expression profiles from wild-type and Foxa1/2-null mice\(^{47}\) (Online Methods). Syntenic genes mapping within 500kb of the credible set at the 19 FOXA2-enriched signals were significantly down-regulated (45.2% decrease) in Foxa1/2 knockout mice (Supplementary Figure 11) compared to those genome-wide (0.021% increase, *p*=0.012), whilst those mapping within 500kb of the other 30 T2D association signals were not (2.25% decrease, *p*=0.20). We observed a consistent down-regulation (39.6% decrease) when considering only those genes mapping closest to each FOXA2-enriched signal, compared to those genome-wide (0.021% increase, *p*=0.0021). Thus, data related to altered gene expression in Foxa1/2 knockout mice support patterns of FOXA2 binding site enrichment in humans.

We next identified specific genes at the 19 FOXA2-enriched association signals that were down-regulated in Foxa1/2 knockout mice, which might represent effector transcripts for these loci (Supplementary Table 14). Several of these genes have been previously implicated as likely effector transcripts in humans, including *TCF7L2*\(^{48,49}\) (57% decrease), *KCNJ11*\(^{7,50}\) (38% decrease), and *SLC30A8*\(^{51}\) (135% decrease). These data also implicate novel candidate effector genes at FOXA2-enriched association signals (Supplementary Table 14). For example, in Foxa1/2 knockout mice, there is a marked down-regulation of *Reg4* (1,415% decrease), which maps to a syntenic region at the FOXA2-enriched *NOTCH2* GWAS locus, highlighting *REG4* as a likely effector transcript in humans. Additional examples of candidate effector genes include *IGF2* at the *KCNQ1* locus (135% decrease), and *CAMK1D* at the *CDC123* locus (81% decrease). Together, these results provide additional support for the importance of FOXA2 binding at a subset of T2D susceptibility loci, and further highlight specific genes through which regulatory variants in these regions may operate.

### DISCUSSION

We have undertaken comprehensive fine-mapping of 39 established T2D susceptibility loci in 27,206 cases and 57,574 controls of European ancestry, and have demonstrated that
multiple distinct association signals in these regions is a common phenomenon. Index variants for just three of the 49 distinct association signals are not common, despite near complete coverage of variation with MAF ≥0.5% in fine-mapping regions after 1000G imputation. 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 via synthetic association. We have demonstrated that seven distinct association signals, mapping to six T2D susceptibility loci represented on the Metabochip, are likely to be driven by coding alleles, including novel index variants mapping to HNF1A and HNF4A. Outside of these regions, our fine-mapping confirms previous reports that T2D association signals are primarily driven by non-coding alleles, with effects that are mediated through gene regulation.

We have demonstrated, by genomic annotation and functional assays, 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. These elements highlight variants and effector transcripts through which association signals are mediated, via altered binding of either FOXA2, directly, or another transcription factor. For example, at the MTNR1B locus, the risk allele of the lone credible variant, rs10830963, which drives the T2D association signal, preferentially binds NEUROD1 in islet-derived cells in vitro, and increases FOXA2-bound enhancer activity in human islet and liver-derived cells. These data are consistent with previous reports correlating the risk allele with higher MTNR1B expression and not loss of function, and suggest altered NEUROD1 binding in islets contributes to T2D susceptibility at this locus. Further experiments will be required to establish that our in vitro findings regarding NEUROD1 binding can be confirmed in vivo. However, our attempts to perform ChIP-Seq in primary islet samples of the defined MTNR1B genotype were repeatedly unsuccessful, owing to a lack of a suitable NEUROD1 antibody. These studies are further complicated by the limited availability of primary human islets, and the slow division rate of human islet derived cell-lines is an impediment to the implementation of genome-editing technologies.

FOXA2 is a pioneer factor that binds native chromatin and bookmarks genomic regions for transcriptional activity, and is involved in pancreatic and hepatic development. FOXA2 is also expressed in other T2D-relevant cell types, such as adipocytes. Future studies will be required to elucidate the extent to which FOXA2 binding events across cell types influence disease risk. Foxa2 null mice have impaired insulin secretion, and common variants at the FOXA2 locus are associated with fasting plasma glucose concentrations. Our findings are thus consistent with the involvement of FOXA2 in maintaining normal glucose homeostasis. Common T2D-associated variants at FOXA2 have also been reported in South Asians, although they do not attain genome-wide significance in the largest GWAS for the disease from multiple ancestry groups, and therefore require further replication. Enrichment of FOXA2 binding has also been reported within genomic intervals containing GWAS signals for endocrine, neuropsychiatric, cardiovascular and cancer traits. Our study has the advantage that we consider only those FOXA2 sites that directly overlap variants that drive association signals by first fine-mapping GWAS loci, thereby providing more targeted credible sets for functional enrichment. Nevertheless, the results of
these studies, taken together, suggest a possible role for FOXA2 across a broad spectrum of complex human phenotypes.

In conclusion, 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. Continued identification of the effector transcripts at these non-coding association signals will require the use of expression QTL and knockout models, in combination with high-throughput experimental data derived from chromatin conformation capture techniques, such as Capture-C. Our findings support the use of transcription factor binding events as a means to partition susceptibility loci, potentially residing in distinct pathways, within disease-relevant cell types. 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

Ethics statement

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

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 phase 1 integrated reference panel (March 2012 release)\(^{18}\) 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 the phase 1 integrated reference panel (all ancestries, March 2012 release) from the 1000 Genomes Project Consortium\(^{18}\), using IMPUTE\(^{21}\) or minimac\(^{22}\). Within each study, well-imputed variants (IMPUTE\(^{21}\) info>0.4 or minimac\(^{22}\) \(r^2>0.3\)) 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\(^{60}\) obtained from 3,598 independent \(r^2<0.05\) QT-interval variants, which were not expected to be associated with T2D\(^{4}\) (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 (\(\hat{\lambda}_{\text{QT}}=1.18\)) 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 \( N_{\text{eff}} = 4 \times N_{\text{cases}} \times N_{\text{controls}} / (N_{\text{cases}} + N_{\text{controls}}) \), 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\(^{23}\) to select index variants in each of the 39 established loci represented on Metabochip with nominal evidence of association \((p_j < 0.001)\) with T2D in an approximate joint regression model. The GCTA model made use 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 Finnish ancestry from FUSION (Supplementary Table 15, Supplementary Figure 12). 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 the phase 1 integrated reference panel (all ancestries, March 2012 release) from the 1000 Genomes Project Consortium\(^{18}\), using IMPUTEv2\(^{21}\) or minimac\(^{22}\). Within each study, well-imputed variants (IMPUTEv2\(^{21}\) info \(\geq 0.4\) or minimac\(^{22}\) \(r^2 \geq 0.3\)) 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\(^{60}\) (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, GCTA\(^{23}\) 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 7). To obtain the association signal attributed to a specific index variant, high-quality variants (IMPUTEv2\(^{21}\) info \(> 0.4\) or minimac\(^{22}\) \(r^2 > 0.3\)) 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 ($\lambda_{QT}=1.18$).

**Credible set construction**

In an ideal fine-mapping experiment, we would calculate the posterior probability of driving each distinct association signal for all variants mapping to a locus. However, the posterior probability is determined by the association signal effect size of the variant and the corresponding standard error, which is also impacted by the quality of imputation across studies, amongst other factors. To minimise the impact of imputation quality on fine-mapping, we therefore retained only those variants that were directly typed and/or well imputed in at least 80% of the total effective sample size. Assuming that the variant driving an association signal meets these quality criteria, the probability that it would be contained within the 99% credible set would be $\sim 0.99$.

For each distinct signal, we first calculated the posterior probability, $\pi_{C_j}$, that the $j$th variant is driving the association, given by

$$\pi_{C_j} = \frac{\Lambda_j}{\sum_k \Lambda_k},$$

where the summation is over all retained variants in the fine-mapping region. In this expression, $\Lambda_j$ is the approximate Bayes’ factor for the $j$th variant, given by

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

where $\beta_j$ and $V_j$ denote the estimated allelic effect (log-OR) and corresponding variance from the meta-analysis 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 $\omega$ denotes the prior variance in allelic effects, taken here to be 0.04. The 99% credible set for each signal was then constructed by: (i) ranking all variants according to their Bayes’ factor, $\Lambda_j$ and (ii) including ranked variants until their cumulative posterior probability of driving the association attained or exceeded 0.99.

**Genomic annotation data and enrichment analyses**

We obtained genomic annotation data for transcription factor binding sites (TFBS) assayed through ChIP experiments from multiple sources. We used sites from the ENCODE Project Consortium for 161 proteins available from the UCSC human genome browser. We also obtained raw ChIP and input sequence data for additional factors assayed in primary pancreatic islets. We then processed these additional factors using protocols employed by the ENCODE Project Consortium. First, sequence reads were aligned to the human
genome (hg19) using BWA with sex-specific references, and were then converted to BAM files using SAMtools after removing duplicate and non-uniquely mapped reads. Binding sites were called from reads of each replicate, as well as reads pooled across all replicates, using SPP. Raw sites from each replicate of a protein were compared using an irreproducible discovery rate (IDR) threshold of 0.02. 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 combined. The complete set of 165 proteins employed in these analyses is presented in Supplementary Table 11. In addition, we obtained FOXA2 ChIP-seq sites that were previously identified in human liver and lifted their positions to hg19.

We obtained annotation data for five histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K36me3, and H3K27me3) and CTCF binding assayed from ChIP experiments. We used data from 9 cell types from ENCODE (Gm12878, K562, Hepg2, Hsmm, Huvec, Nhek, Nhlf, hHesc, and Hmec); we also obtained raw ChIP data assayed in primary pancreatic islets and pre-mature and mature human adipose stromal cells. We mapped reads to hg19 using BWA, and used the resulting mapped reads from these 12 cell types as input to ChromHMM. We assigned states based on the following chromatin signatures: active promoter (H3K4me3 and H3K27ac); strong enhancer 1 (H3K4me3, H3K27ac, and H3K4me1); strong enhancer 2 (H3K27ac and H3K4me1); weak enhancer (H3K4me1); poised promoter (H3K27me3, H3K4me3, and H3K4me1); repressed (H3K27me3); insulator (CTCF); and transcription (H3K36me3). For each cell type, we pooled the three enhancer states into one enhancer category, and the two promoter states into one promoter category. We also identified long non-coding RNA data from the Human Body Map (UCSC genome browser) and from pancreatic islets.

For each genomic annotation, we tested for overall enrichment of the posterior probability that overlapping variants in the 99% credible sets are driving distinct association signals (πC). We first calculated the mean posterior probability (mean πC) over the set of variants overlapping a given annotation. We then generated a null distribution of the mean posterior probability (mean πC) by: (i) shifting the genomic locations of binding sites a random distance within 100kb in either direction; (ii) recalculating the mean posterior probability for 99% credible set 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 considered cell type annotations to be significantly enriched if the p-value was less than 0.05/37 = 0.0014 (Bonferroni correction for 37 annotations). We considered TFBS annotations to be 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 interval intersects a site for at least one other factor), and those that are “unique”. We also partitioned binding sites based on overlap with islet enhancer elements. 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.

We assessed the evidence for intersection in FOXA2 ChIP-seq sites from islets$^{14}$, HepG2$^{10}$,
and liver$^{39}$, genome-wide and overlapping credible set variants, using BEDtools$^{69}$.

**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$^{70}$, which uses up to 100bp surrounding the mid-point of each site. This
resulted in 198 enriched motifs with E-value (expected number of hits) less than 0.05
(Supplementary Table 16). We compared each motif to those known from JASPAR$^{71}$,
ENCODE$^{10}$, and Homer$^{72}$ using Tomtom$^{73}$.

Second, we 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$^{74}$ ($p<0.0001$); and (ii) retaining variants in highly conserved positions (entropy less
than 0.5). For the 12 variants at FOXA2-enriched signals disrupting at least one recognition
motif (Table 3, Supplementary Table 14), we compared their posterior probabilities of
driving the association ($\pi_C$) with those for non-disrupting variants in FOXA2 ChIP-seq sites
at the same signals using a two-sided Wilcoxon rank-sum test.

**Electrophoretic mobility shift assays**

EMSA was performed using nuclear extracts from human HepG2 and EndoC-βH1 cells.
HepG2 cells were the generous gift of the Ratcliffe laboratory$^{75}$ and authenticated by
genotyping in the MHC region. Endo-βH1 cells were obtained from Endocells and have
been previously authenticated$^{44}$. Both cell lines were tested and found negative for
mycoplasma contamination. Nuclear extracts were incubated with$^{32}$ P gamma-ATP end-
labeled double-stranded DNA probes (PerkinElmer, MA). The forward strand probe
sequences used are presented in Supplementary Table 17.

For each lane of the EMSA, 5μg of nuclear extract was incubated with 100 fmol labeled
probes in a 10ul binding reaction containing 10mM Tris-HCl pH7.5, 4% glycerol, 1mM
MgCl$_2$, 0.5mM EDTA, 0.5mM DTT, 50mM NaCl and 1μg 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 1μg
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 Nhel and Xhol. Isolated clones were verified by sequencing.

For luciferase assays, human liver HepG2 and human beta-cell EndoC-βH1 cells were counted and seeded into 24 well trays (Corning) at 1.5×10⁵ (HepG2) or 1.4×10⁵ (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. 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 4,629 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. To account for syntenic differences in gene order between species, we retained only those genes that were: (a) on the same chromosome; and (b) in exactly the same relative order in both mouse and human genomes. At the GIPR locus, one of the genes was ordered differently and thus removed from the analysis. At two loci, *KCNJ11* and *HNF1A*, at least one of the genes was located on a different part of the same chromosome, and at another locus, *GCK*, genes were located on different chromosomes. For these three loci, we retained only those genes that were at the same chromosomal location to the interval covered by the credible set for the association signal for that locus (by lifting over from hg19 to mouse build mm10). Second, for each distinct association signal, we identified the closest gene to the index variant using BEDtools. We then partitioned distinct association signals into those with evidence for enriched FOXA2 binding (fold-enrichment >1) and those without, counting each gene only once in a given group. For each analysis, we converted the fold-changes to percentages, and compared the percent change in expression using a one-sided Wilcoxon rank-sum test between genes in each partition and all 4,629 protein coding genes.
Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors


Affiliations

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
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. Leibniz Institute for Age Research, Fritz Lipmann Institute, Jena, Germany
20. Medical Systems Biology, Ulm University, Ulm, Germany
21Finnish Institute for Molecular Medicine (FIMM), Helsinki, Finland 22Institute for Medical Informatics, Biometry and Epidemiology, University Hospital of Essen, Essen, Germany 23Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden 24deCODE Genetics, Amgen inc., Reykjavik, Iceland 25Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA 26Department of Nutrition, Harvard School of Public Health, Boston, MA, USA 27Channing Division of Network Medicine, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA, USA 28Department of Epidemiology, School of Public Health and Tropical Medicine, Tulane University, New Orleans, Louisiana 29Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands 30Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA 31Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA 32Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA 33Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA 34The Genetics of Obesity and Related Metabolic Traits Program, The Icahn School of Medicine at Mount Sinai, New York, NY, USA 35Perelman School of Medicine, Department of Pharmacology, University of Pennsylvania, Philadelphia, PA, USA 36Lund University Diabetes Centre, Department of Clinical Science Malmo, Scania University Hospital, Lund University, Malmo, Sweden 37Centro Cardiologico Monzino, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS), Milan, Italy 38Dipartimento di Scienze Farmacologiche e Biomolecolari, Università di Milano, Milan, Italy 39INSERM CESP U1018, Villejuif, France 40University Paris Sud 11, UMRS 1018, Villejuif, France 41Faculty of Medicine, University of Iceland, Reykjavik, Iceland 42Landspitali University Hospital, Reykjavik, Iceland 43IFB Adiposity Diseases, University of Leipzig, Leipzig, Germany 44Department of Medicine, University of Leipzig, Leipzig, Germany 45German Institute of Human Nutrition, Potsdam-Rehbruecke, Germany 46National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA 47The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA 48Endocrinology-Diabetology Unit, Corbeil-Essonnes Hospital, Corbeil-Essonnes, France 49Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA 50Collaborative Studies Coordinating Center, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA 51Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore 52Diabetes Research Centre, Biomedical Research Institute, University of Dundee, Ninewells Hospital, Dundee, UK 53Pharmacogenomics Centre, Biomedical Research Institute, University of Dundee, Ninewells Hospital, Dundee, UK 54Lund University Diabetes Centre, Department of Clinical Science Malmo, NovoNordisk Scandinavia AB, Malmo, Sweden 55Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland 56Department of General Practice and Primary Health Care, University of Helsinki, Helsinki, Finland 57Unit of General
Practice, Helsinki University General Hospital, Helsinki, Finland 58Folkhalsan Research Center, Helsinki, Finland 59Division of Endocrinology, Children’s Hospital, Boston, MA, USA 60Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA 61CNRS-UMR-8199, Institute of Biology and Lille 2 University, Pasteur Institute, Lille, France 62National Heart, Lung, and Blood Institute’s Framingham Heart Study, Framingham, MA, USA 63Division of Endocrinology and Metabolism, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA, USA 64Department of Clinical Sciences, Lund University, Malmo, Sweden 65Department of Public Health and Clinical Medicine, Umea University, Umea, Sweden 66Division of Cardiovascular Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden 67Kuopio Research Institute of Exercise Medicine, Kuopio, Finland 68Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Dusseldorf, Dusseldorf, Germany 69German Center for Diabetes Research, Partner Site Dusseldorf, Germany 70HUNT Research Center, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger, Norway 71Cardiovascular Genetics, BHF Laboratories, Institute Cardiovascular Sciences, UCL, London, UK 72Program in Genetic Epidemiology and Statistical Genetics, Harvard School of Public Health, Boston, MA, USA 73Steno Diabetes Center, Gentofte, Denmark 74Research Centre for Prevention and Health, Capital Region of Denmark, Copenhagen, Denmark 75Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark 76Faculty of Medicine, University of Aalborg, Aalborg, Denmark 77Hannover Unified Biobank, Hannover Medical School, Hannover, Germany 78Institute of Human Genetics, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany 79Copenhagen University Hospital, Rigshospitalet, Denmark 80Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark 81Institute of Genetic Epidemiology, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany 82BRC Genomics Core Facility, GSTT Foundation Trust, Guy’s & St Thomas’ Hospital, London 83Institute of Human Genetics, University of Bonn, Bonn, Germany 84Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany 85Institute of Neuroscience and Medicine (INM-1), Research Centre Jülich, D-52425 Jülich, Germany 86Department of Medicine I, University Hospital Grosshadern, Ludwig-Maximilians-Universität, Munich, Germany 87Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Neuherberg, Germany 88DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany 89Department of Epidemiology, Murcia Regional Health Council, IMIB-Arrixaca, Murcia, Spain 90CIBER Epidemiología y Salud Pública (CIBERESP), Spain 91Department of Health and Social Sciences, Universidad de Murcia, Murcia, Spain 92Oxford National Institute for Health Research Biomedical Research Centre, Churchill Hospital, Oxford, UK 93Cancer Research and Prevention Institute (ISPO), Florence, Italy 94Department of Internal Medicine, Levanger Hospital, Nord-
Trondelag Health Trust, Levanger, Norway
Department of Endocrinology and Diabetology, University Hospital Dusseldorf, Dusseldorf, Germany
Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA
Boston University Data Coordinating Center, Boston, MA, USA
University Medical Center Utrecht, Utrecht, The Netherlands
Icelandic Heart Association, Kopavogur, Iceland
Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland
Department of Clinical Chemistry and Central Laboratory, University of Ulm, Ulm, Germany
Department of Public Health, Hjelt Institute, University of Helsinki, Helsinki, Finland
Department of Endocrinology, Abdominal Center, Helsinki University Hospital, Helsinki, Finland
Research Program for Diabetes and Obesity, University of Helsinki, Helsinki, Finland
Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands
Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK
European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK
Division of Endocrinology, Germans Trias i Pujol University Hospital and Research Institute, Badalona, Spain
Josep Carreras Leukaemia Research Institute, Badalona, Spain
CIBER de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Barcelona, Spain
Department of Medicine, Imperial College London, London, UK
Institut d’Investigacions Biomediques August Pi i Sunyer, Centre Esther Koplowitz, Barcelona, Spain
The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA
Department of Medicine, Harvard Medical School, Boston, MA, USA
Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA
Diabetes Research Center, Diabetes Unit, Massachusetts General Hospital, Boston, MA, USA
Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA
Human Genome Sequencing Center at Baylor College of Medicine, Houston, TX, USA
Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, MN, USA
Netherlands Genomics Initiative, Netherlands Consortium for Healthy Ageing and Center for Medical Systems Biology, Rotterdam, The Netherlands
General Medicine Division, Massachusetts General Hospital, MA, USA
Institute of Biomedicine/Physiology, University of Eastern Finland, Kuopio, Finland
Department of Clinical Physiology and Nuclear Medicine, Kuopio University Hospital, Kuopio, Finland
Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
Department of Medical Sciences, Uppsala University Hospital, Uppsala, Sweden
Faculty of Medicine, Institute of Health Sciences, University of Oulu, Oulu, Finland
Unit of General Practice, Oulu University Hospital, Oulu, Finland
South Ostrobothnia Central Hospital, Seinajoki, Finland
Finnish Diabetes Association, Tampere, Finland
Pirkanmaa District Hospital, Tampere, Finland
Department of Medicine, Central Finland Central Hospital, Jyvaskyla, Finland
Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia
Clinic of Cardiology, West German Heart Centre, University Hospital of Essen, University Duisburg-Essen, Essen,
Acknowledgments

Funding for the research undertaken in this study has been received from: Academy of Finland (including grant numbers 77299, 102318, 10493, 118065, 123885, 124243, 129293, 129680, 136895, 139635, 211119, 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; Association of Danish Pharmacies; Augustinus Foundation; Becket Foundation; British Diabetes Association (BDA) Research; British Heart Foundation; Central Norway Health Authority; Central Finland Hospital District; Center for Inherited Disease Research (CIDR); City of Kuopio; City of Leutkirch; Copenhagen County; the Danish Centre for Evaluation and Health Technology Assessment; Danish Council for Independent Research; Danish Heart Foundation; Danish Research Councils; Deutsche Forschungsgemeinschaft (including project ER 155/6-2); the Diabetes Research Foundation; Diabetes UK; Doris Duke Charitable Foundation; Erasmus Medical Center; Erasmus University; Estonian Government (SF0180142s08); European Commission (including ENGAGE HEALTH-F4-2007-201413, FP7-201413, FP7-245536, EXGENESIS LSHM-CT-2004-005272, FP6 LSHM_CT_2006_037197, LSHM-CT-2007-037273, Directorate C-Public Health 2004310, DG XII); European Regional Development Fund; Federal Ministry of Education and Research, Germany (including FKZ 01GI1128 and FKZ 01EO1001); Federal Ministry of Health, Germany; Finnish Diabetes Association; Finnish Diabetes Research Foundation; Finnish Foundation for Cardiovascular Research; Finnish Medical Society; Folkhalsan Research Gaulton et al. Page 21

Nat Genet. Author manuscript; available in PMC 2016 May 18.
References


Gaulton et al.  Page 22

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. There was stronger enrichment in: (i) FOXA2 sites overlapping a ChIP-seq site for another protein compared to unique sites; (ii) sites identified in primary islets compared to HepG2 or primary liver cells; and (iii) sites overlapping islet enhancers compared to those that did 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; at 15 signals this enrichment was nominally-significant ($p<0.05$). (D) FOXA2-bound variants disrupting recognition motifs have an increased probability of being causal.
Figure 2. The lone variant in the 99% credible set at the MTNR1B locus affects FOXA2-bound enhancer activity

(A) The intronic variant, rs10830963, has 99.8% probability of driving the association signal at the MTNR1B locus. This variant overlaps a FOXA2 binding site, and the risk allele G is predicted to create 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-βH1 cell extracts. Proteins were bound to both alleles. In the presence of a NEUROD1 antibody, only the risk allele band was super-shifted, and in the presence of an unlabelled NEUROD1 consensus probe, the signal was competed away. NE: nuclear extract.

Nat Genet. Author manuscript; available in PMC 2016 May 18.
(C, D) The 224bp sequence surrounding each allele was cloned into a luciferase reporter construct containing a minimal promoter and tested for luciferase activity in (C) EndoC-βH1 and (D) HepG2 cells (n=3 for each cell type). Results are presented as mean ± standard error. The risk allele had significantly increased enhancer activity over the protective allele in both forward and reverse orientations in both cell types.
Established T2D susceptibility loci with multiple distinct signals of association at locus-wide significance in the GCTA joint regression model ($p_{1}<10^{-5}$).

<table>
<thead>
<tr>
<th>Locus</th>
<th>Index variant</th>
<th>Chr</th>
<th>Position (b37)</th>
<th>Risk allele</th>
<th>Other allele</th>
<th>Metabochip GCTA joint model 27,206 cases and 57,574 controls</th>
<th>Validation GCTA joint model 19,662 cases and 115,140 controls</th>
<th>Combined GCTA joint model 46,868 cases and 172,714 controls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RAF</td>
<td>OR (95% CI)</td>
<td>$p_{1}$</td>
</tr>
<tr>
<td>DGKB</td>
<td>rs10276674</td>
<td>7</td>
<td>14,922,007</td>
<td>C</td>
<td>T</td>
<td>0.183</td>
<td>1.08 (1.04–1.11)</td>
<td>4.5×10^{-6}</td>
</tr>
<tr>
<td></td>
<td>rs1974620</td>
<td>7</td>
<td>15,065,467</td>
<td>T</td>
<td>C</td>
<td>0.519</td>
<td>1.06 (1.04–1.09)</td>
<td>1.6×10^{-6}</td>
</tr>
<tr>
<td>CDKN2B</td>
<td>rs10811660</td>
<td>9</td>
<td>22,134,068</td>
<td>G</td>
<td>A</td>
<td>0.830</td>
<td>1.32 (1.27–1.38)</td>
<td>2.4×10^{-44}</td>
</tr>
<tr>
<td></td>
<td>rs10757283</td>
<td>9</td>
<td>22,134,172</td>
<td>T</td>
<td>C</td>
<td>0.437</td>
<td>1.14 (1.10–1.17)</td>
<td>7.4×10^{-18}</td>
</tr>
<tr>
<td>KCNQ1</td>
<td>chr11:2692322D</td>
<td>11</td>
<td>2,692,322</td>
<td>D</td>
<td>R</td>
<td>0.374</td>
<td>1.08 (1.05–1.10)</td>
<td>3.5×10^{-8}</td>
</tr>
<tr>
<td></td>
<td>rs2283220</td>
<td>11</td>
<td>2,755,548</td>
<td>A</td>
<td>G</td>
<td>0.661</td>
<td>1.06 (1.03–1.09)</td>
<td>0.000016</td>
</tr>
<tr>
<td></td>
<td>rs2237895</td>
<td>11</td>
<td>2,857,194</td>
<td>C</td>
<td>A</td>
<td>0.428</td>
<td>1.08 (1.05–1.11)</td>
<td>6.6×10^{-7}</td>
</tr>
<tr>
<td></td>
<td>rs74046911</td>
<td>11</td>
<td>2,858,636</td>
<td>C</td>
<td>T</td>
<td>0.951</td>
<td>1.32 (1.24–1.40)</td>
<td>1.7×10^{-17}</td>
</tr>
<tr>
<td></td>
<td>rs480696</td>
<td>11</td>
<td>2,858,800</td>
<td>G</td>
<td>C</td>
<td>0.707</td>
<td>1.06 (1.03–1.10)</td>
<td>0.00026</td>
</tr>
<tr>
<td>HNF1A</td>
<td>rs1169288</td>
<td>12</td>
<td>121,416,650</td>
<td>C</td>
<td>A</td>
<td>0.334</td>
<td>1.10 (1.07–1.13)</td>
<td>5.4×10^{-30}</td>
</tr>
<tr>
<td></td>
<td>rs1800574</td>
<td>12</td>
<td>121,416,864</td>
<td>T</td>
<td>C</td>
<td>0.027</td>
<td>1.21 (1.11–1.31)</td>
<td>5.2×10^{-6}</td>
</tr>
<tr>
<td></td>
<td>chr12:121440833D</td>
<td>12</td>
<td>121,440,833</td>
<td>R</td>
<td>D</td>
<td>0.416</td>
<td>1.06 (1.03–1.09)</td>
<td>0.000028</td>
</tr>
<tr>
<td>MC4R</td>
<td>chr18:57739289D</td>
<td>18</td>
<td>57,739,289</td>
<td>D</td>
<td>R</td>
<td>0.234</td>
<td>1.05 (1.02–1.09)</td>
<td>0.00079</td>
</tr>
<tr>
<td></td>
<td>rs17006842</td>
<td>18</td>
<td>58,040,624</td>
<td>G</td>
<td>A</td>
<td>0.961</td>
<td>1.13 (1.06–1.21)</td>
<td>0.00033</td>
</tr>
<tr>
<td>GIPR</td>
<td>rs4399645</td>
<td>19</td>
<td>46,166,073</td>
<td>T</td>
<td>C</td>
<td>0.395</td>
<td>1.07 (1.04–1.10)</td>
<td>4.4×10^{-7}</td>
</tr>
<tr>
<td></td>
<td>rs2238689</td>
<td>19</td>
<td>46,178,661</td>
<td>C</td>
<td>T</td>
<td>0.425</td>
<td>1.09 (1.07–1.12)</td>
<td>9.7×10^{-12}</td>
</tr>
<tr>
<td>HNF4A</td>
<td>rs1800961</td>
<td>20</td>
<td>43,042,364</td>
<td>T</td>
<td>C</td>
<td>0.034</td>
<td>1.16 (1.09–1.24)</td>
<td>0.000011</td>
</tr>
</tbody>
</table>

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.

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 loci for which the 99% credible set contains no more than ten variants.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Index variant</th>
<th>Chr</th>
<th>Position (b37)</th>
<th>Risk allele</th>
<th>Other allele</th>
<th>RAF</th>
<th>p-value</th>
<th>OR (95% CI)</th>
<th>99% credible set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SNPs Interval (bp) Interval start (bp) Interval stop (bp)</td>
</tr>
<tr>
<td>MTNR1B</td>
<td>rs10830963</td>
<td>11</td>
<td>92,708,710</td>
<td>G</td>
<td>C</td>
<td>0.283</td>
<td>2.9×10^{-12}</td>
<td>1.10 (1.07–1.13)</td>
<td>1  1  92,708,710 92,708,710</td>
</tr>
<tr>
<td>TCF7L2</td>
<td>rs7903146</td>
<td>10</td>
<td>114,758,349</td>
<td>T</td>
<td>C</td>
<td>0.260</td>
<td>5.8×10^{-13}</td>
<td>1.39 (1.35–1.43)</td>
<td>3  4,279 114,754,071 114,758,349</td>
</tr>
<tr>
<td>KCNQ1</td>
<td>rs74046911</td>
<td>11</td>
<td>2,858,636</td>
<td>C</td>
<td>T</td>
<td>0.951</td>
<td>5.9×10^{-18}</td>
<td>1.33 (1.25–1.42)</td>
<td>3  197 2,858,440 2,858,636</td>
</tr>
<tr>
<td>ZBED3</td>
<td>rs7732130</td>
<td>5</td>
<td>76,435,004</td>
<td>G</td>
<td>A</td>
<td>0.278</td>
<td>6.4×10^{-10}</td>
<td>1.09 (1.06–1.12)</td>
<td>5  10,056 76,424,949 76,435,004</td>
</tr>
<tr>
<td>CDKN2A-B</td>
<td>rs10757283</td>
<td>9</td>
<td>22,134,172</td>
<td>T</td>
<td>C</td>
<td>0.437</td>
<td>2.8×10^{-19}</td>
<td>1.14 (1.11–1.18)</td>
<td>5  1,007 22,133,645 22,134,651</td>
</tr>
<tr>
<td>SLC30A8</td>
<td>rs13266634</td>
<td>8</td>
<td>118,184,783</td>
<td>C</td>
<td>T</td>
<td>0.676</td>
<td>1.3×10^{-18}</td>
<td>1.13 (1.10–1.16)</td>
<td>6  33,133 118,184,783 118,217,915</td>
</tr>
<tr>
<td>CDKN2A-B</td>
<td>rs10811660</td>
<td>9</td>
<td>22,134,068</td>
<td>G</td>
<td>A</td>
<td>0.830</td>
<td>7.0×10^{-43}</td>
<td>1.32 (1.27–1.37)</td>
<td>6  1,397 22,132,698 22,134,094</td>
</tr>
<tr>
<td>HNF1B</td>
<td>rs4430796</td>
<td>17</td>
<td>36,098,040</td>
<td>G</td>
<td>A</td>
<td>0.455</td>
<td>6.3×10^{-12}</td>
<td>1.09 (1.07–1.12)</td>
<td>7  5,791 36,097,775 36,103,565</td>
</tr>
<tr>
<td>CDKAL1</td>
<td>rs35261542</td>
<td>6</td>
<td>20,675,792</td>
<td>A</td>
<td>C</td>
<td>0.280</td>
<td>9.6×10^{-23}</td>
<td>1.15 (1.12–1.18)</td>
<td>8  30,073 20,673,880 20,703,952</td>
</tr>
<tr>
<td>GLIS3</td>
<td>chr9:4294707:1</td>
<td>9</td>
<td>4,294,707</td>
<td>I</td>
<td>R</td>
<td>0.360</td>
<td>6.5×10^{-4}</td>
<td>1.07 (1.05–1.10)</td>
<td>10 15,453 4,283,137 4,298,589</td>
</tr>
</tbody>
</table>

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.
## Table 3

Motif-altering credible set variants in FOXA2 sites.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Index variant</th>
<th>Motif-altering variant</th>
<th>Chr</th>
<th>Position (b37)</th>
<th>Posterior probability ($\pi_C$)</th>
<th>Motif-altering allele</th>
<th>Chromatin state</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTNR1B</td>
<td>rs10830963</td>
<td>rs10830963</td>
<td>11</td>
<td>92708710</td>
<td>0.998</td>
<td>G</td>
<td>Islet-enhancer, HepG2-enhancer</td>
</tr>
<tr>
<td>TCF7L2</td>
<td>rs7903146</td>
<td>rs7903146</td>
<td>10</td>
<td>114,788,349</td>
<td>0.78</td>
<td>T</td>
<td>Islet-enhancer</td>
</tr>
<tr>
<td>SLC30A8</td>
<td>rs13266634</td>
<td>rs13266634</td>
<td>8</td>
<td>118,184,783</td>
<td>0.29</td>
<td>T</td>
<td>Islet-enhancer</td>
</tr>
<tr>
<td>CDKN2B</td>
<td>rs10811660</td>
<td>rs10811660</td>
<td>9</td>
<td>22,134,068</td>
<td>0.24</td>
<td>A</td>
<td>Islet-enhancer</td>
</tr>
<tr>
<td>CDC123</td>
<td>rs11257658</td>
<td>rs11257655</td>
<td>10</td>
<td>12,307,894</td>
<td>0.21</td>
<td>T</td>
<td>Islet-enhancer, HepG2-enhancer</td>
</tr>
<tr>
<td>JAZF1</td>
<td>rs1513272</td>
<td>rs849133</td>
<td>7</td>
<td>28,192,280</td>
<td>0.042</td>
<td>T</td>
<td>Islet-enhancer</td>
</tr>
<tr>
<td>KCNQ1</td>
<td>rs2283220</td>
<td>rs231907</td>
<td>11</td>
<td>2,752,130</td>
<td>0.031</td>
<td>T</td>
<td>HepG2-enhancer</td>
</tr>
<tr>
<td>FTO</td>
<td>rs9927317</td>
<td>rs9940128</td>
<td>16</td>
<td>53,800,754</td>
<td>0.027</td>
<td>G</td>
<td>Islet-enhancer, HepG2-enhancer</td>
</tr>
<tr>
<td>FTO</td>
<td>rs9927317</td>
<td>rs9939973</td>
<td>16</td>
<td>53,800,568</td>
<td>0.025</td>
<td>G</td>
<td>Islet-enhancer, HepG2-enhancer</td>
</tr>
<tr>
<td>KCNQ1</td>
<td>rs458069</td>
<td>rs78688069</td>
<td>11</td>
<td>2,752,183</td>
<td>0.0006</td>
<td>A</td>
<td>HepG2-enhancer</td>
</tr>
<tr>
<td>KCNQ1</td>
<td>rs458069</td>
<td>rs190728714</td>
<td>11</td>
<td>2,813,084</td>
<td>0.00042</td>
<td>G</td>
<td>Islet-enhancer</td>
</tr>
<tr>
<td>DGKB</td>
<td>rs10276674</td>
<td>rs7798360</td>
<td>7</td>
<td>15,055,972</td>
<td>0.00005</td>
<td>G</td>
<td>-</td>
</tr>
</tbody>
</table>

Chr: chromosome.