Integrating Enhancer RNA signatures with diverse omics data identifies characteristics of transcription initiation in pancreatic islets

Arushi Varshney[1], Yasuhiro Kyono[2], Venkateswaran Ramamoorthi Elangovan[1], Collin Wang[1], Michael R. Erdos[3], Narisu Narisu[3], Ricardo D'Oliveira Albanus[1], Peter Orchard[1], Michael L. Stitzel[4], Francis S. Collins[3], Jacob O. Kitzman[1,5], Stephen C. J. Parker[1,5,*]

1 Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, USA

2 Institute for Genomics and Systems Biology, University of Chicago, Chicago, IL, USA

3 National Human Genome Research Institute, NIH, Bethesda, MD, USA

4 The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA

5 Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

* Corresponding author scjp@umich.edu

Abstract

Identifying active regulatory elements and their molecular signatures is critical to understand gene regulatory mechanisms and subsequently better delineating biological mechanisms of complex diseases and traits. Studies have shown that active enhancers can be transcribed into enhancer RNA (eRNA). Here, we identify actively transcribed regulatory elements in human pancreatic islets by generating eRNA profiles using cap analysis of gene expression (CAGE) across 70 islet samples. We identify 9,954 clusters of CAGE tag transcription start sites (TSS) or tag clusters (TCs) in islets, ~20% of which are islet-specific when compared to CAGE TCs across publicly available tissues. Islet TCs are most enriched to overlap genome wide association study (GWAS) loci for islet-relevant traits such as fasting glucose. We integrated islet CAGE profiles with diverse epigenomic information such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) profiles of five histone modifications and accessible chromatin profiles from the assay for transposase accessible chromatin followed by sequencing (ATAC-seq), to understand how the underlying islet chromatin landscape is associated with TSSs. We identify that ATAC-seq informed transcription factor (TF) binding sites (TF 'footprint' motifs) for the RFX TF family are highly enriched in transcribed regions occurring in enhancer chromatin states, whereas TF footprint motifs for the ETS family are highly enriched in transcribed regions within promoter chromatin states. Using massively parallel reporter assays in a rat pancreatic islet beta cell line, we tested the activity of 3,378 islet CAGE elements and found that 2,279 (~67.5%) show significant regulatory activity (5% FDR). We find that TCs within accessible enhancer show higher enrichment to overlap T2D GWAS loci than accessible enhancer annotations alone, suggesting that TC annotations pinpoint active regions within the enhancer chromatin states. This work provides a high-resolution transcriptional regulatory map of human pancreatic islets.

Introduction

Type 2 diabetes (T2D) is a complex disease that results from an interplay of factors such as pancreatic islet dysfunction and insulin resistance in peripheral tissues such as fat and muscle. GWASs to date have identified >240 loci that modulate risk for T2D (Mahajan *et al.* 2018). However, these SNPs mostly occur in non protein-

coding regions and are highly enriched to overlap islet-specific enhancer regions (Parker *et al.* 2013; Pasquali *et al.* 2014; Quang *et al.* 2015; Thurner *et al.* 2018). These findings suggest the variants likely affect gene expression rather than directly altering protein structure or function. Moreover, due to the correlated structure of common genetic variations across the genome, GWAS signals are usually marked by numerous SNPs in high linkage disequilibrium (LD). Therefore, identifying causal SNP(s) is extremely difficult using genetic information alone. These factors have impeded our understanding of the molecular mechanisms by which genetic variants modulate gene expression in orchestrating disease.

In order to understand gene regulatory mechanisms, it is essential to identify regulatory elements at high genomic resolution. Active regulatory elements can be delineated by profiling covalent modifications of the histone H3 subunit such as H3 lysine 27 acetylation (H3K27ac) which is associated with enhancer activity (Creyghton *et al.* 2010; Zhou *et al.* 2011), H3 lysine 4 trimethylation (H3K4me3) which is associated with promoter activity among others (Mikkelsen *et al.* 2007; Adli *et al.* 2010; Zhou *et al.* 2011). However, such histone modification based methods identify regions of the genome that typically span hundreds of base pairs. Since TF binding can affect gene expression, TF accessible regions of the chromatin within these broader regulatory elements enable higher-resolution identification of the functional DNA bases. Numerous studies have utilized diverse chromatin information in pancreatic islets to nominate causal gene regulatory mechanisms (Fadista *et al.* 2014; Bunt *et al.* 2015; Varshney *et al.* 2017; Roman *et al.* 2017; Thurner *et al.* 2018).

Studies have shown that a subset of enhancers are transcribed into enhancer RNA (eRNA), and that transcription is a robust predictor of enhancer activity (Andersson *et al.* 2014; Mikhaylichenko *et al.* 2018). eRNAs are nuclear, short, mostly-unspliced, 5' capped and usually non-polyadenylated (Andersson *et al.* 2014). eRNAs have generally shown to be bidirectionally transcribed with respect to the regulatory element (Kim *et al.* 2010; Melgar *et al.* 2011; Andersson *et al.* 2014), however, unidirectional transcription at enhancers has also been reported (Koch *et al.* 2011). Previous studies have indicated that these transcripts could be stochastic output of Pol2 and TF machinery at active regions, whereas in some cases, the transcripts could serve important functions such as sequestering TFs or potentially assisting in chromatin looping (Li *et al.* 2013; Kaikkonen *et al.* 2013; Hsieh *et al.* 2014; Yang *et al.* 2016). Therefore, identifying the location of transcription initiation can pinpoint active regulatory elements.

Genome-wide sequencing of 5′ capped RNAs using CAGE can detect TSSs and thereby profile transcribed promoter and enhancer regions (Kim *et al.* 2010; Andersson *et al.* 2014). CAGE-identified enhancers are two to three times more likely to validate in functional reporter assays than non-transcribed enhancers detected by chromatin-based methods (Andersson *et al.* 2014). An advantage of CAGE is that it can be applied on RNA samples from hard to acquire biological tissue such as islets and does not require live cells that are imperative for other TSS profiling techniques such as GRO-cap seq (Core *et al.* 2008, 2014; Lopes *et al.* 2017). The functional annotation of the mammalian genome (FANTOM) project (The FANTOM Consortium *et al.* 2014) has generated an exhaustive CAGE expression atlas across 573 primary cell types and tissues, including the pancreas. However, pancreatic islets that secrete insulin and are relevant for T2D and related traits, constitute only ~1% of the pancreas tissue. Therefore, the bulk pancreas transcriptome does not accurately represent the islet enhancer transcription landscape. Motivated by these reasons, we profiled the islet transcriptome using CAGE. Here, we present the islet CAGE TSS atlas of pancreatic islets, validate this atlas using a massively

parallel reporter assay, and perform integrative analyses across diverse genetic and epigenomic data sets to reveal how islet TSSs are associated with T2D and related traits.

Results

The CAGE landscape in human pancreatic islets

We analyzed transcriptomes in 70 human pancreatic islet total RNA samples obtained from unrelated organ donors by performing CAGE. To enrich for the non poly-adenylated and smaller (<1kb) eRNA transcripts, we performed polyA depletion and fragment size selection (<1kb, see Methods). CAGE libraries were prepared according to the no-amplification non-tagging CAGE libraries for Illumina next-generation sequencers (nAnT-iCAGE) protocol (Murata *et al.* 2014), and an 8 bp unique molecular identifier (UMI) was added during reverse transcription to identify PCR duplicates. We sequenced CAGE libraries, performed pre-alignment quality control (QC), mapped to the hg19 reference genome, performed post-alignment QC, and identified CAGE tags. We selected 57 samples that passed our QC measures (see Methods) for all further analyses. To identify regions with high density of transcription initiation events, we identified CAGE TCs using the paraclu (Frith *et al.* 2008) method in each islet sample. We then identified a consensus set of aggregated islets TCs by merging TCs across samples in a strand-specific manner and retaining TC segments that were supported by at least 10 individual samples (see Methods, Supplementary Figure 1). We identified 9,954 tag clusters with median length of 176 bp (Supplement Figure 2), spanning a total genomic territory of ~2.4 Mb. To analyze characteristics of islet TCs and explore the chromatin landscape underlying these regions, we utilized publicly available ChIP-seq data for five histone modifications along with ATAC-seq data in islets (Varshney *et al.* 2017). We integrated the datasets for histone modifications, namely, promoter associated H3K4me3, enhancer associated H3K4me1, active promoter and enhancer associated H3K27ac, transcribed gene-associated H3K36me3 and repressed chromatin associated H3K27me3 in islets along with corresponding publicly available ChIP-seq datasets for Skeletal Muscle, Adipose and Liver (included for other ongoing projects) using ChromHMM (Ernst and Kellis 2010, 2012; Ernst *et al.* 2011). This analysis produced 11 distinct and recurrent chromatin states (Supplement Figure 3), including promoter, enhancer, transcribed, and repressed states. Figure 1A shows an example locus in the intronic region of the *ST18* gene where a TC identified in islets overlaps the active TSS chromatin state and an ATAC-seq peak. The regulatory activity of this element was validated by the VISTA project in an *in vivo* reporter assay in mouse embryos (Visel *et al.* 2007).

We next compared the islet TCs with CAGE peaks identified across diverse cell/tissue types by the FANTOM project. Using 988 CAGE libraries across human cell lines/tissue types, the FANTOM project identified CAGE peaks using a decomposition-based peak identification (DPI) approach (The FANTOM Consortium *et al.* 2014). They then classified peaks with a CAGE TSS with more than 10 read counts in at least 1 sample and 1 tags per million (TPM) as 'robust' CAGE peaks. We observed that 79.5% of Islet TC segments overlapped (at least 1bp) with the FANTOM robust peaks, and the total overlapping region comprised 25.5% of the total islet TC territory (Figure 1B). To compare islet TCs with TCs in individual FANTOM tissues, we identified TCs in 118 FANTOM human tissues with publicly available CAGE TSSs using the paraclu method with the same parameters as Islets (see Methods). For each islet TC segment, we then calculated the number of FANTOM tissues in which TCs overlapped the segment. We observed that ~20% of Islet TCs were unique to islets, whereas about ~60% of islet TCs were shared across 60 or more FANTOM tissues (Figure 1C). We highlight an example locus where an islet TC in the *AP1G2* gene occurs in active TSS chromatin states across multiple tissues, and overlaps shared ATAC-seq peaks in islet and the lymphoblastoid cell line GM12878 (Buenrostro

et al. 2013) (Figure 1D). This region was also identified as a TC in the FANTOM tissues (Figure 1D, green box). Another islet TC ~34kb away, however, occurs in a region lacking gene annotations, and overlaps a more islet-specific active enhancer chromatin state and ATAC-seq peak (Figure 1D, blue box). This region was not identified as a TC in the 118 analyzed FANTOM tissues. Collectively, these results highlight that CAGE profiling in islets identifies islet-specific sites of active transcription initiation.

We next asked if islet TCs preferentially overlapped certain genomic annotations. We computed the enrichment of islet TCs to overlap islet annotations such as active TSS, enhancer etc. chromatin states and islet ATAC-seq peaks. We also included 'common' annotations such as known gene promoters, coding, untranslated (UTR) regions, or annotations such as super enhancers, or histone-modification ChIP-seq peaks that were aggregated across multiple cell types. We observed that islet TCs were highly enriched to overlap islet active TSS chromatin states (fold enrichment = 69.72, P value = 0.0001, Figure 1E, Supplementary Table 1). This result is expected since CAGE profiles transcription start sites where the underlying chromatin is more likely to resemble the 'active TSS' chromatin state. TCs were also enriched to overlap islet ATAC-seq peaks (fold enrichment = 53.8, P value = 0.0001, Figure 1E), signifying that the identified transcription initiation sites constitute accessible chromatin where TFs can bind.

To gauge if these transcribed elements could be relevant for diverse disease/traits, we computed enrichment for islet TCs to overlap GWAS loci for 116 traits from the NHGRI catalog (Buniello *et al.* 2019) and other relevant studies (Udler *et al.* 2018). We observed that traits such as Fasting Glucose (FGlu) (fold enrichment = 7.05, P value = 3.30×10^{-4}), metabolic traits (fold enrichment = 6.46, P value = 2.03×10^{-4}) were among the most highly enriched, highlighting the relevance of these transcribed elements for islet biology (Figure. 1F, Supplementary Table 2). GWAS loci for T2D were also enriched in islet TCs (fold enrichment = 2.48, P value = 0.02). Because T2D is orchestrated through a complex interplay between islet beta cell dysfunction and insulin resistance in peripheral tissues, we reasoned that some underlying pathways in T2D might be more relevant in islets than others. To explore this rationale, we utilized results from a previous study that analyzed GWAS data for T2D along with 47 other diabetes related traits and identified clusters of T2D GWAS signals (Udler *et al.* 2018). Interestingly, we observe that GWAS loci in the islet beta cell and proinsulin cluster were highly enriched to overlap islet TCs (fold enrichment = 5.62, P value = 0.004), whereas loci in the insulin resistance cluster were depleted (fold enrichment = 0.97; Figure 1F, Supplementary Table 2). These results suggest that islet TCs comprise active regulatory elements relevant for traits specifically related to islet function.

Integrating CAGE TCs with epigenomic information

We further explored CAGE profiles relative to the underlying chromatin landscape to identify characteristics of transcription initiation. We first overlayed CAGE profiles over ATAC-seq data. Aggregated CAGE signal over ATAC-seq narrow peak summits highlighted patterns of transcription initiation just downstream of the ATAC-seq peak summit on both strands (Figure 2A). Conversely, on anchoring the ATAC-seq signal over islet TC centers we observed that the summit of the ATAC-seq signal lies upstream of the TC center (Figure 2B). We next asked if TF binding sites were more enriched to occur upstream or downstream with respect to the TC. We utilized TF footprint motifs previously identified using islet ATAC-seq data and TF DNA binding position weight matrices (PWMs) (Varshney *et al.* 2017). These footprint motifs represent putative TF binding sites that are also supported by accessible chromatin profiles, as opposed to TF motif matches that are only informed by DNA sequence. We observe that most TF footprint motifs were more enriched to overlap the 500 bp TC upstream region compared to the 500 bp downstream region relative to TCs (Figure 2C). These observations indicate that the region just upstream of the TC is highly accessible where more TF binding events occur.

We next explored the characteristics of TCs that occurred in the two main regulatory classes - promoter and enhancers, relative to each other. We focussed on transcribed, accessible regions in promoter or enhancer chromatin states - namely TCs overlapping ATAC-seq peaks in promoter (active, weak or flanking TSS) chromatin states or enhancer (active, weak or genic enhancer) chromatin states. We considered the proximity of these elements to known gene TSSs and further classified the segments as TSS proximal or distal using a 5kb distance threshold from the nearest protein coding genes (Gencode V19) (Harrow *et al.* 2012). We then explored the chromatin landscape at these regions across 98 Roadmap Epigenomics cell types for which chromatin state annotations are publicly available (18 state 'extended model', see Methods) (The Roadmap Epigenomics Consortium *et al.* 2015). We observed that TSS proximal islet TCs in accessible islet promoter chromatin states (N = 7,064 segments) were nearly ubiquitously identified as promoter chromatin states across roadmap cell types (Figure 2D, left). A subset of TSS distal islet TCs in accessible islet TSS chromatin states (out of total N = 443 segments) however were more specific for pancreatic islets (Figure 2D, right). In contrast, we observed that islet TCs in accessible islet enhancer chromatin states, both proximal (N = 254 segments) and distal (N = 289 segments) to known gene TSSs were more specifically identified as enhancer states in pancreatic islets (Figure 2E). This pattern was more clear for Roadmap pancreatic islet segmentations compared to the whole pancreas segmentations (Figure 2D and E, labelled) which further emphasizes the differences between epigenomic profiles for islets vs the whole pancreas tissue.

Having observed differences in cell type-specificities in islet TCs in promoter vs enhancer states, we next asked if TFs displayed preferences to bind in these regions. We observed that footprint motifs for the regulatory factor X (RFX) TF family were highly enriched (>3 fold, P value = 0.0001) in TCs in accessible enhancers (Figure 2F). On the other hand, TCs in accessible promoter regions were highly enriched to overlap footprint motifs of the E26 transformation-specific (ETS) TF family (Figure 2F). We observe divergent aggregate CAGE profiles over TF footprint motifs enriched in enhancers for example RFX5_known8 footprint motifs in 5kb TSS distal regions and ELK4_1 motif (Figure 2G and H). Taken together, these results describe the differences in the characteristics of transcription initiation sites based on the underlying chromatin context.

Experimental validation of transcribed regions

We next sought to experimentally validate the transcriptional activity of islet CAGE-profiled regions. We utilized a massively parallel reporter assay (MPRA) platform wherein thousands of elements can be simultaneously tested by including unique barcode sequences for each element and determining the transcriptional regulatory activity using sequencing-based barcode quantification (Melnikov *et al.* 2012; Arnold *et al.* 2013). This approach is also known as the self-transcribing active regulatory region sequencing (STARR-seq) assay. We generated a library of 7,188 CAGE elements (198 bp each, see Methods) and cloned these along with a library of 16 bp barcode sequences into the STARR-seq vector downstream of the GFP gene such that the barcode would itself be transcribed. In our setup, therefore, one CAGE element was represented by multiple barcodes. We transfected the STARR-seq libraries into rat beta cell insulinoma (INS1 832/13) cell line in triplicate, extracted DNA and RNA and sequenced the barcodes. We added 6 bp unique molecular identifier (UMI) sequences before the PCR amplification of the RNA libraries to enable accounting for PCR duplicates while quantifying true biological RNA copies. We selected barcodes that were observed with at least 10 DNA counts and non zero RNA counts in at least one replicate, and selected CAGE elements that were observed with at least two of such qualifying barcodes. This filtering procedure resulted in 3,378 CAGE elements. We observed high correlation between the normalized sum of RNA counts of the CAGE element barcodes across the three

biological replicates (pearson r = 0.97 Supplementary Figure 4). We quantified the enhancer reporter activity of the qualifying CAGE elements by modeling the RNA and DNA barcode counts using generalized linear models (GLMs) implemented in the MPRAnalyze package (Ashuach *et al.* 2019). We then tested for significant transcriptional activity against a null model (see Methods). We observed that 67.4% (N = 2,279) of the testable CAGE elements showed significant regulatory activity (5% FDR) (Figure 3A (top)). On classifying CAGE elements based on the underlying chromatin landscape such as - promoter (active, weak or flanking TSS) chromatin state, enhancer (active, weak or genic enhancer) chromatin state or other chromatin state overlap in islets, we observed that a larger fraction of CAGE elements overlapping the promoter states had significant transcriptional activity compared to elements overlapping enhancer states, which was in turn higher than CAGE elements in other chromatin states (Figure 3A (bottom)). We also observed that the CAGE elements in promoter chromatin states had higher STARR-seq activity Z scores compared to the elements in enhancer chromatin states (Wilcoxon rank sum test P = 1.02×10^{-6}) (Figure 3B). Z scores for the CAGE elements that overlapped ATAC-seq peaks were significantly higher than the elements that did not occur in peaks (Wilcoxon rank sum test P = 5.50×10^{-16}) (Figure 3C). Z scores for CAGE elements 5kb proximal to protein-coding gene TSSs (Gencode V19) were higher than CAGE elements that were distal to gene TSS locations (Wilcoxon rank sum test P = 5.38×10^{-9}) (Figure 3D).

We then asked if sequence-based features of the CAGE elements such as the occurrence of TF motifs could predict the activities of these elements in the STARR-seq assay. We trained a lasso regression model with the CAGE element STARR-seq Z scores as the response variable and the TF motif scan scores for 540 representative TF motifs in each CAGE element as predictors (see Methods). We report the TF motifs with the top 30 lasso regression coefficients in (Figure 3E, Supplementary table 3). We observed that TF motifs from the ETS family showed positive lasso coefficients, indicating that these sequence elements are associated with high transcriptional activity. Indeed, we earlier observed that these motifs were also highly enriched to occur in TCs in accessible promoter chromatin state regions (Figure 2F). Other TF motifs with positive lasso coefficients included CEBP, YY1, and NRF-1 (Figure 3E). NRF-1, for instance, has been identified to be relevant for islet biology, as its target genes are downregulated in diabetic individuals (Patti *et al.* 2003); knockdown of this gene in the mouse insulinoma cell line (MIN6) and beta cell specific Nrf1-knockout mice resulted in impaired glucose responsiveness, elevated basal insulin release and decreased glucose-stimulated insulin secretion (GSIS) (Zheng *et al.* 2015).

In Figure 3F, we highlight an islet TC for which we tested three elements (Figure 3F, STARR-seq elements track), which occurred in active TSS and enhancer states and overlapped an ATAC-seq peak. All three elements showed significant transcriptional activity in our assay (Z score > 2.94, P values < 0.001). Interestingly, while there are no known gene TSS annotations in this region, clear islet polyA+ mRNA-seq profiles overlapping the CAGE signal can be observed here. The active elements overlapped TF motifs that showed positive lasso regression coefficients (Figure 3F, positive lasso motifs track) but also motifs that showed negative lasso regression coefficients (Figure 3F, negative lasso motifs track).

We highlight another example locus where a TC is identified in the promoter region of the genes *DCAF16* and *NCAPG* (Figure 3G). The TC lies close to two T2D GWAS SNPs rs7667864 and rs2074974. These two SNPs are in high LD (rs7667864 r2 = 0.97, rs2074974 r2 = 0.96) with the lead SNP rs12640250 at this GWAS locus named *LCORL* (Supplementary Figure 5). This element showed significant activity in the STARR-seq assay (Z score = 18.48, p value = 1.56×10^{-76}), Figure 3G). Several TF motifs that showed positive lasso regression

coefficients occur in this region (Figure 3G, positive lasso motifs track), whereas other motifs with negative lasso regression coefficients occur nearby (Figure 3G, negative lasso motifs track). Through these analyses we could experimentally validate a large proportion of testable CAGE elements for significant transcriptional regulatory activity in a rodent islet beta cell model system; and identify TF motifs associated with high transcriptional activities.

CAGE profiles augment functional genomic annotations to better understand GWAS and eQTL associations

Observing the molecular signature of islet TCs in different epigenomic contexts and validating the activities of these elements, we next asked if islet TCs taken as an additional layer of functional genomic information could supplement our understanding of GWAS or islet eQTL associations. We classified genomic annotations based on layers of epigenomic data such as a) histone modification based chromatin states, b) accessible regions within these states and c) transcribed and accessible regions within these states. We then computed enrichment for T2D GWAS loci (Mahajan *et al.* 2018) to overlap these annotations using a Bayesian hierarchical model implemented in the fGWAS tool (Pickrell 2014). This method utilizes not only the genome wide significant loci but leverages full genome wide association summary statistics such that marginal associations can also be accounted for. We observed that TCs in accessible enhancer regions were the most highly enriched for T2D GWAS signals among the annotations tested (Figure 4A, left). We also computed enrichment for annotations to overlap islet eQTL (Varshney *et al.* 2017) and observed that TCs in accessible regions in both enhancers and promoters were highly enriched (Figure 4A, right). These data suggest that including TC information with other functional genomics data help illuminate relevant regions for the genetic control of gene expression and trait association signals.

We next asked to what extent TCs or ATAC-seq annotations add information above and beyond chromatin state annotations in the GWAS and eQTL enrichment models. We performed conditional analyses in fGWAS where, first, the enrichment parameters for active TSS or active enhancer chromatin states were modeled and fixed to their maximum likelihood values. Second, an additional parameter for either TCs or ATAC-seq peaks was estimated. We then asked if the added enrichment parameter was significant (confidence intervals above zero). We observed that TCs had a higher conditional enrichment over enhancer states for T2D (Figure 4B) than ATAC-seq peaks. TCs also showed a higher conditional enrichment over enhancer and promoter states for islet eQTL loci as compared to ATAC-seq peaks (Figure 4B).

We then sought to utilize this new set of islet TC functional annotations to fine-map GWAS loci and reweight the SNP posterior probabilities of association (PPAs). We performed functional re-weighting of Fasting Glucose GWAS using either chromatin states and TCs or chromatin states and ATAC-seq peak annotations and compared the maximal SNP PPA at each locus (Figure 4C). We observed differences at some loci in the maximal SNP PPA when TCs or ATAC-seq peaks are included in the model (Figure 4C). We highlight one such region near the *ARAP1* gene that includes many variants in high LD. Variants at this T2D and FGlu GWAS locus are identified as eQTL in islets for the *STARD10* gene (Bunt *et al.* 2015) but not for *ARAP1*. Two SNPs - the GWAS and eQTL index SNP rs11603334 and rs1552224 (LD (r^2=1 with rs11603334) lie in the promoter region of *ARAP1* (Figure 4 D, E). Including chromatin state and TC information resulted in rs1552224 with the highest SNP PPA of 0.772. Including chromatin state and ATAC-seq peak information, the PPA for

both rs11603334 and rs1552224 were 0.446. We observed significant activity of the TC element that overlaps rs1552224 in our STARR-seq assay (Z score = 4.90, Z score P value = 4.78×10^{-7}). A previous study showed evidence for rs11603334 to be the causal variant (Kulzer *et al.* 2014), whereas another study pointed towards an indel rs140130268 as more likely causal (Carrat *et al.* 2017). These studies nominating different causal variants highlight the complexity at this locus. Our analyses demonstrate the utility of identifying transcription initiation sites to demarcate active regulatory elements in islets.

Discussion

We profiled TSSs in human pancreatic islets using CAGE. Islet TCs were enriched to occur in promoter chromatin states and ATAC-seq peaks, which expectedly reflects the chromatin landscape at regions where transcription initiation occurs. Comparison of islet CAGE TCs with those identified across diverse tissues revealed that 20% of islet TCs were islet-specific. Furthermore, comparing the chromatin landscape underlying the islet TCs across multiple cell types and tissues indicated that TCs that occur distal to known TSSs of protein coding genes are comprised of more islet-specific promoter and enhancer chromatin states. Our analyses also highlighted the differences in the chromatin architecture underlying islet TC in islets vs whole pancreas tissues, which further demonstrate the need for molecular profiling in the islet tissue to better understand islet biology. Islet TCs were enriched to overlap GWAS loci of fasting glucose and specifically the islet beta cell related components of T2D GWAS signals, while being depleted for the insulin resistance related components of T2D GWAS signals. These analyses demonstrate that islet TCs mark active, specific and relevant islet regulatory elements.

Surveying the TF footprint motifs occurring in TCs revealed that several ETS family footprint motifs were highly enriched in transcribed and accessible promoter regions, and that these motifs were also strong predictors of the elements' activity in the STARR-seq assay. The regulatory potential of ETS family motifs has been described in the literature. One study demonstrated that orienting for islet eQTL SNPs occurring in ETS footprint motifs, the base preferred in the motifs was significantly more often associated with increased expression of the target gene (Viñuela *et al.* 2019). Another study utilizing an MPRA assay with tiled sequences in HepG2 and K562 cell lines also observed high regulatory activities of ETS motifs (Ernst *et al.* 2016). Our concordant findings with these studies using completely orthogonal datasets highlight the robustness and utility of the CAGE dataset. Our work also revealed that transcribed and accessible enhancer regions were most enriched to overlap TF footprint motifs for the RFX family of TFs. We previously showed that RFX footprint motifs are confluently disrupted by T2D GWAS risk alleles (Varshney *et al.* 2017), which are enriched to occur in islet-specific enhancer regions. These observations together highlight the role of islet specific enhancer regions, and the potential of ATAC-seq and CAGE profiling to pinpoint the active regulatory nucleotides within enhancer regions.

Utilizing the STARR-seq enhancer MPRA approach, we observed that 67.4% of testable CAGE elements showed significant transcriptional activity which again highlights how CAGE profiling identifies active regulatory elements. A larger proportion of CAGE elements that occurred in promoter chromatin states showed significant transcriptional activity and higher activity Z scores than CAGE elements occurring in active enhancer or other chromatin states. It is interesting to note that our STARR-seq platform where the tested element is cloned downstream of the GFP gene is traditionally considered an 'enhancer' reporter assay. While STARR-seq vectors are episomal and do not recapitulate the underlying chromatin context, our results nevertheless show that sequences associated with native promoter chromatin state landscape can show strong enhancer activity when cloned downstream of the reporter gene. Indeed, T2D and FGlu GWAS SNPs rs11603334 and

rs1552225 occur in the promoter region of the *ARAP1* gene are also eQTL for the *STARD10* gene and are potentially causal variants. Here, we note that only a small fraction of CAGE TCs identified in our work (0.4%) overlapped the active enhancer chromatin state. Studies have shown that gene distal transcripts are more unstable, which would therefore be difficult to profile from a total RNA sample. Of course, given the relative instability of enhancer RNAs, enhancer chromatin-like sites may be actively transcribed but fall below the limits of detection of CAGE. Therefore, it is plausible that islet CAGE profiling from total RNA samples would comprise more stable promoter-associated RNA transcripts and have a lesser representation of weaker transcripts originating from enhancer regions. In our previous work (Varshney *et al.* 2018), we showed that genetic variants in more cell type-specific enhancer regions have lower effects on gene expression than the variants occurring in more ubiquitous promoter regions. This aspect is in line with our observation that enhancer chromatin state regions comprise a lesser proportion of active transcription initiation sites and lower transcriptional activities relative to promoter state regions.

Integrating CAGE TC information with GWAS and eQTL data also revealed the potential of the CAGE dataset to better understand the mechanisms underlying these associations. We reasoned that if CAGE TCs represent active sites within regulatory elements, GWAS or eQTL variants would be highly enriched at these sites and conversely, GWAS/eQTL SNPs occurring at these sites would more likely be causal. Indeed, regions supported by TCs, ATAC-seq peaks and enhancer chromatin states (transcribed, accessible enhancer regions) were most enriched to overlap T2D GWAS loci. This enrichment was higher than in regions only informed by ATAC-seq peaks and enhancer chromatin states, indicating that the small set of TCs in enhancer regions actually delineate highly relevant elements. Our work demonstrates that transcription initiation information profiled using CAGE in islets can be used in addition to other relevant epigenomic information such as histone mark informed chromatin states and chromatin accessibility in nominating relevant variants and biological mechanisms.

Data Availability

Raw and processed Islet CAGE data has been submitted to dBGaP (accession no. phs001188.v1.p1), raw and processed STARR-seq data has been submitted to GEO (GSE137693).

Materials and Methods

Islet Procurement and Processing

Islet samples from organ donors were received from the Integrated Islet Distribution Program, the National Disease Research Interchange (NDRI), and Prodo-Labs. Islets were shipped overnight from the distribution centers. Upon receipt, we prewarmed islets to 37 °C in shipping media for 1–2 h before harvest; ~2,500–5,000 islet equivalents (IEQs) from each organ donor were harvested for RNA isolation. We transferred 500–1,000 IEQs to tissue culture-treated flasks and cultured them as in the work in (Gershengorn *et al.* 2004).

RNA isolation, CAGE-seq library preparation and sequencing

Total RNA from 2000-3000 islet equivalents (IEQ) was extracted and purified using Trizol (Life Technologies). RNA quality was confirmed with Bioanalyzer 2100 (Agilent); samples with RNA integrity number (RIN) > 6.5

were prepared for CAGE sequencing. 1ug Total RNA samples were sent to DNAFORM, Japan where CAGE libraries were generated. The library preparation included polyA negative selection and size selection (<1000bp). Stranded CAGE libraries were generated for each islet sample using the no-amplification non-tagging CAGE libraries for Illumina next-generation sequencers (nAnT-iCAGE) protocol (Murata *et al.* 2014). Each islet CAGE library was barcoded, pooled into 24-sample batches, and sequenced over multiple lanes of HiSeq 2000 to obtain paired-end 126 bp sequences. All procedures followed ethical guidelines at the National Institutes of Health (NIH).

CAGE data mapping and processing

We processed Islet CAGE data uniformly with CAGE data for other tissues included in separate ongoing projects. Because read lengths differed across libraries, we trimmed all reads to 51 bp using fastx_trimmer (FASTX Toolkit v. 0.0.14). Adapters and technical sequences were trimmed using trimmomatic (v. 0.38; paired-end mode, with options ILLUMINACLIP:adapters.fa:1:30:7:1:true). To remove potential *E. coli* contamination, we mapped to the *E. coli* chromosome (genome assembly GCA_000005845.2) with bwa mem (v. 0.7.15; options: -M). We then removed read pairs that mapped in a proper pair (with mapq >= 10) to *E. coli.* We mapped the remaining reads to hg19 using STAR (v. 2.5.4b; default parameters) (Dobin *et al.* 2013). We pruned the mapped reads to high quality autosomal read pairs (using samtools view v. 1.3.1; options -f 3 -F 4 -F 8 -F 256 -F 2048 -q 255)(Li *et al.* 2009). We then performed UMI-based deduplication using umitools dedup (v. 0.5.5; --method directional).

We selected 57 Islet samples with strandedness measures >0.85 calculated from QoRTS (Hartley and Mullikin 2015) for all downstream analyses.

Tag cluster identification

We used the paralu method to identify clusters of CAGE start sites (CAGE tag clusters) (Frith *et al.* 2008). The algorithm uses a density parameter d and identifies segments that maximize the value of (Number of events - d * size of the segment (bp)). Here, large values of d would favor small, dense clusters and small values of d would favor larger more rarefied clusters. The method identifies segments over all values of d beginning at the largest scale, where d = 0, where all of the events are merged into one big cluster. It then calculates the density (events per nucleotide) of every prefix and suffix of the big cluster. The lowest value among all of these densities is the maximum value of d for the big cluster because at higher values of d the big cluster will no longer be a maximal-scoring segment (because zero-scoring prefixes or suffixes are not allowed).

We called TCs in each individual sample using raw tag counts, requiring at least 2 tags at each included start site and allowing single base-pair tag clusters ('singletons') if supported by >2 tags. We then merged the tag clusters on each strand across samples. For each resulting segment, we calculated the number of islet samples in which TCs overlapped the segment. We included the segment in the consensus TCs set if it was supported by independent TCs in at least 10 individual islet samples. This threshold was selected based on comparing the number of tag clusters with the number of samples across which support was required to consider the segment (Figure 1 - figure supplment 1). We then filtered out regions blacklisted by the ENCODE consortium due to poor mappability (wgEncodeDacMapabilityConsensusExcludable.bed and wgEncodeDukeMapabilityRegionsExcludable.bed) using bedtools subtract to obtain the final set of Islet tag cluster regions used in all downstream analyses.

FANTOM CAGE datasets

We obtained the set of 'robust CAGE peaks' identified by the FANTOM 5 consortium (The FANTOM Consortium *et al.* 2014) using CAGE libraries (CAGE sequencing on HeliScope Single Molecule Sequencer (hCAGE)) of 988 human cell lines or tissues (http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/hg19.cage_peak_phase1and2combined_coord.bed.gz). These peaks were identified using the decomposition-based peak identification (DPI) method (The FANTOM Consortium *et al.* 2014), followed by filtering 'robust' peaks that included a CAGE tag (TSS) with more than 10 read counts in at least 1 sample and 1 tags per million (TPM). For a more direct comparison of islet TCs with TCs from other tissues, we downloaded the CAGE transcription start site (CTSS) data for 118 tissue types (from http://fantom.gsc.riken.jp/5/datafiles/latest/basic/human.tissue.hCAGE/) and called tag clusters for each tissue sample using the paraclu method (Frith *et al.* 2008) as described above, with the same parameters.

Chromatin state analysis

We collected publicly available cell/tissue ChIP-seq data for H3K27ac, H3K27me3, H3K36me3, H3K4me1, and H3K4me3 and input for Islets, Adipose and Skeletal Muscle (Supplementary Table 4). Data for Adipose, Skeletal Muscle and Liver tissues were included in the joint model for other ongoing projects. We performed read mapping and integrative chromatin-state analyses in a manner similar to that of our previous reports and followed quality control procedures reported by the Roadmap Epigenomics Study (The Roadmap Epigenomics Consortium *et al.* 2015). Briefly, we trimmed reads across datasets to 36bp and overrepresented adapter sequences as shown by FASTQC (version v0.11.5) using cutadapt (version 1.12) (Martin 2011). We then mapped reads using BWA (version 0.5.8c), removed duplicates using samtools (Li *et al.* 2009), and filtered for mapping quality score of at least 30. To assess the quality of each dataset, we performed strand cross-correlation analysis using phantompeakqualtools (v2.0; code.google.com/p/phantompeakqualtools) (Landt *et al.* 2012). We converted bam files for each dataset to bed using the bamToBed tool. To more uniformly represent datasets with different sequencing depths across histone marks and tissues, we randomly subsampled each dataset bed file to the mean depth for that mark across the four included tissues. This allowed comparable chromatin state territories across tissues and ensured that chromatin state territories were not heavily driven by high sequencing depth. Chromatin states were learned jointly for the three cell types using the ChromHMM (version 1.10) hidden Markov model algorithm at 200-bp resolution to five chromatin marks and input (Ernst and Kellis 2010, 2012; Ernst *et al.* 2011). We ran ChromHMM with a range of possible states and selected a 11-state model, because it most accurately captured information from higher-state models and provided sufficient resolution to identify biologically meaningful patterns in a reproducible way. We have used this state selection procedure in previous analyses (Scott *et al.* 2016; Varshney *et al.* 2017). To assign biological function names to our states that are consistent with previously published states, we performed enrichment analyses in ChromHMM comparing our states with the states reported previously (Varshney *et al.* 2017) for the four matched tissues. We assigned the name of the state that was most strongly enriched in each of our states (Supplementary Figure 3).

ATAC-seq data analysis

We used previously published chromatin accessibility data profiled using ATAC-seq in islets from two human organ donor samples (Varshney *et al.* 2017). For each sample, we trimmed reads to 36 bp (to uniformly process ATAC-seq from other tissues for ongoing projects) and removed adapter sequences using Cutadapt (version 1.12) (Martin 2011), mapped to hg19 used bwa-mem (version 0.7.15-r1140) (Li 2013), removed

duplicates using Picard (http://broadinstitute.github.io/*picard*) and filtered out regions blacklisted by the ENCODE consortium due to poor mappability (wgEncodeDacMapabilityConsensusExcludable.bed and wgEncodeDukeMapabilityRegionsExcludable.bed). For each tissue we subsampled both samples to the same depth so that each tissue had overall similar genomic region called as peaks. We used MACS2 (https://github.com/taoliu/MACS), version 2.1.0, with flags "-g hs–nomodel–shift -100–extsize 200 -B–broad–keep-dup all," to call peaks and retained all broad-peaks that satisfied a 1% FDR.

Overlap enrichment between TCs and annotations

We calculated the enrichment for Islet TCs to overlap annotations such as different Islet chromatin states, Islet ATAC-seq peaks and various 'common' annotations. Common annotations imply annotations that don't vary across cell types such as coding gene regions, intronic regions or annotations created by merging epigenomic data such as histone modification peaks across cell types. We utilized 29 total static annotation bed files supplied by (Finucane *et al.* 2015) (https://data.broadinstitute.org/alkesgroup/LDSCORE/baseline_bedfiles.tgz). These included coding, un-translated regions (UTRs), promoter and intronic regions obtained from UCSC (Kent *et al.* 2002); the histone marks monomethylation (H3K4me1) and trimethylation (H3K4me3) of histone H3 at lysine 4 and acetylation of histone H3 at lysine 9 (H3K9ac) (The ENCODE project Consortium 2012; Trynka *et al.* 2013; The Roadmap Epigenomics Consortium *et al.* 2015) and acetylation of histone H3 at lysine 27 (H3K27ac) (Hnisz *et al.* 2013; Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014); open chromatin, as reflected by DNase I hypersensitivity sites (DHSs) (Trynka *et al.* 2013; Gusev *et al.* 2014); combined chromHMM and Segway predictions(Hoffman *et al.* 2013), which partition the genome based on distinct and recurring patterns of histone marks into seven underlying chromatin states; regions that are conserved in mammals(Lindblad-Toh *et al.* 2011; Ward and Kellis 2012); super-enhancers, which are large clusters of highly active enhancers(Hnisz *et al.* 2013); and enhancers with balanced bidirectional capped transcripts identified using CAGE in the FANTOM5 panel of samples, (called Enhancer (Andersson)) (Andersson *et al.* 2014). Histone marks included in the static annotation set included merged histone mark data from different cell types into a single annotation.

Enrichment for overlap between each Islet tag clusters and regulatory annotations was calculated using the Genomic Association Tester (GAT) tool (Heger *et al.* 2013). To ask if two sets of regulatory annotations overlap more than that expected by chance, GAT randomly samples segments of one regulatory annotation set from the genomic workspace (hg19 chromosomes) and computes the expected overlaps with the second regulatory annotation set. We used 10,000 GAT samplings for each enrichment run. GAT outputs the observed overlap between segments and annotation along with the expected overlap and an empirical p-value.

Enrichment of GWAS loci in TCs

We downloaded the GWAS data for various traits from the NHGRI website on June 12, 2018 (file gwas_catalog_v1.0.2-associations_e92_r2018-05-29.tsv from \url{https://www.ebi.ac.uk/gwas/docs/file-downloads}). We selected genome-wide significant GWAS SNPs ($P < 5x10^{-8}$) for traits for which the study included European samples. To retain independent signals, we LD pruned the list of SNPs to retain SNPs with the most significant P values that had LD r^2 <0.2 between each pair. This procedure was performed using the PLINK (v1.9) tool (Purcell *et al.* 2007; Chang *et al.* 2015) –clump option and 1000 genomes phase 3 vcf files (downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502), subsetted to the European samples as reference. We selected traits that had >30 independent signals for following analyses. We also downloaded data from (Udler *et al.* 2018) which included clusters of T2D GWAS loci such as beta cell

function, insulin resistance etc. This study clustered T2D GWAS loci by analyzing GWAS for T2D along with 47 other diabetes related traits.

Enrichment for GWAS variants for different traits in Islet TCs was calculated using the Genomic Regulatory Elements and Gwas Overlap algoRithm (GREGOR) tool (version 1.2.1) (Schmidt *et al.* 2015). Since the causal SNP(s) for the traits are not known, GREGOR allows considering the input lead SNP along with SNPs in high LD (based on the provided R2THRESHOLD parameter) while computing overlaps with genomic features (such as islet TCs). Therefore, as input to GREGOR, we supplied SNPs that were not in high LD with each other. For each input SNP, GREGOR selects ~500 control SNPs that match the input SNP for minor allele frequency (MAF), distance to the nearest gene, and number of SNPs in LD. Fold enrichment is calculated as the number of loci at which an input SNP (either lead SNP or SNP in high LD) overlaps the feature over the mean number of loci at which the matched control SNPs (or SNPs in high LD) overlap the same features. This process accounts for the length of the features, as longer features will have more overlap by chance with control SNP sets. Specific parameters for the GWAS enrichment were: GREGOR: r2 threshold = 0.8. LD window size = 1Mb; minimum neighbor number = 500, population = European.

Aggregate signal

We generated the ATAC-seq density plot over islet TC midpoints using the Agplus tool (version 1.0) (Maehara and Ohkawa 2015). We used the ATAC-seq signal track for reads per 10 Million to aggregate over stranded TCs.

To obtain CAGE tracks, we merged CAGE bam files for islet samples that passed QC (see CAGE data processing section) and obtained the read 1 start sites or TSSs. To better visualise the CAGE signal, we then flanked each TSS 10bp upstream and downstream and normalized the TSS counts to 10M mapped reads. We generated CAGE density plots over ATAC-seq narrow peak summits by using the agplus tool.

To obtain aggregate CAGE signal over TF footprint motifs, we oriented the CAGE signal with respect to the footprint taken on the plus strand. We used HTSeq GenomicPosition method (Anders *et al.* 2015) to obtain the sum of CAGE signal at each base pair relative to the footprint motif mid point.

Enrichment for Islet TF footprint motifs to overlap TC-related annotations

We compared the enrichment of Islet TF footprint motifs in several TC-related annotations such as upstream and downstream 500bp regions of TCs and Islet TCs that occurred in accessible enhancer states vs those that occurred in accessible promoter states using the GAT tool similarly as described above (Heger *et al.* 2013). TF footprint motifs are occurrences of TFs motifs (obtained from databases of DNA binding motifs for several TFs) in accessible chromatin regions (identified from assays such as ATAC-seq). We utilized previously published islet TF footprint motifs (Varshney *et al.* 2017), which were generated using ATAC-seq data in two islet samples and DNA binding motif information for 1,995 publicly available TF motifs (Jolma *et al.* 2013; Kheradpour and Kellis 2014; Mathelier *et al.* 2016).

We obtained the 500bp upstream or downstream regions of each TC (upstream/downstream regions were determined based on TC strand; TC region itself was not included). We generated the list of regions where islet TCs overlapped ATAC-seq peaks and any enhancer states (Active enhancer, Weak enhancer, or Genic enhancer) using BEDTools intersect - we referred to these as ‘TCs in accessible enhancers’. Similarly, we also

generated the list of regions where islet TCs overlapped ATAC-seq peaks and any TSS/promoter states (Active TSS, Weak TSS, Flanking TSS) - we referred to these as 'TCs in accessible promoters'. as segments.

In the GAT analyses, the TC-related annotations were considered as 'annotations' (argument -a) and footprint motif occurrences for each known motif were considered as 'segments' (argument -s). Since the TF footprint motifs only occur in ATAC-seq peaks, we considered these peaks as the 'workspace' (argument -w) to sample segments. We used 10,000 GAT samplings for each enrichment run. We accounted for the 1,995 footprint motifs being tested against each TC-related annotation by performing an FDR correction with the Benjajmini-Yekutieli method (Benjamini and Yekutieli 2001) using the stats.multitest.multipletests function from the statsmodels library in Python (Seabold and Perktold 2010). Significant enrichment was considered at 5% FDR threshold.

Comparison of features with Roadmap chromatin states

We downloaded the chromatin state annotations identified in 127 human cell types and tissues by the Roadmap epigenomics project (The Roadmap Epigenomics Consortium *et al.* 2015) after integrating ChIP-seq data for five histone 3 lysine modifications (H3K4me3, H3K4me1, H3K36me3, H3K9me3 and H3K27me3) that are associated with promoter, enhancer, transcribed and repressed activities, across each cell type. For each TC feature, for example, TCs in ATAC-seq peaks within islet enhancer chromatin states, we identified segments occurring proximal to (within 5kb) and distal from (further than 5kb) known protein coding gene TSS (gencode V19) (Harrow *et al.* 2012). For each such segment, we identified the maximally overlapping chromatin state across 98 cell types publicly available from the Roadmap Epigenomics project in their 18 state 'extended' model using BEDtools intersect. We then ordered the segments using clustering (hclust function in R) based on the gower distance metric (daisy function in R) for the roadmap state assignments across 127 cell types.

Experimental validation using STARR-seq

1. **Selection of CAGE elements**

We generated a library of islet CAGE elements to test in the STARR-seq assay by using two approaches. First, we identified clusters of CAGE tags in each islet sample by simply concatenating tags that occurred within 20bp. We retained clusters with at least two tags in each islet sample. We then merged these cluster coordinates across samples and retained clusters supported by at least 15 samples, representing a highly reproducible set. Second, we complemented this approach by also including the set of FANTOM 5 'robust' CAGE peaks that were also supported by CAGE tags in at least 15 samples. 94% of the selected CAGE robust peak regions were already included in the selected CAGE 20 bp clusters; we reasoned that the remaining 6% of CAGE peaks represented relevant and reproducible CAGE elements missed by the 20 bp concatenation approach. We therefore took the union of these two sets of CAGE elements and created 198 bp oligo sequences centered on each element for cloning into the STARR-seq vector. When a CAGE element was longer than 198 bp, we tiled 198 bp oligos over the element, offset by 100 bp. Through this approach, we included a total of 7,188 CAGE elements (each 198 bp long). We note that these CAGE elements represent slightly different coordinates from the TCs coordinates presented elsewhere in the paper that were identified using the paraclu method. While the paraclu approach of calling TCs was adopted after the STARR-seq experiments were already performed, 6,810 (94.7%) of the CAGE elements included in STARR-seq experiment overlapped the final set of TCs presented in the manuscript.

We synthesized 230-bp oligos (198bp CAGE element flanked by 16bp anchor sequences) (Agilent Technologies). We PCR-amplified oligos to add homology arms for Gibson assembly cloning, and gel-purified the PCR products (i.e. inserts) using the Zymoclean Gel DNA Recovery Kit (Zymo Research). We used the NEBuilder HiFi DNA Assembly Kit (NEB) to assemble the purified inserts and the backbone of STARR-seq plasmid (previously digested with EcoRV). After column purification of the reaction using the DNA Clean and Concentrator-5 kit (Zymo), we transformed it into 10beta electrocompetent cells (NEB), and obtained 1.39 million unique transformants.

We post-barcoded the library by first digesting the library with PmeI, and then by setting up Gibson assembly to insert 16-bp random nucleotides ('barcodes') at the PmeI restriction site. After column purification, we transformed the reaction into electrocompetent cells (NEB), and obtained 1.44 million unique transformants. We prepared the library plasmid for transfection using the ZymoPURE Plasmid Maxiprep Kit (Zymo).

2. Electroporation, RNA isolation and cDNA synthesis

We electroporated 50 ug of barcoded STARR-seq library into 25 million the 832/13 rat insulinoma cell line for each biological replicate (3 replicates), and harvested the cells twenty-four hours later. We isolated total RNA using TRIZOL reagent (Life Technologies) following the manufacturer's protocol up to phase separation. After phase separation, we transferred the aqueous phase of the solution to a new 1.5 mL Eppendorf tube, added 1:1 volume of 100% ethanol, and then column-purified using the Direct-zol RNA Miniprep Kit (Zymo Research Corporation, Irvine, CA) following the manufacturer's protocol. We further purified mRNA using Dynabeads oligo(dT) beads (Thermo Fisher, Carlsbad, CA). We treated 2 ug of mRNA with RNase-free DNaseI (Invitrogen, Carlsbad CA) to eliminate possible plasmid DNA contamination, and then reverse transcribed 1ug into cDNA using the SuperScript III First-Strand Synthesis kit (Invitrogen) with a custom primer that specifically recognizes 'STARR transcripts' (i.e. mRNA that had been transcribed from the STARR-seq plasmids). The other 1ug of mRNA was used in a enzyme-negative reaction to determine. To eliminate any residual plasmid contamination in cDNA samples, we treated cDNA with DpnI (NEB), and purified the reaction using the DNA Clean and Concentrator-5 kit (Zymo).

3. Construction of Illumina sequencing library

We constructed Illumina sequencing library via two serial rounds of PCR. In the first round, we used a primer set to specifically amplify STARR transcripts using cDNA as starting material. In the second round, we used a primer set to append the P5/P7 Illumina sequences using the PCR product from the first round as starting material. In both rounds, we PCR-amplified the fragments until the amplification curve reached a mid-log phase, and then purified the products for subsequent steps using the DNA Clean and Concentrator-5 kit (Zymo).

4. **CAGE element-barcode pairing**

To identify the barcodes corresponding to each CAGE element in the STARR-seq plasmid, 1ng of each library constructed was used in a polymerase chain reaction with primers flanking the allele and the barcode to generate fragments which were subsequently gel verified and extracted using Zymo gel extraction kit (Zymo). 25ng of the purified product was subjected to self-ligation at 16 °C overnight in a total volume of 50uls and column purified using Qiagen (Qiagen) PCR purification kit as per manufacturers recommendations. The purified fragments were subsequently treated with 10U of Plasmid-Safe ATP-Dependent DNase for 1 hour in the presence of 25mM ATP to remove unligated linear DNA fragments. 1ul of the recircularized fragments

were subjected to another round of PCR resulting in a smaller fragment. Briefly an aliquot of this product was diluted 1:10 in DNAse/RNase free water and amplified in a PCR reaction, with Illumina P5 and P7 adapters, until saturation to generate libraries. The libraries were subsequently column purified, quantified and sequenced.

5. Data analysis

The STARR-seq barcode sequencing data included the input DNA barcode library along with three cDNA barcode libraries representing three biological replicated. We processed this data through a custom pipeline which quantified barcode counts while accounting for sequencing errors. We extracted the DNA barcodes from the input DNA library (first 16 bp of the read-1 fastq file) and clustered these at an edit distance of 0 followed by computing the DNA counts for each read group of DNA sequencing files. We then aggregated the read groups and collapsed counts for barcodes using the sequence clustering algorithm Starcode (https://github.com/gui11aume/starcode) (Zorita *et al.* 2015). We repeated this process for the cDNA barcode counts for each replicate, with the added step of removing PCR-duplicated barcodes using the UMI information (UMI sequence was the reverse complement of the first 6bp of the read-2 fastq file). The pipeline is shared at \url{https://github.com/ParkerLab/STARR-seq-Analysis-Pipeline}.

We matched the barcodes with CAGE inserts using results from CAGE insert-barcode pairing experiment (data file “cage_insert_barcode_pairing.tsv” in GEO GSE137693). We first retained barcodes with at least 10 DNA counts, and further retained CAGE elements that had at least two such qualifying barcodes. This was the set of N=3,446 CAGE elements quantifiable in our assay. To quantify STARR-seq activities from these count-based data, we used the tool MPRAnalyze (version 1.3.1) (\url{https://github.com/YosefLab/MPRAnalyze}) (Ashuach *et al.* 2019) that models DNA and RNA counts in a negative binomial generalized linear model. This approach is more robust than using metrics such as the aggregated ratio, which is the ratio of the sum of RNA counts across barcodes divided by the sum of DNA counts across barcodes and loses the statistical power provided by multiple barcodes per tested element; and the mean ratio, which is the mean of the observed RNA/DNA ratios across barcodes which can be quite sensitive to low counts and noise. We corrected for library depth for the three replicates using upper quartile normalization via the ‘estimateDepthFactors’ function in MPRAnalyze. STARR-seq activity is quantified by estimating the transcription rate for each element in the dataset, followed by identifying active elements that induce a higher transcription by testing against a null. MPRAnalyze fits two nested generalized linear models - the DNA model estimates plasmid copy numbers, and the RNA model is estimates transcription rate. We included barcode information in the DNA model which allows different estimated counts for each barcode, and increases the statistical power of the model. Replicate information was included in the RNA model. MPRAnalyze then tests the transcriptional activity of each element against a null distribution and computing Z and Median-Absolute-Deviation (MAD) scores. The null is based on the assumption that the mode of the distribution of transcription rate estimates is the center of the null distribution, and that values lower than the mode all belong to the null. Thus, values lower than the mode are used to estimate the variance of the null.

Lasso regression

We used lasso regression to model TC element STARR-seq activity z scores as a function of TF motif occurrences within the TC elements. Lasso regression is useful when a large number of features such as hundreds of TF motifs in this case are included because it imposes a constraint on the model parameters

causing regression coefficients for some variables to shrink toward zero. Features with non-zero regression coefficients are most strongly associated with the response variable.

We utilized a set of 1,995 TF motifs including their position weight matrices (PWMs), available from ENCODE, JASPAR and Jolma datasets (Jolma *et al.* 2013; Kheradpour and Kellis 2014; Mathelier *et al.* 2016), which we have also used previously (Scott *et al.* 2016; Varshney *et al.* 2017). In order to reduce motif redundancy, we performed PWM clustering in our motif database using the matrix-clustering tool from RSAT (Castro-Mondragon *et al.* 2017), with parameters -lth cor 0.7 -lth Ncor 0.7. For each of the 540 clusters obtained, we retained the motif with the highest total PWM information content. Because STARR-seq is an episomal assay and doesn't recapitulate the native chromatin context, we quantified overlaps of each TC element with sequence motif scans rather than ATAC-seq informed footprint motifs. We scanned each of these motifs on the hg19 reference using FIMO (Grant *et al.* 2011). We used the nucleotide frequencies from the hg19 reference and the default p value cutoff of 10^{-4}.

To quantify motif occurrences within each TC element, we considered the -log10(P value) of each motif occurrence from FIMO. Since the FIMO motif scan p-values depend on the motif length and information content etc., these log transformed P values are not directly comparable across motifs. We therefore inverse normalized the -log10(P values) for occurrences of each motif using the RNOmni package (version 0.7.1) to obtain motif 'scores' on a comparable normal scale. P value = 1 was included for each motif to obtain the score corresponding to no motif occurrence on the transformed scale. For each TF motif, we aligned the hg19 scan occurrences with each islet TC STARR-seq element using BedTools intersect and recorded the corresponding motif scores. We added the scores for TC elements that overlapped multiple occurrences for the said motif. We again inverse normalized the motif overlap score vector across the input CAGE elements for each TF motif so that the regression coefficients could be comparable across motifs. The lasso regression was run using the glmnet package (version 2.0-16) with default parameters (specifically, alpha=1, which corresponds to the lasso regression). Lambda was determined automatically by glmnet as the lambda that belonged to the model with the lowest mean cross validated error.

fGWAS analyses and fine-mapping

We used the fGWAS (version 0.3.6) (Pickrell 2014) tool to compute enrichment of GWAS and islet eQTL data in TC-related annotations along with computing conditional enrichment and fine mapping analyses. fGWAS employs a Bayesian hierarchical model to determine shared properties of loci affecting a trait. The model uses association summary level data, divides the genome into windows generally larger than the expected LD patterns in the population. The method assumes that there is either a single causal SNP in a window or none. The model defines the prior probabilities that an association lies in a genomic window and that a SNP within it is causal. These probabilities are allowed to depend on genomic annotations, and are estimated based on enrichment patterns of annotations across the genome using a Bayes approach.

We obtained publicly available summary data for T2D GWAS (Mahajan *et al.* 2018) and islet eQTL (Varshney *et al.* 2017) and organized it in the format required by fGWAS. We used fGWAS with default parameters for enrichment analyses for individual annotations in Figure 4 A. For each individual annotation, the model provided the maximum likelihood enrichment parameter. Annotations were considered as significantly enriched if the log2(parameter estimate) and respective 95% confidence intervals were above zero or significantly

depleted if the log2(parameter estimate) and respective 95% confidence intervals were below zero. We performed conditional analyses using the '-cond' option.

To reweight GWAS summary data based on functional annotation overlap, we used the '-print' option in an fGWAS model run after including multiple annotations that were individually significantly enriched. We included Active TSS, active enhancer, stretch enhancer, quiescent and polycomb repressed annotations along with ATAC-seq or TCs in a model to derive enrichment priors which can then be used to evaluate both the significance and functional impact of associated variants in GWAS regions; such that variants overlapping more enriched annotations carry extra weight.

References

Adli M., J. Zhu, and B. E. Bernstein, 2010 Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. Nat. Methods 7: 615–618. https://doi.org/10.1038/nmeth.1478

Anders S., P. T. Pyl, and W. Huber, 2015 HTSeq--a Python framework to work with high-throughput sequencing data. Bioinforma. Oxf. Engl. 31: 166–169. https://doi.org/10.1093/bioinformatics/btu638

Andersson R., C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, *et al.*, 2014 An atlas of active enhancers across human cell types and tissues. Nature 507: 455. https://doi.org/10.1038/nature12787

Arnold C. D., D. Gerlach, C. Stelzer, Ł. M. Boryń, M. Rath, *et al.*, 2013 Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. Science 339: 1074–1077. https://doi.org/10.1126/science.1232542

Ashuach T., D. S. Fischer, A. Kreimer, N. Ahituv, F. J. Theis, *et al.*, 2019 MPRAnalyze: statistical framework for massively parallel reporter assays. Genome Biol. 20: 183. https://doi.org/10.1186/s13059-019-1787-z

Benjamini Y., and D. Yekutieli, 2001 The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29: 1165–1188. https://doi.org/10.1214/aos/1013699998

Buenrostro J. D., P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, 2013 Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10: 1213–1218. https://doi.org/10.1038/nmeth.2688

Buniello A., J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, *et al.*, 2019 The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.

Nucleic Acids Res. 47: D1005–D1012. https://doi.org/10.1093/nar/gky1120

Bunt M. van de, J. E. M. Fox, X. Dai, A. Barrett, C. Grey, *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: e1005694. https://doi.org/10.1371/journal.pgen.1005694

Carrat G. R., M. Hu, M.-S. Nguyen-Tu, P. Chabosseau, K. J. Gaulton, *et al.*, 2017 Decreased STARD10 Expression Is Associated with Defective Insulin Secretion in Humans and Mice. Am. J. Hum. Genet. 100: 238–256. https://doi.org/10.1016/j.ajhg.2017.01.011

Castro-Mondragon J. A., S. Jaeger, D. Thieffry, M. Thomas-Chollier, and J. van Helden, 2017 RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. Nucleic Acids Res. 45: e119–e119. https://doi.org/10.1093/nar/gkx314

Chang C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, *et al.*, 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4: 1–16. https://doi.org/10.1186/s13742-015-0047-8

Core L. J., J. J. Waterfall, and J. T. Lis, 2008 Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. Science 322: 1845–1848. https://doi.org/10.1126/science.1162228

Core L. J., A. L. Martins, C. G. Danko, C. Waters, A. Siepel, *et al.*, 2014 Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat. Genet. 46: 1311–1320. https://doi.org/10.1038/ng.3142

Creyghton M. P., A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, *et al.*, 2010 Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. U. S. A. 107: 21931–21936. https://doi.org/10.1073/pnas.1016071107

Dobin A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, *et al.*, 2013 STAR: ultrafast universal RNA-seq aligner. Bioinforma. Oxf. Engl. 29: 15–21. https://doi.org/10.1093/bioinformatics/bts635

Ernst J., and M. Kellis, 2010 Discovery and characterization of chromatin states for systematic annotation of

the human genome. Nat. Biotechnol. 28: 817–825. https://doi.org/10.1038/nbt.1662

Ernst J., P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, *et al.*, 2011 Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43–49. https://doi.org/10.1038/nature09906

Ernst J., and M. Kellis, 2012 ChromHMM: automating chromatin state discovery and characterization. Nat. Methods 9: 215–216. https://doi.org/10.1038/nmeth.1906

Ernst J., A. Melnikov, X. Zhang, L. Wang, P. Rogov, *et al.*, 2016 Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nat. Biotechnol. 34: 1180–1190. https://doi.org/10.1038/nbt.3678

Fadista J., P. Vikman, E. O. Laakso, I. G. Mollet, J. L. Esguerra, *et al.*, 2014 Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. Proc. Natl. Acad. Sci. U. S. A. 111: 13924–13929.

Finucane H. K., B. Bulik-Sullivan, A. Gusev, G. Trynka, Y. Reshef, *et al.*, 2015 Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. 47: 1228–1235. https://doi.org/10.1038/ng.3404

Frith M. C., E. Valen, A. Krogh, Y. Hayashizaki, P. Carninci, *et al.*, 2008 A code for transcription initiation in mammalian genomes. Genome Res. 18: 1–12. https://doi.org/10.1101/gr.6831208

Gershengorn M. C., A. A. Hardikar, C. Wei, E. Geras-Raaka, B. Marcus-Samuels, *et al.*, 2004 Epithelial-to-Mesenchymal Transition Generates Proliferative Human Islet Precursor Cells. Science 306: 2261–2264.

Grant C. E., T. L. Bailey, and W. S. Noble, 2011 FIMO: scanning for occurrences of a given motif. Bioinformatics 27: 1017–1018. https://doi.org/10.1093/bioinformatics/btr064

Gusev A., S. H. Lee, G. Trynka, H. Finucane, B. J. Vilhjálmsson, *et al.*, 2014 Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. 95: 535–552. https://doi.org/10.1016/j.ajhg.2014.10.004

Harrow J., A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, *et al.*, 2012 GENCODE: The reference

human genome annotation for The ENCODE Project. Genome Res. 22: 1760–1774. https://doi.org/10.1101/gr.135350.111

Hartley S. W., and J. C. Mullikin, 2015 QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. BMC Bioinformatics 16: 224. https://doi.org/10.1186/s12859-015-0670-5

Heger A., C. Webber, M. Goodson, C. P. Ponting, and G. Lunter, 2013 GAT: a simulation framework for testing the association of genomic intervals. Bioinformatics 29: 2046–2048. https://doi.org/10.1093/bioinformatics/btt343

Hnisz D., B. J. Abraham, T. I. Lee, A. Lau, V. Saint-André, *et al.*, 2013 Super-Enhancers in the Control of Cell Identity and Disease. Cell 155: 934–947. https://doi.org/10.1016/j.cell.2013.09.053

Hoffman M. M., J. Ernst, S. P. Wilder, A. Kundaje, R. S. Harris, *et al.*, 2013 Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 41: 827–841. https://doi.org/10.1093/nar/gks1284

Hsieh C.-L., T. Fei, Y. Chen, T. Li, Y. Gao, *et al.*, 2014 Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. Proc. Natl. Acad. Sci. 111: 7319–7324. https://doi.org/10.1073/pnas.1324151111

Jolma A., J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, *et al.*, 2013 DNA-Binding Specificities of Human Transcription Factors. Cell 152: 327–339. https://doi.org/10.1016/j.cell.2012.12.009

Kaikkonen M. U., N. J. Spann, S. Heinz, C. E. Romanoski, K. A. Allison, *et al.*, 2013 Remodeling of the Enhancer Landscape during Macrophage Activation Is Coupled to Enhancer Transcription. Mol. Cell 51: 310–325. https://doi.org/10.1016/j.molcel.2013.07.010

Kent W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, *et al.*, 2002 The human genome browser at UCSC. Genome Res. 12: 996–1006. https://doi.org/10.1101/gr.229102

Kheradpour P., and M. Kellis, 2014 Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. 42: 2976–2987. https://doi.org/10.1093/nar/gkt1249

Kim T.-K., M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, *et al.*, 2010 Widespread transcription at neuronal activity-regulated enhancers. Nature 465: 182–187. https://doi.org/10.1038/nature09033

Koch F., R. Fenouil, M. Gut, P. Cauchy, T. K. Albert, *et al.*, 2011 Transcription initiation platforms and GTF

recruitment at tissue-specific enhancers and promoters. Nat. Struct. Mol. Biol. 18: 956–963. https://doi.org/10.1038/nsmb.2085

Kulzer J. R., M. L. Stitzel, M. A. Morken, J. R. Huyghe, C. Fuchsberger, *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: 186–197. https://doi.org/10.1016/j.ajhg.2013.12.011

Landt S. G., G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, *et al.*, 2012 ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 22: 1813–1831. https://doi.org/10.1101/gr.136184.111

Li H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio.

Li W., D. Notani, Q. Ma, B. Tanasa, E. Nunez, *et al.*, 2013 Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature 498: 516–520. https://doi.org/10.1038/nature12210

Lindblad-Toh K., M. Garber, O. Zuk, M. F. Lin, B. J. Parker, *et al.*, 2011 A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476–482. https://doi.org/10.1038/nature10530

Lopes R., R. Agami, and G. Korkmaz, 2017 GRO-seq, A Tool for Identification of Transcripts Regulating Gene Expression. Methods Mol. Biol. Clifton NJ 1543: 45–55. https://doi.org/10.1007/978-1-4939-6716-2_3

Maehara K., and Y. Ohkawa, 2015 agplus: a rapid and flexible tool for aggregation plots. Bioinformatics 31: 3046–3047. https://doi.org/10.1093/bioinformatics/btv322

Mahajan A., D. Taliun, M. Thurner, N. R. Robertson, J. M. Torres, *et al.*, 2018 Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat. Genet. 50: 1505. https://doi.org/10.1038/s41588-018-0241-6

Martin M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17: 10–12. https://doi.org/10.14806/ej.17.1.200

Mathelier A., O. Fornes, D. J. Arenillas, C. Chen, G. Denay, *et al.*, 2016 JASPAR 2016: a major expansion and

update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 44: D110–D115. https://doi.org/10.1093/nar/gkv1176

Matys V., O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, *et al.*, 2006 TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 34: D108-110. https://doi.org/10.1093/nar/gkj143

Melgar M. F., F. S. Collins, and P. Sethupathy, 2011 Discovery of active enhancers through bidirectional expression of short transcripts. Genome Biol. 12: R113. https://doi.org/10.1186/gb-2011-12-11-r113

Melnikov A., A. Murugan, X. Zhang, T. Tesileanu, L. Wang, *et al.*, 2012 Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. 30: 271–277. https://doi.org/10.1038/nbt.2137

Mikhaylichenko O., V. Bondarenko, D. Harnett, I. E. Schor, M. Males, *et al.*, 2018 The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Genes Dev. 32: 42–57. https://doi.org/10.1101/gad.308619.117

Mikkelsen T. S., M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, *et al.*, 2007 Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553–560. https://doi.org/10.1038/nature06008

Murata M., H. Nishiyori-Sueki, M. Kojima-Ishiyama, P. Carninci, Y. Hayashizaki, *et al.*, 2014 Detecting expressed genes using CAGE. Methods Mol. Biol. Clifton NJ 1164: 67–85. https://doi.org/10.1007/978-1-4939-0805-9_7

Parker S. C. J., M. L. Stitzel, D. L. Taylor, J. M. Orozco, M. R. Erdos, *et al.*, 2013 Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc. Natl. Acad. Sci. 110: 17921–17926. https://doi.org/10.1073/pnas.1317023110

Pasquali L., K. J. Gaulton, S. A. Rodríguez-Seguí, L. Mularoni, I. Miguel-Escalada, *et al.*, 2014 Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat. Genet. 46: 136–143. https://doi.org/10.1038/ng.2870

Patti M. E., A. J. Butte, S. Crunkhorn, K. Cusi, R. Berria, *et al.*, 2003 Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and

NRF1. Proc. Natl. Acad. Sci. U. S. A. 100: 8466–8471. https://doi.org/10.1073/pnas.1032913100

Pickrell J. K., 2014 Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. Am. J. Hum. Genet. 94: 559–573. https://doi.org/10.1016/j.ajhg.2014.03.004

Purcell S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, *et al.*, 2007 PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81: 559–575. https://doi.org/10.1086/519795

Quang D. X., M. R. Erdos, S. C. J. Parker, and F. S. Collins, 2015 Motif signatures in stretch enhancers are enriched for disease-associated genetic variants. Epigenetics Chromatin 8: 23. https://doi.org/10.1186/s13072-015-0015-7

Roman T. S., M. E. Cannon, S. Vadlamudi, M. L. Buchkovich, B. N. Wolford, *et al.*, 2017 A Type 2 Diabetes–Associated Functional Regulatory Variant in a Pancreatic Islet Enhancer at the ADCY5 Locus. Diabetes 66: 2521–2530. https://doi.org/10.2337/db17-0464

Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014 Biological insights from 108 schizophrenia-associated genetic loci. Nature 511: 421–427. https://doi.org/10.1038/nature13595

Schmidt E. M., J. Zhang, W. Zhou, J. Chen, K. L. Mohlke, *et al.*, 2015 GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. Bioinformatics 31: 2601–2606. https://doi.org/10.1093/bioinformatics/btv201

Scott L. J., M. R. Erdos, J. R. Huyghe, R. P. Welch, A. T. Beck, *et al.*, 2016 The genetic regulatory signature of type 2 diabetes in human skeletal muscle. Nat. Commun. 7: ncomms11764. https://doi.org/10.1038/ncomms11764

Seabold S., and J. Perktold, 2010 Statsmodels: Econometric and Statistical Modeling with Python. Proc. 9th Python Sci. Conf. 5.

The ENCODE project Consortium, 2012 An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature 489: 57–74. https://doi.org/10.1038/nature11247

The FANTOM Consortium, A. R. R. Forrest, H. Kawaji, M. Rehli, J. K. Baillie, *et al.*, 2014 A promoter-level mammalian expression atlas. Nature 507: 462. https://doi.org/10.1038/nature13182

The Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, *et al.*, 2015 Integrative analysis of 111 reference human epigenomes. Nature 518: 317–330. https://doi.org/10.1038/nature14248

Thurner M., M. van de Bunt, J. M. Torres, A. Mahajan, V. Nylander, *et al.*, 2018 Integration of human pancreatic islet genomic data refines regulatory mechanisms at Type 2 Diabetes susceptibility loci. eLife 7: e31977. https://doi.org/10.7554/eLife.31977

Trynka G., C. Sandor, B. Han, H. Xu, B. E. Stranger, *et al.*, 2013 Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet. 45: 124–130. https://doi.org/10.1038/ng.2504

Udler M. S., J. Kim, M. von Grotthuss, S. Bonàs-Guarch, J. B. Cole, *et al.*, 2018 Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLOS Med. 15: e1002654. https://doi.org/10.1371/journal.pmed.1002654

Varshney A., L. J. Scott, R. P. Welch, M. R. Erdos, P. S. Chines, *et al.*, 2017 Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. Proc. Natl. Acad. Sci. 114: 2301–2306. https://doi.org/10.1073/pnas.1621192114

Varshney A., H. VanRenterghem, P. Orchard, A. P. Boyle, M. L. Stitzel, *et al.*, 2018 Cell Specificity of Human Regulatory Annotations and Their Genetic Effects on Gene Expression. Genetics genetics.301525.2018. https://doi.org/10.1534/genetics.118.301525

Viñuela A., A. Varshney, M. van de Bunt, R. B. Prasad, O. Asplund, *et al.*, 2019 Influence of genetic variants on gene expression in human pancreatic islets – implications for type 2 diabetes. bioRxiv 655670. https://doi.org/10.1101/655670

Visel A., S. Minovitsky, I. Dubchak, and L. A. Pennacchio, 2007 VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic Acids Res. 35: D88–D92. https://doi.org/10.1093/nar/gkl822

Ward L. D., and M. Kellis, 2012 Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science 337: 1675–1678. https://doi.org/10.1126/science.1225057

Yang Y., Z. Su, X. Song, B. Liang, F. Zeng, *et al.*, 2016 Enhancer RNA-driven looping enhances the transcription of the long noncoding RNA DHRS4-AS1, a controller of the DHRS4 gene cluster. Sci.

Rep. 6: 20961. https://doi.org/10.1038/srep20961

Zheng H., J. Fu, P. Xue, R. Zhao, J. Dong, *et al.*, 2015 CNC-bZIP Protein Nrf1-Dependent Regulation of Glucose-Stimulated Insulin Secretion. Antioxid. Redox Signal. 22: 819–831. https://doi.org/10.1089/ars.2014.6017

Zhou V. W., A. Goren, and B. E. Bernstein, 2011 Charting histone modifications and the functional organization of mammalian genomes. Nat. Rev. Genet. 12: 7–18. https://doi.org/10.1038/nrg2905

Zorita E., P. Cuscó, and G. J. Filion, 2015 Starcode: sequence clustering based on all-pairs search. Bioinformatics 31: 1913–1919. https://doi.org/10.1093/bioinformatics/btv053

Figure Legends:

Figure 1: Islet CAGE tag cluster identification. A: Genome browser view of the intronic region of the *ST18* gene as an example locus where an islet TC overlaps an islet ATAC-seq peak and active TSS chromatin state. This TC also overlaps an enhancer element which was validated by the VISTA project (Visel *et al.* 2007). Also shown is the human-mouse-rat conserved TF binding site (TFBS) track from the Transfac Matrix Database (Matys *et al.* 2006). B: Base-pair level overlap between islet CAGE TC territory and FANTOM robust CAGE peaks (identified from 988 CAGE libraries across diverse human cell lines/tissues). C: Fraction of islet TCs that overlap TCs identified in at most x FANTOM tissues. D: Genome browser view of an example locus near the *AP1G2* gene that highlights an islet TC that is also identified in FANTOM tissues (FANTOM TCs track depicts TCs called across 118 human tissues) (green box), occurs in a ATAC-seq peak region in both islets and lymphoblastoid cell line GM12878 (ATAC-seq track) and overlaps active TSS chromatin states across numerous other tissues. Another islet TC ~34 kb distal to the *AP1G2* gene is not identified as a TC in other FANTOM tissues, occurs in an islet ATAC-seq peak and a more islet-specific active enhancer chromatin state region (blue box). E: Enrichment of islet TCs to overlap islet chromatin state annotations and other common annotations. Error bars represent the 95% confidence intervals. Colors represent significant enrichment after Bonferroni correction accounting for 40 total annotations (see Methods for references of common annotations, Supplementary Table 1), nominal enrichment (P value < 0.05) or non-significant enrichment. F: Enrichment of islet TCs to overlap GWAS loci of various disease/traits. Number of independent loci for each trait are noted in parentheses. Colors represent significant enrichment after Bonferroni correction accounting for total 116 traits (see Methods for GWAS data sources, Supplementary Table 2), nominal enrichment (P value < 0.05) or non-significant enrichment. Acronyms: T2D = type 2 diabetes, ASD = Autism Spectrum Disorder, LDL = Low density lipoprotein, HDL = High density lipoprotein, BMI = Body mass index.

Figure 2. Integrating Islet CAGE TCs with other epigenomic information reveals characteristics of transcription initiation. A: Aggregate CAGE profiles over ATAC-seq peak summits. B: Aggregate ATAC-seq profile over TC midpoints. C: Enrichment of TF footprint motifs to overlap 500 bp upstream region (y axis) vs 500 bp

downstream region (x axis) of islet TCs. Colors denote if a TF footprint motif was significantly enriched (5% FDR correction, Benjajmini-Yekutieli method) to overlap only upstream regions, only downstream regions, both or none. D: Chromatin state annotations across 98 Roadmap Epigenomics cell types (using the 18 state 'extended model' (The Roadmap Epigenomics Consortium *et al.* 2015) for TC segments that occur in islet promoter (active, weak, flanking TSS) chromatin states and overlap ATAC-seq peaks. These segments were segregated into those occurring 5kb proximal (left, N=7,064 TC segments) and distal (right, N=443 TC segments) to known protein coding gene TSS (Gencode V19). E: Chromatin state annotations across 98 Roadmap Epigenomics cell types for TC segments that occur in islet enhancer (active, weak or genic enhancer) chromatin states state and overlap ATAC-seq peaks, segregated into those occurring 5kb proximal (left, N=254 TC segments) and distal (right, N=289 TC segments) to known protein coding gene TSS. F: Enrichment of TF footprint motifs to overlap TCs occurring in accessible enhancer (active, weak or genic enhancer) chromatin state regions (y axis) vs TCs occurring in accessible promoter (active, weak, flanking TSS) chromatin state regions (x axis). Colors denote if a TF footprint motif was significantly enriched (5% FDR correction, Benjajmini-Yekutieli method) to overlap only TCs in accessible enhancer regions, only TCs in accessible promoter regions, both or none. G: Aggregate CAGE profiles centered and oriented relative to RFX5_known8 footprint motifs occurring in 5kb TSS distal regions. H: Aggregate CAGE profiles centered and oriented relative to ELK4_1 footprint motifs.

Figure 3. Experimental validation of CAGE elements using STARR-seq assay: A: (Top) Number and fraction of CAGE elements that show significant (5% FDR), nominal ($P < 0.05$) or non-significant transcriptional activity in the STARR-seq assay performed in rat beta cell insulinoma (INS1 832/13) cell line model. (Bottom) Proportion of CAGE elements overlapping promoter (active, weak or flanking TSS), enhancer (active, weak or genic enhancer) or other chromatin states that showed significant transcriptional activity in the STARR-seq assay. B. STARR-seq activity Z scores for CAGE elements overlapping in promoter, enhancer or other chromatin states. C: STARR-seq activity Z scores for CAGE elements that overlap ATAC-seq peak vs CAGE elements that do not overlap peaks. D: STARR-seq activity Z scores for CAGE elements based on position relative to known protein coding gene TSSs (5kb TSS proximal or distal, Gencode V19). E: (Top) An overview of the lasso regression model to predict the STARR-seq activity Z scores of CAGE elements as a function of the TF motif scan scores within the element. (Bottom) Top 30 TF motifs with non-zero coefficients from the model. F: An example locus on chr17, where the nearest gene *RPH3AL* lies ~6kb away, an islet TC overlaps active TSS and enhancer chromatin states and an ATAC-seq peak. Elements overlapping this TC showed significant transcriptional activity in the STARR-seq assay. The CAGE profile coincides with islet mRNA profile that is detected despite no known gene annotation in the region and the nearest protein coding gene is ~6kb away. Also shown are occurrences of TF motifs with positive or negative lasso regression coefficients from the analysis in E. G: The promoter locus of the genes *DCAF16* and *NCAPG*, where an islet TC is identified in the vicinity of two T2D GWAS SNPs. The TC overlaps an ATAC-seq peak and active TSS chromatin states and the TC element showed significant activity in the STARR-seq assay. Also shown are TF motifs with positive or negative lasso regression coefficients from the analysis in E

Figure 4. Islet TCs supplement functional understanding of GWAS and eQTL associations and help nominate causal variants: Enrichment of T2D GWAS (left) or islet eQTL (right) loci in annotations that comprise different levels of epigenomic information, including including chromatin state, ATAC-seq and TCs. Annotations defined using combinations of these datasets are depicted with different colors on the y axis. Enrichment was calculated using fGWAS (Pickrell 2014) using summary statistics from GWAS (left) (Mahajan *et al.* 2018) or

islet eQTL (right) (Varshney *et al.* 2017). Error bars denote the 95% confidence intervals. C: fGWAS conditional enrichment analysis testing the contribution of islet TC or ATAC-seq peak annotations after conditioning on histone-only based annotations such as active TSS and active enhancer chromatin states in islets. D: Maximum SNP PPA per FGlu GWAS locus after functional re-weighting using a model with islet chromatin states and ATAC-seq peak annotations (x axis) or chromatin states and TC (y axis) annotations. Continued on the next page. E: Genome browser view of the *STARD10* gene locus where T2D and Fasting Glucose GWAS SNPs and eQTL SNPs for the *STARD10* gene occur (left). *STARD10* eQTL Lead and LD r^2>0.8 proxy SNPs are shown in the eQTL SNP track. Genome browser view on the right shows the region zooming in on the lead eQTL SNP rs11603334 and another SNP rs1552225 (LD r^2=1 with the lead SNP) which overlaps an islet TC. Functional reweighting of Fasting Glucose GWAS data using chromatin state, ATAC-seq and TC data resulted in the PPA of the SNP rs1552225 = 0.772.

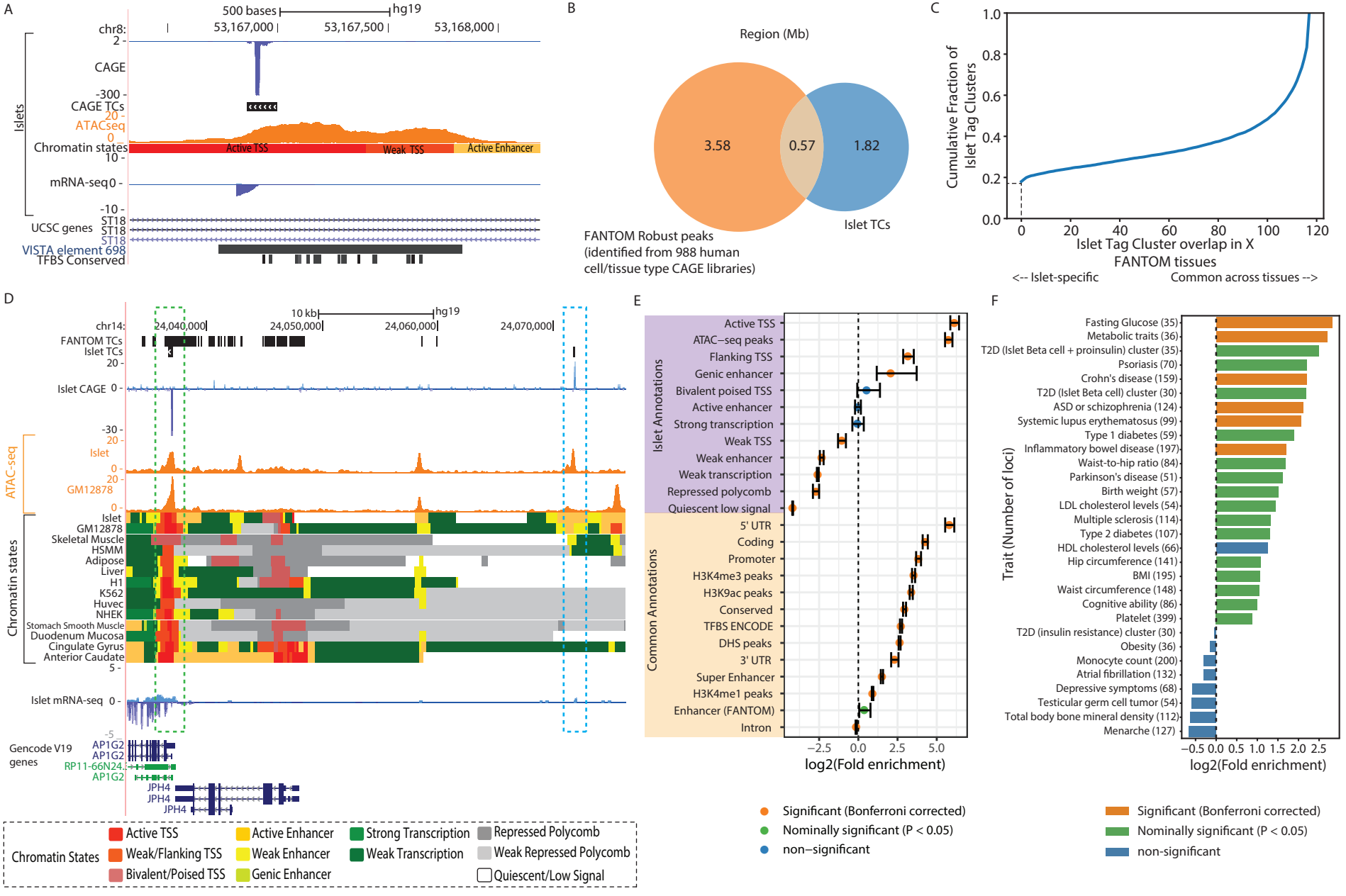
Supplementary Figure 1: Islet TC identification using CAGE data across multiple samples. TC segments called using the paraclu method in each of the 57 selected islet samples were merged in a strand specific manner. Shown here is the number of merged TC segments that overlap TCs in x or more islet samples. We required TC overlap in a minimum of 10 islet samples to include a segment in the aggregate list of islet TCs.

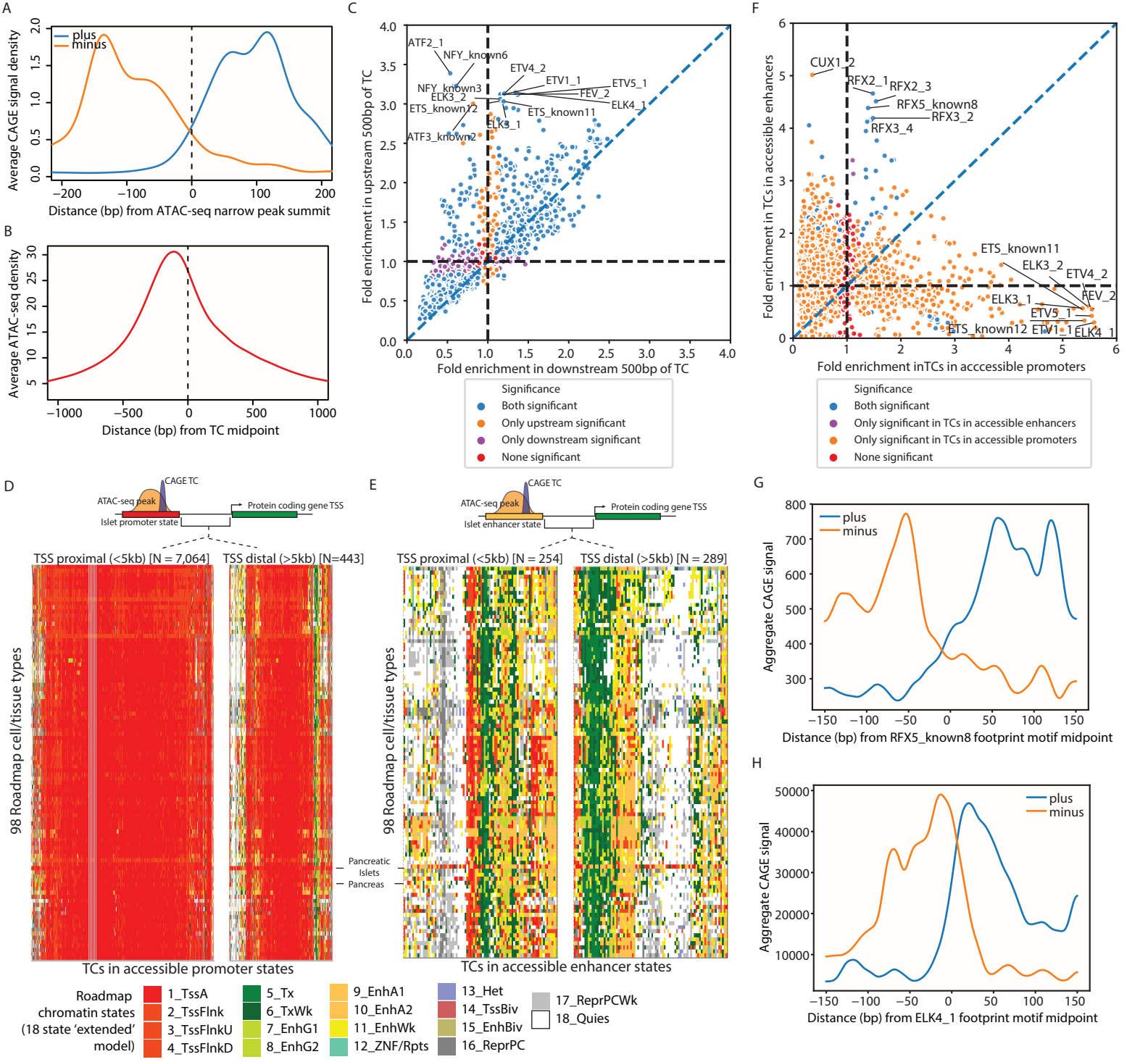
Supplementary Figure 2: Distribution of islet TC lengths.

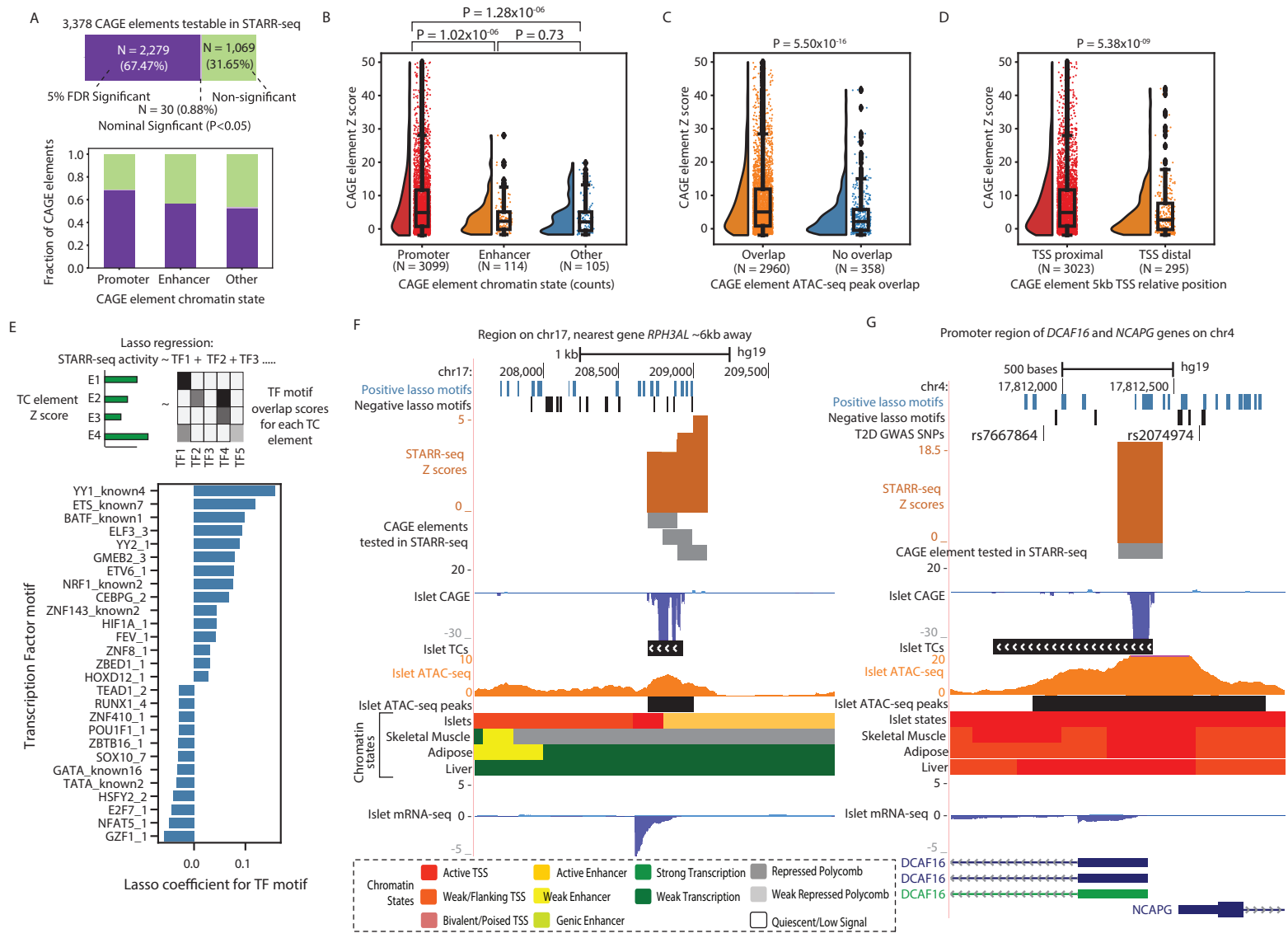
Supplementary Figure 3: 11 chromatin state model. Shown are the emission probabilities of each of the five histone marks, chromatin state annotation and the percent genome coverage of each state

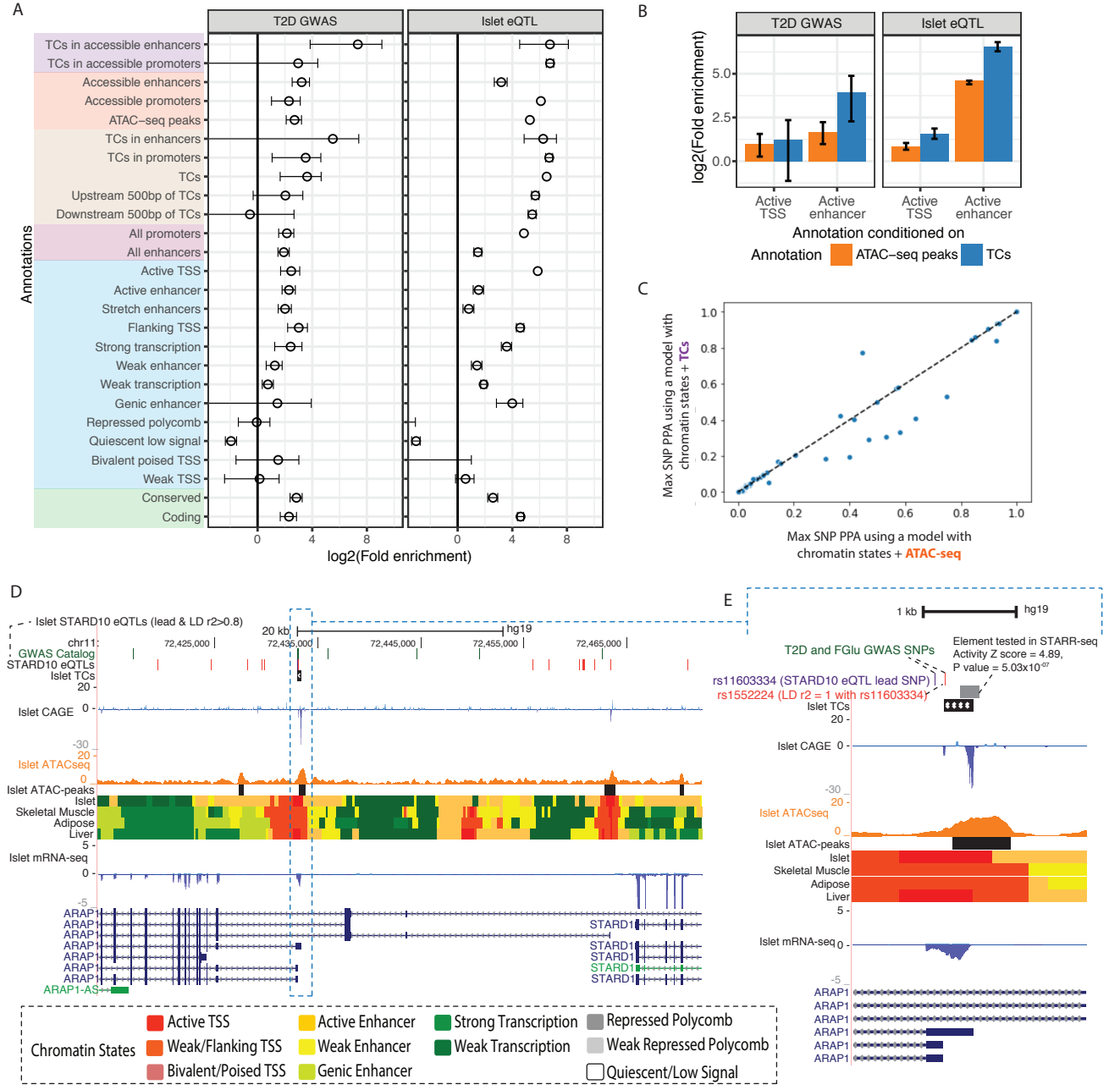
Supplementary Figure 4. Correlation between replicates for normalized total RNA counts for CAGE inserts.

Supplementary Figure 5. *LCORL* T2D GWAS locus. Genome browser shot of the *LCORL* T2D GWAS locus showing the lead SNP (green, GWAS SNP track) along with r2 > 0.8 proxy SNPs. Also shown are islet CAGE, TC, ATAC-seq, chromatin state annotations.


A
500 bases
hg19
chr8:
53,167,000
53,167,500
53,168,000
Islets
CAGE
CAGE TCs
ATACseq
Chromatin states
Active TSS
Weak TSS
Active Enhancer
mRNA-seq
UCSC genes
ST18
VISTA element 698
TFBS Conserved
B
Region (Mb)
3.58
0.57
1.82
Islet TCs
FANTOM Robust peaks
(identified from 988 human
cell/tissue type CAGE libraries)
C
Cumulative Fraction of
Islet Tag Clusters
Islet Tag Cluster overlap in X
FANTOM tissues
<-- Islet-specific
Common across tissues -->
D
10 kb
hg19
chr14:
24,040,000
24,050,000
24,060,000
24,070,000
FANTOM TCs
Islet TCs
Islet CAGE
ATAC-seq
Islet
GM12878
Chromatin states
Skeletal Muscle
HSMM
Adipose
Liver
H1
K562
Huvec
NHEK
Stomach Smooth Muscle
Duodenum Mucosa
Cingulate Gyrus
Anterior Caudate
Islet mRNA-seq
Gencode V19
genes
AP1G2
RP11-66N24.
JPH4
Chromatin States
Active TSS
Weak/Flanking TSS
Bivalent/Poised TSS
Active Enhancer
Weak Enhancer
Genic Enhancer
Strong Transcription
Weak Transcription
Repressed Polycomb
Weak Repressed Polycomb
Quiescent/Low Signal
E
Islet Annotations
Active TSS
ATAC–seq peaks
Flanking TSS
Genic enhancer
Bivalent poised TSS
Active enhancer
Strong transcription
Weak TSS
Weak enhancer
Weak transcription
Repressed polycomb
Quiescent low signal
Common Annotations
5' UTR
Coding
Promoter
H3K4me3 peaks
H3K9ac peaks
Conserved
TFBS ENCODE
DHS peaks
3' UTR
Super Enhancer
H3K4me1 peaks
Enhancer (FANTOM)
Intron
log2(Fold enrichment)
Significant (Bonferroni corrected)
Nominally significant (P < 0.05)
non–significant
F
Trait (Number of loci)
Fasting Glucose (35)
Metabolic traits (36)
T2D (Islet Beta cell + proinsulin) cluster (35)
Psoriasis (70)
Crohn's disease (159)
T2D (Islet Beta cell) cluster (30)
ASD or schizophrenia (124)
Systemic lupus erythematosus (99)
Type 1 diabetes (59)
Inflammatory bowel disease (197)
Waist-to-hip ratio (84)
Parkinson's disease (51)
Birth weight (57)
LDL cholesterol levels (54)
Multiple sclerosis (114)
Type 2 diabetes (107)
HDL cholesterol levels (66)
Hip circumference (141)
BMI (195)
Waist circumference (148)
Cognitive ability (86)
Platelet (399)
T2D (insulin resistance) cluster (30)
Obesity (36)
Monocyte count (200)
Atrial fibrillation (132)
Depressive symptoms (68)
Testicular germ cell tumor (54)
Total body bone mineral density (112)
Menarche (127)
log2(Fold enrichment)
Significant (Bonferroni corrected)
Nominally significant (P < 0.05)
non-significant


A
plus
minus
Average CAGE signal density
Distance (bp) from ATAC-seq narrow peak summit
B
Average ATAC-seq density
Distance (bp) from TC midpoint
C
Fold enrichment in upstream 500bp of TC
Fold enrichment in downstream 500bp of TC
ATF2_1
NFY_known6
ETV4_2
ETV1_1
ETV5_1
NFY_known3
ELK3_2
FEV_2
ETS_known12
ELK4_1
ETS_known11
ELK3_1
ATF3_known2
Significance
Both significant
Only upstream significant
Only downstream significant
None significant
F
Fold enrichment in TCs in accessible enhancers
Fold enrichment inTCs in acccessible promoters
CUX1_2
RFX2_1
RFX2_3
RFX5_known8
RFX3_2
RFX3_4
ETS_known11
ELK3_2
ETV4_2
FEV_2
ELK3_1
ETV5_1
ETS_known12
ETV1_1
ELK4_1
Significance
Both significant
Only significant in TCs in accessible enhancers
Only significant in TCs in accessible promoters
None significant
D
CAGE TC
ATAC-seq peak
Protein coding gene TSS
Islet promoter state
TSS proximal (<5kb) [N = 7,064]
TSS distal (>5kb) [N=443]
98 Roadmap cell/tissue types
Pancreatic Islets
Pancreas
TCs in accessible promoter states
E
CAGE TC
ATAC-seq peak
Protein coding gene TSS
Islet enhancer state
TSS proximal (<5kb) [N = 254]
TSS distal (>5kb) [N = 289]
98 Roadmap cell/tissue types
TCs in accessible enhancer states
Roadmap chromatin states (18 state 'extended' model)
1_TssA
2_TssFlnk
3_TssFlnkU
4_TssFlnkD
5_Tx
6_TxWk
7_EnhG1
8_EnhG2
9_EnhA1
10_EnhA2
11_EnhWk
12_ZNF/Rpts
13_Het
14_TssBiv
15_EnhBiv
16_ReprPC
17_ReprPCWk
18_Quies
G
Aggregate CAGE signal
Distance (bp) from RFX5_known8 footprint motif midpoint
H
Aggregate CAGE signal
Distance (bp) from ELK4_1 footprint motif midpoint


A
3,378 CAGE elements testable in STARR-seq
N = 2,279 (67.47%)
N = 1,069 (31.65%)
5% FDR Significant
Non-significant
N = 30 (0.88%)
Nominal Signficant (P<0.05)
Fraction of CAGE elements
Promoter
Enhancer
Other
CAGE element chromatin state
B
P = 1.28x10-06
P = 1.02x10-06
P = 0.73
CAGE element Z score
Promoter (N = 3099)
Enhancer (N = 114)
Other (N = 105)
CAGE element chromatin state (counts)
C
P = 5.50x10-16
Overlap (N = 2960)
No overlap (N = 358)
CAGE element ATAC-seq peak overlap
D
P = 5.38x10-09
TSS proximal (N = 3023)
TSS distal (N = 295)
CAGE element 5kb TSS relative position
E
Lasso regression:
STARR-seq activity ~ TF1 + TF2 + TF3 .....
TC element Z score
E1
E2
E3
E4
TF1
TF2
TF3
TF4
TF5
TF motif overlap scores for each TC element
Transcription Factor motif
YY1_known4
ETS_known7
BATF_known1
ELF3_3
YY2_1
GMEB2_3
ETV6_1
NRF1_known2
CEBPG_2
ZNF143_known2
HIF1A_1
FEV_1
ZNF8_1
ZBED1_1
HOXD12_1
TEAD1_2
RUNX1_4
ZNF410_1
POU1F1_1
ZBTB16_1
SOX10_7
GATA_known16
TATA_known2
HSFY2_2
E2F7_1
NFAT5_1
GZF1_1
Lasso coefficient for TF motif
F
Region on chr17, nearest gene RPH3AL ~6kb away
1 kb
hg19
chr17:
Positive lasso motifs
Negative lasso motifs
STARR-seq Z scores
CAGE elements tested in STARR-seq
Islet CAGE
Islet TCs
Islet ATAC-seq
Islet ATAC-seq peaks
Chromatin states
Islets
Skeletal Muscle
Adipose
Liver
Islet mRNA-seq
Chromatin States
Active TSS
Weak/Flanking TSS
Bivalent/Poised TSS
Active Enhancer
Weak Enhancer
Genic Enhancer
Strong Transcription
Weak Transcription
Repressed Polycomb
Weak Repressed Polycomb
Quiescent/Low Signal
G
Promoter region of DCAF16 and NCAPG genes on chr4
500 bases
hg19
chr4:
Positive lasso motifs
Negative lasso motifs
T2D GWAS SNPs
rs7667864
rs2074974
STARR-seq Z scores
CAGE element tested in STARR-seq
Islet CAGE
Islet TCs
Islet ATAC-seq
Islet ATAC-seq peaks
Islet states
Skeletal Muscle
Adipose
Liver
Islet mRNA-seq
DCAF16
NCAPG


A
T2D GWAS
Islet eQTL
Annotations
TCs in accessible enhancers
TCs in accessible promoters
Accessible enhancers
Accessible promoters
ATAC−seq peaks
TCs in enhancers
TCs in promoters
TCs
Upstream 500bp of TCs
Downstream 500bp of TCs
All promoters
All enhancers
Active TSS
Active enhancer
Stretch enhancers
Flanking TSS
Strong transcription
Weak enhancer
Weak transcription
Genic enhancer
Repressed polycomb
Quiescent low signal
Bivalent poised TSS
Weak TSS
Conserved
Coding
0
4
8
log2(Fold enrichment)
B
T2D GWAS
Islet eQTL
log2(Fold enrichment)
5.0
2.5
0.0
Active TSS
Active enhancer
Annotation conditioned on
Annotation
ATAC−seq peaks
TCs
C
Max SNP PPA using a model with chromatin states + TCs
Max SNP PPA using a model with chromatin states + ATAC-seq
0.0
0.2
0.4
0.6
0.8
1.0
D
Islet STARD10 eQTLs (lead & LD r2>0.8)
20 kb
hg19
chr11
72,425,000
72,435,000
72,445,000
72,455,000
72,465,000
GWAS Catalog
STARD10 eQTLs
Islet TCs
Islet CAGE
Islet ATACseq
Islet ATAC-peaks
Islet
Skeletal Muscle
Adipose
Liver
Islet mRNA-seq
ARAP1
ARAP1-AS
STARD1
STARD1
STARD1
STARD1
E
1 kb
hg19
T2D and FGlu GWAS SNPs
Element tested in STARR-seq
Activity Z score = 4.89,
P value = 5.03x10−07
rs11603334 (STARD10 eQTL lead SNP)
rs1552224 (LD r2 = 1 with rs11603334)
Islet TCs
Islet CAGE
Islet ATACseq
Islet ATAC-peaks
Islet
Skeletal Muscle
Adipose
Liver
Islet mRNA-seq
ARAP1
Chromatin States
Active TSS
Weak/Flanking TSS
Bivalent/Poised TSS
Active Enhancer
Weak Enhancer
Genic Enhancer
Strong Transcription
Weak Transcription
Repressed Polycomb
Weak Repressed Polycomb
Quiescent/Low Signal