Genetic regulatory signatures underlying islet gene expression and type 2 diabetes

Arushi Varshneya,1, Laura J. Scottb,1, Ryan P. Welchb,1, Michael R. Erdosc,1, Peter S. Chinesc,1, Narisu Narisuc,1, Ricardo D’O. Albanus,b, Peter Orchardb,1, Brooke N. Wolfordb, Romy Kursaweb, Swaroopanari Vadlamudia, Maren E. Cannona, John P. Didionc, John Hensleyd, Anthony Kirilusha, NISC Comparative Sequencing Programf,g,2, Lori L. Bonncastle, D. Leland Taylorc,1, Richard Watanaeb, Karen L. Mohlkeb, Michael Boehneh,k, Frances S. Collinsa,1,2, Stephen C. J. Parkead,4,1,3, and Michael L. Stitzele,1,3

*Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109; 1Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109; 2National Human Genome Research Institute, NIH, Bethesda, MD 20892; 3Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI 48109; 4The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032; 5Department of Genetics, University of North Carolina, Chapel Hill, NC 27599; 6NIH Intramural Sequencing Center, National Human Genome Research Institute, NIH, Bethesda, MD 20892; 7European Molecular Biology Laboratory, Wellcome Trust Genome Campus, European Bioinformatics Institute, Hinxton, Cambridgeshire CB10 1SD, United Kingdom; 8Department of Preventive Medicine, University of Southern California Keck School of Medicine, Los Angeles, CA 90089; and 9Department 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-eQTLs). 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 (RXF) footprints. Remarkably, within and across independent loci, T2D risk alleles that overlap with RXF footprints uniformly disrupt the RXF 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 RXF regulatory grammar plays a significant role in the genetic component of T2D predisposition.


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).

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 (RXF) is predicted to bind in an islet-specific manner. Genetic variants that increase T2D risk are predicted to disrupt RXF binding, providing a molecular mechanism to explain how the genome can influence the epigenome, modulating gene expression and ultimately T2D risk.

Significance

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. 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 interactions 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

chromatin | diabetes | eQTL | epigenome | footprint
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 6,602,281 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 for the two sets of islet samples and combined the cis-eQTL results (8) by sequencing (ChIP-seq) data for five histone modifications 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).

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 KCNK16 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 theparkeralab.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.

Varshney et al.
localization when tested in a mouse model (13). Our analysis revealed that rs1535500 is not associated with KCNK17 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 rs342471110) that are located in an islet promoter chromatin state, and all but rs34247110 are located in an ATAC-seq peak (Fig. 1A). 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. 1F). 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 (KCNK6). Together, these results illustrate how integrated analyses help to identify potential causal variation associated with islet expression that can be used to enable easy, in-depth exploration of our results, we created an interactive islet cis-eQTL and chromatin-state browser (theparkergenetics.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 $\geq 3$ kb long. These segments tend to mark cell identity regions and have been shown to have tissue-specific GWAS SNPs (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 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 chromatin 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. S6A), which is expressed in islets with high specificity (iESI = 0.78), overlapped islet-specific enhancer states (Fig. S6B). 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 cell types/assays. 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. S6C). 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. S6D). Likewise, in other tissues, tissue-specific enhancer clusters were enriched for the corresponding tissues’ stretch enhancers (Fig. S6D). Next, we asked if islet cis-eQTLs were enriched in specific enhancer clusters and observed enrichment in multiple clusters (Fig. S6E). 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 ($P = 1.2 \times 10^{-8}$, fold enrichment = 1.91) (Fig. S6E). 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 identity regulatory regions, such as promoters and enhancers, but lack the specific chromatin context that is important for regulatory activity linked to one gene (Fig. S6). As an example, we found that results illustrate how integrated analyses help to identify potential causal variation associated with islet expression that can be used to enable easy, in-depth exploration of our results, we created an interactive islet cis-eQTL and chromatin-state browser (theparkergenetics.org/tools/isleteqtl).
Footprints (Fig. 3A) expression by modulating TF binding. We found that T2D-GWAS SNPs that could regulate gene motifs. PWM. Row 2: PWM genetically reconstructed using the overrepresented alleles (and extent of overrepresentation) for SNPs with significant ATAC-seq allelic motifs compared with their enrichment in TF nonfootprint motifs. TFs for which footprint and nonfootprint motifs overlap four or more eQTL SNPs are served enrichment of T2D-associated SNPs in islet FOXA2 ChIP-seq peaks (21). We obbers are enriched. A recent study found enrichment of T2D risk alleles may explain why similar motifs from many RFX family mem-
evolutionarily conserved DNA binding domains (20), which The RFX family of TFs recognizes X-box motifs and has highly significant enrichment of T2D-associated SNPs in islet non-
cis-eQTLs, we next 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. 2D, 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. 3A 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. 3A). The RFX family of TFs recognizes X-box motifs and has highly evolved conservatively 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. 3B). 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. 3B 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. 1B, C, and E) and ABCB9 (PITPNM2 locus) (Fig. 1C). 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 confluence 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...
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 present 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 K+ signaling and GSIS by simultaneously reactivating 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. Li 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 contiguous 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.
This study was supported by National Institute of Diabetes and Digestive and Kidney Diseases Grants F31HL127984 (to M.E.C.), U10DK062370 (to M.B.), ZIAHG000204 (to F.S.C.), R00DK099240 (to S.C.J.P.), SR00DK092251 (to M.L.S.), and R01DK093757, U10DK05561, and R01DK072193 (to K.L.M.) and American Diabetes Association Pathway to Stop Diabetes Grant 1-14-AD1 (to S.C.J.P.). This research was supported, in part, by the Intramural Research Program of the National Human Genome Research Institute, NIH.