# Genetic regulatory signatures underlying islet gene expression and type 2 diabetes

Arushi Varshney[a,1], Laura J. Scott[b,1], Ryan P. Welch[b,1], Michael R. Erdos[c,1], Peter S. Chines[c], Narisu Narisu[c], Ricardo D'O. Albanus[d], Peter Orchard[d], Brooke N. Wolford[d], Romy Kursawe[e], Swarooparani Vadlamudi[f], Maren E. Cannon[f], John P. Didion[c], John Hensley[d], Anthony Kirilusha[c], NISC Comparative Sequencing Program[g,2], Lori L. Bonnycastle[c], D. Leland Taylor[c,h], Richard Watanabe[i,j], Karen L. Mohlke[f], Michael Boehnke[b,1], Francis S. Collins[c,1,3], Stephen C. J. Parker[a,d,1,3], and Michael L. Stitzel[e,1]

[a]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109; [b]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109; [c]National Human Genome Research Institute, NIH, Bethesda, MD 20892; [d]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI 48109; [e]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032; [f]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599; [g]NIH Intramural Sequencing Center, National Human Genome Research Institute, NIH, Bethesda, MD 20892; [h]European Molecular Biology Laboratory, Wellcome Trust Genome Campus, European Bioinformatics Institute, Hinxton, Cambridgeshire CB10 1SD, United Kingdom; [i]Department of Preventive Medicine, University of Southern California Keck School of Medicine, Los Angeles, CA 90089; and [j]Department of Physiology and Biophysics, University of Southern California Keck School of Medicine, Los Angeles, CA 90089

Contributed by Francis S. Collins, December 31, 2016 (sent for review August 11, 2016; reviewed by Bradley E. Bernstein and Andrew S. McCallion)

**Genome-wide association studies (GWAS) have identified >100 independent SNPs that modulate the risk of type 2 diabetes (T2D) and related traits. However, the pathogenic mechanisms of most of these SNPs remain elusive. Here, we examined genomic, epigenomic, and transcriptomic profiles in human pancreatic islets to understand the links between genetic variation, chromatin landscape, and gene expression in the context of T2D. We first integrated genome and transcriptome variation across 112 islet samples to produce dense *cis*-expression quantitative trait loci (*cis*-eQTL) maps. Additional integration with chromatin-state maps for islets and other diverse tissue types revealed that *cis*-eQTLs for islet-specific genes are specifically and significantly enriched in islet stretch enhancers. High-resolution chromatin accessibility profiling using assay for transposase-accessible chromatin sequencing (ATAC-seq) in two islet samples enabled us to identify specific transcription factor (TF) footprints embedded in active regulatory elements, which are highly enriched for islet *cis*-eQTL. Aggregate allelic bias signatures in TF footprints enabled us de novo to reconstruct TF binding affinities genetically, which support the high-quality nature of the TF footprint predictions. Interestingly, we found that T2D GWAS loci were strikingly and specifically enriched in islet Regulatory Factor X (RFX) footprints. Remarkably, within and across independent loci, T2D risk alleles that overlap with RFX footprints uniformly disrupt the RFX motifs at high-information content positions. Together, these results suggest that common regulatory variations have shaped islet TF footprints and the transcriptome and that a confluent RFX regulatory grammar plays a significant role in the genetic component of T2D predisposition.**

chromatin | diabetes | eQTL | epigenome | footprint

Type 2 diabetes (T2D) is a complex disease characterized by pancreatic islet dysfunction and insulin resistance in peripheral tissues; >90% of T2D SNPs identified through genome-wide association studies (GWASs) reside in nonprotein coding regions and are likely to perturb gene expression rather than alter protein function (1). In support of this finding, we and others recently showed that T2D GWAS SNPs are significantly enriched in enhancer elements that are specific to pancreatic islets (2–4). The critical next steps to translate these islet enhancer T2D genetic associations into mechanistic biological knowledge are (*i*) identifying the putative functional SNP(s) from all of those that are in tight linkage disequilibrium (LD), (*ii*) localizing their target gene(s), and (*iii*) understanding the direction of effect (increased or decreased target gene expression) conferred by the risk allele. Two recent studies analyzed genome variation and gene expression variation across human islet samples to identify *cis*-expression quantitative trait loci (*cis*-eQTLs) that linked T2D GWAS SNPs to target genes (5, 6). However, the transcription factor (TF) molecular mediators of the islet *cis*-eQTLs remain poorly understood and represent important links to upstream pathways that will help untangle the regulatory complexity of T2D.

## Results

**Integrated Analysis of Islet Transcriptome and Epigenome Data.** To build links between SNP effects on regulatory element use and gene expression in islets, we performed strand-specific mRNA sequencing of 31 pancreatic islet tissue samples (Table S1) to an average depth of 100 million paired end reads. In parallel, we

### Significance

**The majority of genetic variants associated with type 2 diabetes (T2D) are located outside of genes in noncoding regions that may regulate gene expression in disease-relevant tissues, like pancreatic islets. Here, we present the largest integrated analysis to date of high-resolution, high-throughput human islet molecular profiling data to characterize the genome (DNA), epigenome (DNA packaging), and transcriptome (gene expression). We find that T2D genetic variants are enriched in regions of the genome where transcription Regulatory Factor X (RFX) is predicted to bind in an islet-specific manner. Genetic variants that increase T2D risk are predicted to disrupt RFX binding, providing a molecular mechanism to explain how the genome can influence the epigenome, modulating gene expression and ultimately T2D risk.**

Author contributions: A.V., L.J.S., M.R.E., R.W., K.L.M., M.B., F.S.C., S.C.J.P., and M.L.S. designed research; A.V., L.J.S., R.P.W., M.R.E., R.K., S.V., N.C.S.P., S.C.J.P., and M.L.S. performed research; P.O. and J.H. contributed new reagents/analytic tools; A.V., L.J.S., R.P.W., P.S.C., N.N., R.D.A., P.O., B.N.W., M.E.C., J.P.D., A.K., L.L.B., D.L.T., and S.C.J.P. analyzed data; and A.V., L.J.S., S.C.J.P., and M.L.S. wrote the paper.

Reviewers: B.E.B., Harvard Medical School, Broad Institute; and A.S.M., Johns Hopkins University School of Medicine.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the dbGaP (accession no. phs001188.v1.p1; FUSION Tissue Biopsy Study—Islet Expression and Regulation by RNAseq and ATACseq).

[1]A.V., L.J.S., R.P.W., M.R.E., M.B., F.S.C., S.C.J.P., and M.L.S. contributed equally to this work.

[2]A complete list of the NISC Comparative Sequencing Program can be found in *SI Materials and Methods*.

[3]To whom correspondence may be addressed. Email: collinsf@od.nih.gov or scjp@umich.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1621192114/-/DCSupplemental.

analyzed unstranded mRNA sequencing (mRNA-seq) data for 81 islet samples from a previous study (5). We subjected both datasets to the same quality control and processing. We additionally completed dense genotyping of 31 islet samples and downloaded genotypes for 81 previously described islet samples (5). Phasing and imputation yielded a final set of 6,060,203 autosomal SNPs present in both datasets with an overall minor allele count >10. To identify SNPs affecting gene expression within 1 Mb of the most upstream transcription start site (TSS), we performed separate *cis*-eQTL analyses for the two sets of islet samples and combined the *cis*-eQTL results via meta-analysis. We identified 3,964 unique autosomal *cis*-eQTL lead SNPs for 3,993 genes at a 5% false discovery rate (FDR).

Next, we integrated chromatin immunoprecipitation followed by sequencing (ChIP-seq) data for five histone modifications across islets (2, 7) and 30 diverse tissues with publicly available datasets (Table S2) (8–10) using ChromHMM (9). This analysis produced 13 unique and recurrent chromatin states (Fig. 1*A* and Fig. S1), including promoter, enhancer, transcribed, and repressed regions. To identify specific regulatory element sites within these chromatin states, we profiled open chromatin in two islets using the assay for transposase-accessible chromatin sequencing (ATAC-seq) (11) (Fig. 1*A* and Table S1). Our high-depth ATAC-seq data (>1.4 billion reads for both islets) allowed us to identify TF DNA footprints using the CENTIPEDE algorithm (12). We assigned regulatory state and TF footprint status to every islet *cis*-eQTL based on the annotation of SNPs with $r^2 > 0.8$ with the lead SNP (Fig. 1*B*). We used iterative conditional analyses (7) to identify 28 T2D and related quantitative trait GWAS SNPs that could be islet *cis*-eQTL signals (Fig. 1*C* and Datasets S1 and S2). Given the modest *cis*-eQTL signals at most of these loci, conditional analysis in larger islet samples will likely change this list.

As an example, T2D GWAS index SNP rs1535500 occurs at the *KCNK16* locus, and the risk allele results in a glutamate substitution at alanine 277. This change was implicated in increasing the KCNK16 basal channel activity and cell surface

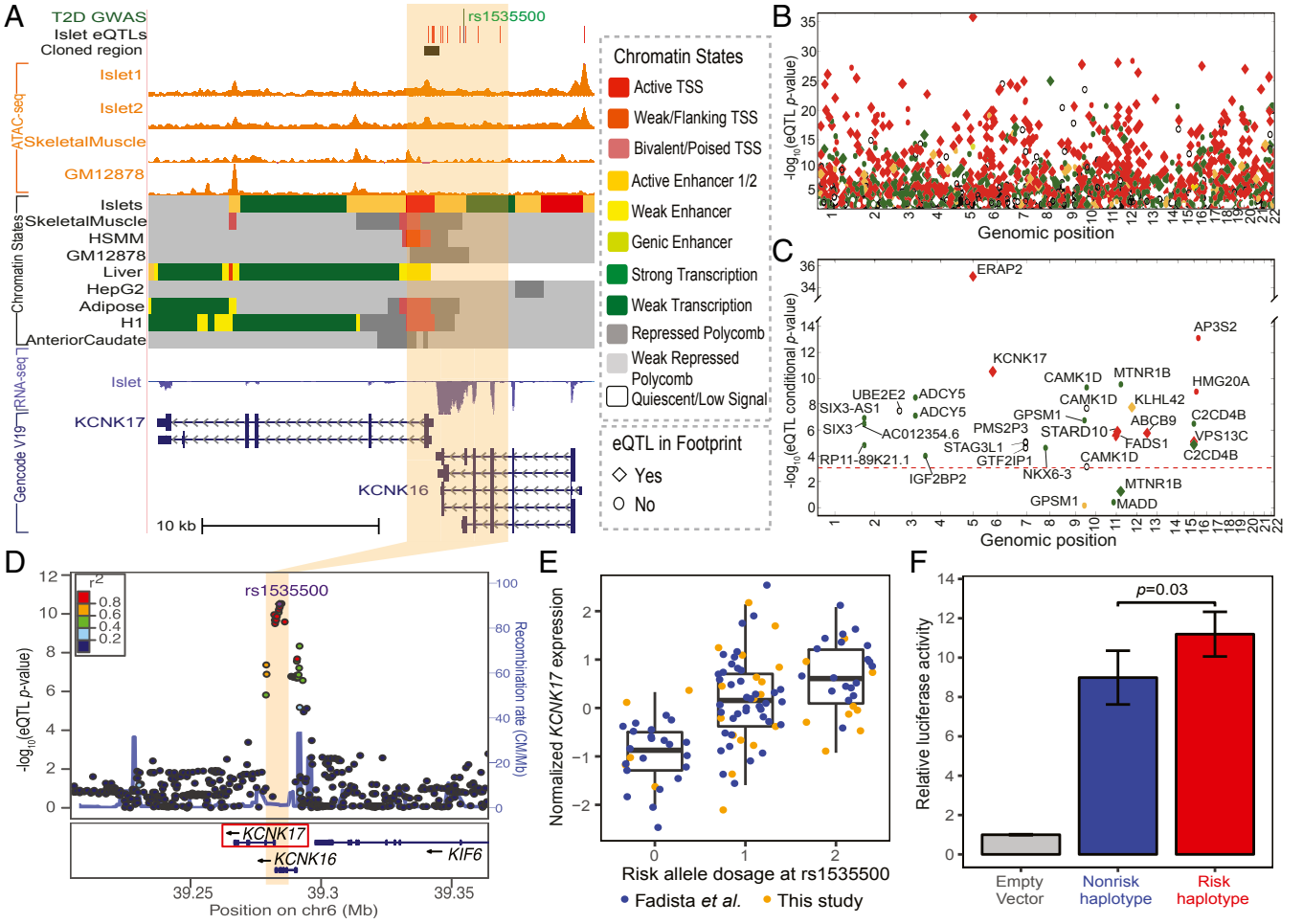



**Fig. 1.** Integrated genomic, epigenomic, and transcriptomic analyses of human pancreatic islets. (*A*) An overview of diverse molecular profiling data types used in this study. Integrative molecular profiling (open chromatin, ATAC-seq; chromatin states; RNA-seq) highlights islet-specific signatures at the *KCNK17* locus. (*B*) Plot of strength of association (*y* axis) for significant islet *cis*-eQTLs colored by chromatin-state annotation (*A*) by chromosomal location (*x* axis); diamonds indicate SNPs overlapping ATAC-seq footprints. An interactive version of this plot can be found at theparkerlab.org/tools/isleteqtl/. (*C*) Plot of strength of islet *cis*-eQTL association for T2D and related trait GWAS SNPs after conditional analysis to identify variants likely independent of stronger *cis*-eQTL signals for the same gene by chromosomal position and annotated as in *B*. The plot includes all GWAS SNP–gene pairs with FDR < 0.05 in original *cis*-eQTL analysis. The dotted red line represents the *P* value threshold for FDR < 0.05 based on the conditional analysis. (*D*) Islet *cis*-eQTL associated with *KCNK17* expression highlighted for comparison with molecular profiling tracks in *A*. (*E*) Plot of normalized *KCNK17* expression in islet samples and *cis*-eQTL risk allele dosage. (*F*) Functional validation of *KCNK17 cis*-eQTL at its promoter region. The haplotype containing alleles associated with T2D risk and increased *KCNK17* expression (rs10947804-C, rs12663159-A, rs146060240-G, and rs34247110-A) shows higher transcriptional activity than the haplotype with nonrisk alleles. The cloned region is indicated at the top of *A*. Relative luciferase activity is given as mean ± SD of four to five independent clones per haplotype normalized to empty vector. Significance was evaluated using a two-sided *t* test.

localization when tested in a mouse model (13). Our analysis revealed that rs1535500 is not associated with *KCNK16* expression (Fig. S2). Interestingly, the rs1535500 risk allele is associated with increased expression of the neighboring potassium channel gene *KCNK17* (Fig. 1 *D* and *E*); rs1535500 is in high LD ($r^2 > 0.95$), with four SNPs (rs10947804, rs12663159, rs146060240, and rs34247110) that are located in an islet promoter chromatin state, and all but rs34247110 are located in an ATAC-seq peak (Fig. 1*A*). Motivated by the overlap with islet regulatory annotations, we cloned two different copies of the 473-bp DNA sequence surrounding these SNPs: one containing the T2D risk alleles for each of four SNPs (risk haplotype) and the other containing the nonrisk alleles (nonrisk haplotype). We performed luciferase reporter assays in the mouse insulinoma (MIN6) beta cell line to test the transcriptional activity of these two clones. Both clones exhibited promoter activity, but the T2D risk haplotype showed significantly greater ($P = 0.03$) transcriptional activity than the nonrisk haplotype (Fig. 1*F*). This result suggests that one or more of these T2D risk variants cause increased regulatory activity in islets. These findings highlight a complex functional genetic architecture for a single haplotype that results in regulatory activity linked to one gene (*KCNK17*) and coding variation in another (*KCNK16*). Together, these results illustrate how integrated analyses help to identify potential causal SNPs associated with islet expression and T2D risk. To enable easy, in-depth exploration of our results, we created an interactive islet *cis*-eQTL and chromatin-state browser (theparkerlab.org/tools/isleteqtl/).

**Common and Islet-Specific Gene *cis*-eQTLs Are Enriched in Different Chromatin States.** To understand the regulatory architecture of islet *cis*-eQTLs, we measured their co-occurrence with different classes of chromatin states across diverse tissues, including stretch enhancers, defined as enhancer chromatin states ≥3 kb long. These segments tend to mark cell identity regions and have been shown to harbor tissue-specific GWAS SNPs (2, 14). We calculated genome-wide enrichment for *cis*-eQTL overlaps with these features while controlling for minor allele frequency, distance to TSS, and the number of SNPs in LD (15). *cis*-eQTLs were enriched in active chromatin states, such as promoter, and genic enhancer in islets, whereas inactive states, such as polycomb repressed, were depleted for such overlaps across multiple tissues (Fig. S3). Reasoning that this common enrichment pattern across diverse tissues may be largely driven by *cis*-eQTLs of commonly expressed genes, we sought to classify *cis*-eQTLs by the islet expression specificity of their associated genes. To measure gene expression specificity in islets, we analyzed RNA-seq data from 16 additional tissues from the Illumina Human Body Map 2.0 project. We used an information theory approach to define the islet expression specificity index (iESI) (Fig. S4) (7). iESI values near zero represent lowly and/or ubiquitously expressed genes, whereas values near one represent genes that are highly and specifically expressed in islets. We divided genes into quintiles based on ascending iESI (Fig. S4). We assigned *cis*-eQTLs for these genes to their respective iESI quintile and measured enrichment of each set in chromatin annotations. Interestingly, although *cis*-eQTLs across iESI quintile bins were similarly enriched in islet promoter states, *cis*-eQTL enrichment in active and stretch enhancer states increased concomitantly with iESI (Fig. S5). As an example, we found that the *cis*-eQTL for the *KCNA6* gene (Fig. S6*A*), which is expressed in islets with high specificity (iESI = 0.78), overlapped islet-specific enhancer states (Fig. S6*B*). This *cis*-eQTL does not overlap a known T2D GWAS locus. When we restricted our enrichment analysis to ATAC-seq peaks in islet stretch enhancer states, we saw a stronger trend toward increasing enrichment by iESI quintile (Fig. S5). These results indicate a strong link between active regulatory chromatin architecture and the genetic control of cell-specific gene expression.

To further identify and dissect regulatory regions critical for islet-specific gene expression, we sought to distinguish between shared and tissue-specific enhancer chromatin states. We performed *k*-means clustering for active enhancer chromatin states across 31 cells/tissues. This method segregated enhancer regions based on activity across diverse tissues; for example, cluster 13 is islet-specific, whereas cluster 3 is liver-specific (Fig. S6*C*). We compared these enhancer clusters with stretch enhancer annotations across tissues and found that tissue-specific clusters, such as the islet-specific cluster 13, indeed displayed high enrichment for islet stretch enhancers (Fig. S6*D*). Likewise, in other tissues, tissue-specific enhancer clusters were enriched for the corresponding tissues' stretch enhancers (Fig. S6*D*). Next, we asked if islet *cis*-eQTLs were enriched in specific enhancer clusters and observed enrichment in multiple clusters (Fig. S6*E*). We then stratified the *cis*-eQTLs by iESI quintile and repeated this analysis. Notably, islet *cis*-eQTLs for genes in iESI quintile 5 only showed significant enrichment in the islet-specific enhancer cluster 13 ($P$ value = $1.2 \times 10^{-8}$, fold enrichment = 1.91) (Fig. S6*E*). Together, these results show that islet tissue-specific genetic regulatory architecture is enriched in islet-specific enhancers and stretch enhancers.

**Islet Expression Quantitative Trait Loci Are Enriched in Islet ATAC-Seq Peaks and DNA Footprints.** Chromatin-state maps identify regulatory regions, such as promoters and enhancers, but lack the resolution to pinpoint specific sites that may be bound and regulated by a TF. To refine the link between genetic variation, TF binding sites, and gene expression, we leveraged the high-resolution ATAC-seq data to identify in vivo putative TF binding sites using CENTIPEDE as previously described (7, 12). This approach detected high-quality footprints for many TFs, including the general CCCTC-binding factor (CTCF) and the TF Regulatory Factor X (RFX) (Fig. 2 *A* and *B*). Notably, we detect RFX footprints in islet stretch enhancers near the islet-specific (iESI = 0.94) TF *RFX6* (Fig. 2*A*), suggesting an autoregulatory mechanism that, based on recent studies (3, 16), may indicate that *RFX6* is an islet core transcriptional regulatory gene. Comparing ATAC-seq profiles from islets with those of skeletal muscle tissue (7), adipose tissue (17), and a lymphoblastoid cell line (GM12878) (11), we found that islet ATAC-seq peaks occurred preferentially in islet promoter and enhancer chromatin states (Fig. S7). Islet *cis*-eQTLs were highly enriched in multiple TF footprint motifs but were not in nonfootprint motifs (Fig. 2*C* and Dataset S3). These results suggest a strong link between SNPs at TF binding sites in relevant tissues and gene regulation.

To detect motif occurrences that could be altered by the presence of nonreference alleles, we developed a personalized phased SNP-aware genome motif scanning procedure (*SI Materials and Methods*). This method allowed us to identify motif instances, even when multiple nonreference alleles occur within a few base pairs of each other. We observed significant enrichment for islet *cis*-eQTLs in the set of TF footprint motifs identified only from this haplotype phase-aware scanning approach (that is, the motifs are missed even when a single SNP-aware motif scanning approach is used) in both islet samples (Fig. S8). Given the informative chromatin accessibility allelic analyses in recent studies (18, 19), we next asked if we could recreate known TF position weight matrices (PWMs) (Fig. 2*D*, row 1) based on the allele-specific bias at heterozygous SNPs within TF footprint motifs. We identified every heterozygous site in a given TF footprint motif, calculated the allelic bias in ATAC-seq signal at these positions, and retained all SNPs with significant bias (Fig. 2*D*, row 3 and *SI Materials and Methods*). We genetically reconstructed a PWM using the degree of allelic bias for the overrepresented alleles (Fig. 2*D*, row 2). This allelic bias-based PWM (Fig. 2*D*, row 2) closely matched the canonical PWM for the corresponding TF (Fig. 2*D*, row 1), providing an in vivo verification of the cognate PWM. There was a larger difference in the PWM score for the two alleles of allelic bias SNPs than for the two alleles of matched the 1000 Genomes Project (1000G) SNPs occurring in the same motif (Fig. S9). To further verify that the allelic bias-based genetically reconstructed PWMs were not

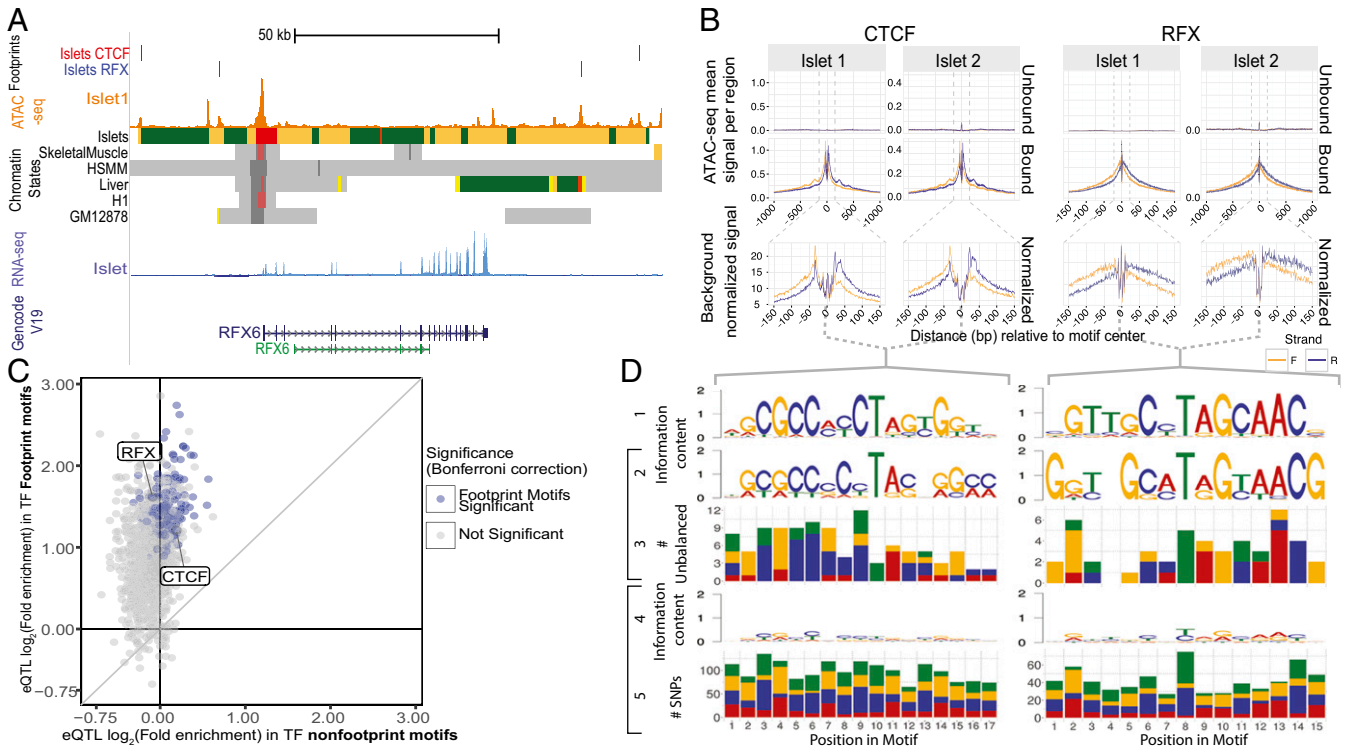



**Fig. 2.** Nucleotide resolution islet ATAC-seq profiling nominates regulatory mechanisms. (*A*) *RFX6* locus with expression (RNA-seq), chromatin states, open chromatin (ATAC-seq), and footprints for CTCF and RFX in islets. (*B*) Density plots indicating normalized sequence coverage of ATAC-seq from two human islet samples at sites overlapping CTCF (motif = CTCF_known2) and RFX (motif = RFX2_4) motifs. (*C*) Log twofold enrichment of islet *cis*-eQTLs in TF footprint motifs compared with their enrichment in TF nonfootprint motifs. TFs for which footprint and nonfootprint motifs overlap four or more eQTL SNPs are shown. Blue shows significant enrichment in footprints only (Bonferroni corrected $P < 0.05$). No significant enrichment was observed in any TF nonfootprint motif. (*D*) Reconstruction of CTCF (motif = CTCF_known2) and RFX (motif = RFX2_4) motifs using ATAC-seq TF footprint allelic bias data. Row 1: original motif PWM. Row 2: PWM genetically reconstructed using the overrepresented alleles (and extent of overrepresentation) for SNPs with significant ATAC-seq allelic bias. Row 3: count of nucleotides in SNPs with significant allelic bias. Row 4: PWM reconstructed using the count of nucleotides for heterozygous SNPs in the TF footprint. Row 5: count of nucleotides in heterozygous SNPs in the TF footprint.

simply reflecting the allelic composition of SNPs in the motifs, we constructed PWMs using the allele count for all TF footprint heterozygous SNPs observed at each position (where each observed SNP contributed two alleles) and found that the resulting PWMs had little information and little similarity to the cognate motifs used to scan across the genome (Fig. 2*D*, rows 4 and 5). Collectively, these results reinforce the potential of ATAC-seq and allelic footprinting analyses to identify relevant and potentially causal TF binding changes in the genetic control of gene expression.

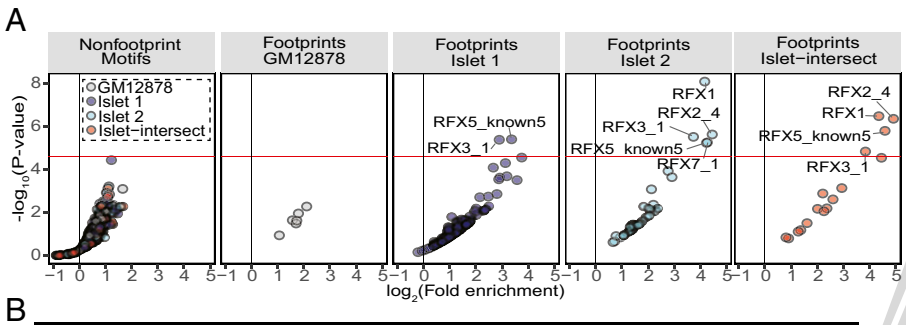
**T2D GWAS Loci Are Enriched in RFX Footprints, and T2D Risk Alleles Disrupt the Motifs at Independent Locations.** Given the strong enrichment for islet *cis*-eQTL in diverse TF footprints, we next sought to identify T2D GWAS SNPs that could regulate gene expression by modulating TF binding. We found that T2D-associated SNPs were significantly enriched in islet RFX TF footprints (Fig. 3*A* and Dataset S4). In contrast, we did not see significant enrichment of T2D-associated SNPs in islet nonfootprint RFX TF motifs or GM12878 TF footprints (Fig. 3*A*). The RFX family of TFs recognizes X-box motifs and has highly evolutionarily conserved DNA binding domains (20), which may explain why similar motifs from many RFX family members are enriched. A recent study found enrichment of T2D GWAS SNPs in islet FOXA2 ChIP-seq peaks (21). We observed enrichment of T2D-associated SNPs in islet FOX TF footprints, although none passed the Bonferroni threshold of $2.5 \times 10^{-5}$ (Dataset S4).

Studies of autoimmune disease have found that disease-associated variants often occur near but not in TF motifs (22). We, therefore, asked if T2D-associated SNPs were enriched in regions flanking RFX footprints motifs ($n = 22$). We found that regions flanking RFX footprint motifs were enriched for T2D-associated SNPs and that the enrichment decreased with increasing distance from footprint motifs (Fig. S10). The flanking enrichment was lower than in the RFX TF footprints. In contrast, we did not see enrichment of T2D-associated SNPs in nonfootprint RFX TF motifs or the regions flanking the nonfootprint RFX TF motifs (Fig. S10).

We next assessed the potential effects of the risk and nonrisk alleles for nine T2D-associated SNPs at five independent loci on RFX TF binding (Fig. 3*B*). For each SNP, the nonrisk allele was the highest probability nucleotide in the RFX PWM, and thus, the risk allele was predicted to disrupt the motif (Fig. 3 *B* and *C*, black boxes). At two of five loci, the T2D GWAS risk alleles were associated with significantly increased gene expression in our conditional eQTL analysis: *KCNK17* (*KCNK16* locus) (Fig. 1 *B*, *C*, and *E*) and *ABCB9* (*PITPNM2* locus) (Fig. 1*C*). Other loci might not have been detectable as *cis*-eQTLs because of state-specific regulation or small effect sizes. The observation that T2D risk alleles at multiple loci confluently disrupt RFX footprint motifs provides a hypothesis that could explain the mechanism of a subset of T2D-associated variants.

## Discussion

We have integrated genome, epigenome, and transcriptome variation and created maps to better understand the genetic control of islet gene expression. Comparison of these maps with T2D GWAS SNPs has helped identify potential disease mechanisms. For example, the risk allele of the coding SNP rs1535500 has been implicated to increase KCNK16 activity and cell surface




B

| T2D GWAS locus | T2D lead SNP | LD SNP | Lead-LD SNP $r^2$ | Chromatin State* | Overlapping RFX Footprint | Base in motif at SNP position† | Non-risk allele | Risk Allele |
|---|---|---|---|---|---|---|---|---|
| *KCNK16* | rs1535500 | rs10947804 | 0.967 | Act. TSS | RFX2_4 | A | A | G |
| *MPHOSPH9* | rs1727313 | rs655293 | 0.948 | Act. Enh 1 | RFX7_1 | A | A | G |
| | | rs1617156 | 0.948 | Act. Enh 2 | RFX2_4 | C | C | T |
| | | rs1716165 | 1 | Act. Enh 2 | RFX5_known1 | A | A | G |
| *WFS1* | rs4458523 | rs5018647 | 1 | Wk. Transc. | RFX5_disc1 | A | A | C |
| | | rs4689391 | 1 | Act. Enh. 2 | RFX5_known9 | G | G | A |
| *CDKN2A/B* | rs10811661 | rs10811660 | 0.834 | Wk. Transc. | RFX5 | A | A | G |
| *ZFAND3* | rs9470794 | rs57995712 | 1 | Act. Enh. 2 | RFX1 | T | T | G |
| | | rs9470802 | 0.915 | Wk. Transc. | RFX3_1 | C | C | G |

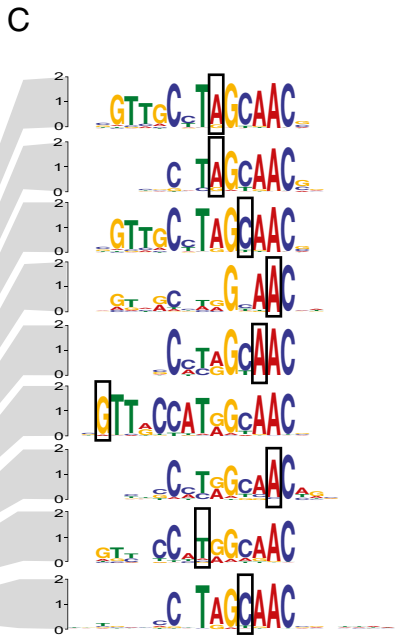



**Fig. 3.** T2D GWAS enrichment at islet footprints reveals confluent RFX motif disruption. (*A*) T2D GWAS SNPs are significantly enriched in RFX motifs in islet footprints but not in control motifs or footprints from a nondisease-relevant cell type (GM12878). TF motifs for which footprints overlap four or more T2D GWAS SNPs are shown. The red line indicates Bonferroni multiple testing threshold. (*B*) T2D-associated SNPs that overlap high information content (>1 bit) positions in RFX motifs. The highest scoring RFX footprints are reported for each T2D GWAS SNP. Act. Enh., active enhancer; Act. TSS, active TSS; Wk. Transc., weak transcribed. *Chromatin-state annotation overlapping the SNP. †Because RFX motifs in *C* are organized by alignment to the longest RFX3_1 motif, motifs overlapping rs10947804 and rs1716165 correspond to the reverse complement sequence. Therefore, risk and nonrisk alleles are also reported as reverse complement relative to the plus strand sequence. (*C*) Alignment of highest scoring RFX footprint at each SNP; the boxes indicate the SNP overlap positions. Note that, in every case, the risk allele disrupts that motif.

localization in a mouse model (13). Other risk alleles in SNPs in high LD with rs153550 are associated with increased expression of the neighboring potassium channel gene *KCNK17*, which is not in the mouse genome. KCNK16 and KCNK17 are two pore domain "background" $K^+$ channels, members of the TWIK-related alkaline pH-activated $K^+$ channel family (23, 24). Both genes are expressed in islets with high specificity (*KCNK16* iESI = 0.98; *KCNK17* iESI = 0.76). *KCNK16* has been implicated in regulating electrical excitability and glucose-stimulated insulin secretion (GSIS) (13). It is possible that the T2D risk haplotype at this locus may have multiple effects that collectively disrupt islet $K^+$ signaling and GSIS by simultaneously overactivating *KCNK16* and overexpressing *KCNK17*.

We find that T2D GWAS-associated SNPs are significantly enriched in RFX TF footprint motifs. We find consistent disruption of islet RFX footprint motifs by T2D risk alleles, including at the *KCNK17* locus. Lizio et al. (25) found that knockdown of *RFX6* results in increased expression of *KCNK17*, which is consistent with the T2D risk allele disrupting TF binding and increasing target gene expression. At other T2D GWAS loci, such as the *MPHOSPH9* locus (index SNP rs1727313), two or three T2D GWAS SNPs in high LD are each predicted to have risk alleles that coordinately disrupt independent RFX footprint motifs (Fig. 3 *B* and *C*). We and others (2, 26, 27) previously described the presence of multiple SNPs in enhancers at individual GWAS loci. Our results build on this concept to include the possibility of multiple confluent disruptions of similar TF motifs in the same locus. Collectively, these results indicate that T2D risk may, in part, be propagated through genetic modulation of RFX binding in islets. Indeed, our study shortlists only a subset of T2D-associated variants as candidates that should be functionally dissected in vivo.

Among the RFX TFs, *RFX6* is expressed in islets with high specificity (iESI = 0.94) (Fig. S11) and involved in pancreatic progenitor specification, endocrine cell differentiation, maintenance of beta cell functional identity, and control of glucose homeostasis (28–30). Beta cell-specific deletion of *RFX6* results in impaired insulin secretion (31, 32). Individuals who are heterozygous for a frameshift mutation in *RFX6* have increased 2-h glucose levels (33). Importantly, rare autosomal recessive mutations that alter DNA-contacting amino acids in the DNA binding domain of *RFX6* result in Mitchell–Riley syndrome, which is characterized by neonatal diabetes (29). Although *RFX6* was not in our motif library, a recent report found it to be highly similar to the other RFX family motifs (25), consistent with the expectation for highly conserved DNA binding domains (20). Our findings could represent a connection between rare coding variation in the islet master TF *RFX6* (30, 31) and common noncoding variations in multiple target sites for this TF. The impact of these variations mirror the expected physiological effect, with coding variants that result in neonatal diabetes and noncoding variants that result in later-onset T2D. This study implicates impaired RFX-dependent transcriptional responses in genetic susceptibility to T2D and nominates mechanistic hypotheses about the molecular genetic pathogenesis of this complex disease. Following up on the reported loci to functionally validate this hypothesis could help in better understanding T2D mechanisms. Given that most other GWAS SNPs are noncoding, this approach could be used to identify other master TF and multiple target site relationships.

## Materials and Methods

A detailed description of computational and experimental analyses is provided in *SI Materials and Methods*. Briefly, we conducted high-depth, strand-specific mRNA-seq and dense genotyping in human islets followed by *cis*-eQTL analysis. We integrated the *cis*-eQTL maps with chromatin-state annotations generated from ChIP-seq datasets for different histone modifications across diverse cell types. We profiled open chromatin in two islet samples using ATAC-seq and carried out TF footprinting using a library of motifs.

**ACKNOWLEDGMENTS.** We thank additional members of our laboratories and Finland–United States Investigation of NIDDM Genetics (FUSION) Study investigators for helpful comments on and critiques of the study and

GENETICS

manuscript. This study was supported by National Institute of Diabetes and Digestive and Kidney Diseases Grants F31HL127984 (to M.E.C.), U01DK062370 (to M.B.), ZIAHG000024 (to F.S.C.), R00DK099240 (to S.C.J.P.), 5R00DK092251 (to M.L.S.), and R01DK093757, U01DK105561, and R01DK072193 (to K.L.M.) and American Diabetes Association Pathway to Stop Diabetes Grant 1-14-INI-07 (to S.C.J.P.). This research was supported, in part, by the Intramural Research Program of the National Human Genome Research Institute, NIH.

1. Mohlke KL, Boehnke M (2015) Recent advances in understanding the genetic architecture of type 2 diabetes. *Hum Mol Genet* 24(R1):R85–R92.
2. Parker SCJ, et al.; NISC Comparative Sequencing Program; National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program Authors; NISC Comparative Sequencing Program Authors (2013) Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci USA* 110(44):17921–17926.
3. Pasquali L, et al. (2014) Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* 46(2):136–143.
4. Trynka G, et al. (2013) Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* 45(2):124–130.
5. 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(38):13924–13929.
6. van de Bunt M, et al. (2015) Transcript expression data from human islets links regulatory signals from genome-wide association studies for type 2 diabetes and glycemic traits to their downstream effectors. *PLoS Genet* 11(12):e1005694.
7. Scott LJ, et al. (2016) The genetic regulatory signature of type 2 diabetes in human skeletal muscle. *Nat Commun* 7:11764.
8. Kundaje A, et al.; Roadmap Epigenomics Consortium (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518(7539):317–330.
9. Ernst J, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473(7345):43–49.
10. Mikkelsen TS, et al. (2010) Comparative epigenomic analysis of murine and human adipogenesis. *Cell* 143(1):156–169.
11. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10(12):1213–1218.
12. Pique-Regi R, et al. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* 21(3):447–455.
13. Vierra NC, et al. (2015) Type 2 diabetes-associated K+ channel TALK-1 modulates β-cell electrical excitability, second-phase insulin secretion, and glucose homeostasis. *Diabetes* 64(11):3818–3828.
14. Quang DX, Erdos MR, Parker SCJ, Collins FS (2015) Motif signatures in stretch enhancers are enriched for disease-associated genetic variants. *Epigenetics Chromatin* 8(1):23.
15. Schmidt EM, et al. (2015) GREGOR: Evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* 31(16):2601–2606.
16. Saint-André V, et al. (2016) Models of human core transcriptional regulatory circuitries. *Genome Res* 26(3):385–396.
17. Allum F, et al.; Multiple Tissue Human Expression Resource Consortium (2015) Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. *Nat Commun* 6:7211.
18. Moyerbrailean GA, et al. (2016) Which genetics variants in DNase-Seq footprints are more likely to alter binding? *PLoS Genet* 12(2):e1005875.
19. Maurano MT, et al. (2015) Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat Genet* 47(12):1393–1401.
20. Aftab S, Semenec L, Chu JS-C, Chen N (2008) Identification and characterization of novel human tissue-specific RFX transcription factors. *BMC Evol Biol* 8(1):226.
21. Gaulton KJ, et al.; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2015) Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat Genet* 47(12):1415–1425.
22. Farh KK-H, et al. (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518(7539):337–343.
23. Girard C, et al. (2001) Genomic and functional characteristics of novel human pancreatic 2P domain K(+) channels. *Biochem Biophys Res Commun* 282(1):249–256.
24. Lotshaw DP (2007) Biophysical, pharmacological, and functional characteristics of cloned and native mammalian two-pore domain K+ channels. *Cell Biochem Biophys* 47(2):209–256.
25. Lizio M, et al.; FANTOM consortium (2015) Mapping mammalian cell-type-specific transcriptional regulatory networks using KD-CAGE and ChIP-seq data in the TC-YIK cell line. *Front Genet* 6:331.
26. Corradin O, et al. (2014) Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res* 24(1):1–13.
27. Guo C, et al. (2015) Coordinated regulatory variation associated with gestational hyperglycaemia regulates expression of the novel hexokinase HKDC1. *Nat Commun* 6:6069.
28. Zhu Z, et al. (2016) Genome editing of lineage determinants in human pluripotent stem cells reveals mechanisms of pancreatic development and diabetes. *Cell Stem Cell* 18(6):755–768.
29. Smith SB, et al. (2010) Rfx6 directs islet formation and insulin production in mice and humans. *Nature* 463(7282):775–780.
30. Soyer J, et al. (2010) Rfx6 is an Ngn3-dependent winged helix transcription factor required for pancreatic islet cell development. *Development* 137(2):203–212.
31. Piccand J, et al. (2014) Rfx6 maintains the functional identity of adult pancreatic β cells. *Cell Reports* 9(6):2219–2232.
32. Chandra V, et al. (2014) RFX6 regulates insulin secretion by modulating Ca2+ homeostasis in human β cells. *Cell Reports* 9(6):2206–2218.
33. Huopio H, et al. (2016) Clinical, genetic, and biochemical characteristics of early-onset diabetes in the Finnish population. *J Clin Endocrinol Metab* 101(8):3018–3026.
34. Gershengorn MC, et al. (2004) Epithelial-to-mesenchymal transition generates proliferative human islet precursor cells. *Science* 306(5705):2261–2264.
35. Manichaikul A, et al. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26(22):2867–2873.
36. Dobin A, et al. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15–21.
37. Hartley SW, Mullikin JC (2015) QoRTs: A comprehensive toolset for quality control and data processing of RNA-Seq experiments. *BMC Bioinformatics* 16(1):224.
38. Jun G, et al. (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* 91(5):839–848.
39. Anders S, Pyl PT, Huber W (2015) HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31(2):166–169.
40. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 44(8):955–959.
41. Auton A, et al.; 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526(7571):68–74.
42. Delaneau O, Zagury J-F, Marchini J (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 10(1):5–6.
43. Fuchsberger C, Abecasis GR, Hinds DA (2015) minimac2: Faster genotype imputation. *Bioinformatics* 31(5):782–784.
44. Shabalin AA (2012) Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28(10):1353–1358.
45. Stegle O, Parts L, Durbin R, Winn J (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLOS Comput Biol* 6(5):e1000770.
46. Stegle O, Parts L, Piipari M, Winn J, Durbin R (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* 7(3):500–507.
47. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 100(16):9440–9445.
48. Willer CJ, Li Y, Abecasis GR (2010) METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26(17):2190–2191.
49. Welter D, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42(Database issue):D1001–D1006.
50. Miyazaki J, et al. (1990) Establishment of a pancreatic β cell line that retains glucose-inducible insulin secretion: Special reference to expression of glucose transporter isoforms. *Endocrinology* 127(1):126–132.
51. Kulzer JR, et al. (2014) A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. *Am J Hum Genet* 94(2):186–197.
52. Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* 28(8):817–825.
53. Ernst J, Kellis M (2012) ChromHMM: Automating chromatin-state discovery and characterization. *Nat Methods* 9(3):215–216.
54. Rozowsky J, et al. (2011) AlleleSeq: Analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* 7(1):522.
55. van de Geijn B, McVicker G, Gilad Y, Pritchard JK (2015) WASP: Allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* 12(11):1061–1063.
56. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS (2007) Quantifying similarity between motifs. *Genome Biol* 8(2):R24.