Supplementary Information

A transcription start site map in human pancreatic islets reveals functional regulatory signatures

Arushi Varshney[1,2], Yasuhiro Kyono[1,3], Venkateswaran Ramamoorthi Elangovan[1], Collin Wang[1,4], Michael R. Erdos[5], Narisu Narisu[5], Ricardo D'Oliveira Albanus[1], Peter Orchard[1], Michael L. Stitzel[6], Francis S. Collins[5], Jacob O. Kitzman[1,2], Stephen C. J. Parker[1,2,*]

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

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

3 Current address: Tempus Labs, Inc., Chicago, IL, USA

4 Current address: Columbia University, New York, NY, USA

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

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

* Corresponding author:
Stephen C. J. Parker,
Associate Professor,
Department of Computational Medicine & Bioinformatics,
Department of Human Genetics,
University of Michigan
100 Washtenaw Ave.
2049 Palmer Commons Building
Ann Arbor, MI 48109
734-647-3144 (phone)
734-615-6553 (fax)
scjp@umich.edu

CAGE data 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 (minimum) 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 contamination, we mapped to the *E. coli* chromosome (genome assembly GCA_000005845.2) with bwa mem (v. 0.7.15; options: -M) and 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) (1). 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)(2). 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 (3) for all downstream analyses.

Tag cluster identification

We used the paraclu method to identify clusters of CAGE tags (CAGE TCs) (4). 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 TCs ('singletons') if supported by >2 tags. We then merged the TCs 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 TCs with the number of samples across which support was required to consider the segment (Supplementary Figure 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 TC regions used in all downstream analyses.

FANTOM CAGE datasets

We obtained the set of 'robust CAGE peaks' identified by the FANTOM 5 consortium (5) 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 (5), 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 TCs for each tissue sample using the paraclu method (4) as described above, with the same parameters.

Chromatin state analysis

We generated ChromHMM states in islets along with 30 other tissues previously (6) However, since the statistical power to identify enriched regions depends on the ChIP-seq read depth, in this previous study we uniformly downsampled the ChIP seq datasets to the depth of 20 million reads across the 31 tissues. This allowed comparable chromatin state territories across tissues and ensured that chromatin state territories were not heavily driven by high sequencing depth. Other studies have presented independent chromHMM states in islets, however, there isn't a consensus set of states (7,8). For the current manuscript, we aimed to compare islets and three other relevant tissues (Skeletal Muscle, Liver and Adipose). To maximize power to detect enriched states in this smaller subset of tissues in our study, we downsampled read depth to the mean depth across the four tissues within each mark and then generated harmonized ChromHMM states across the tissue types. These subsampling depths were H3K27ac = 28,123,470, H3K27me3 = 53,603,910, H3K36me3 = 63,763,910, H3K4me1 = 63,441,280 and H3K4me3 = 62,220,940. We list references for the ChIP-seq datasets we utilized in Supplementary table 3. We performed read mapping and integrative chromatin-state analyses in a manner similar to that of our previous reports (6,9) and followed quality control procedures reported by the Roadmap Epigenomics Study (10). 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) (11). We then mapped reads using BWA (version 0.5.8c), removed duplicates using samtools (2), 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) (12). We converted bam files for each dataset to bed using the bamToBed tool, followed by randomly subsampling each dataset bed file to the thresholds mentioned above. 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 (13–15). 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. To assign names to chromatin states that are consistent with previously published states, we performed enrichment analyses in ChromHMM comparing our states with the states reported previously (6) for the four matched tissues. We assigned each state with the state name that was most strongly enriched to overlap that state.

ATAC-seq data analysis

We used previously published chromatin accessibility data profiled using ATAC-seq in islets from two human organ donor samples (6). 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) (11), mapped to hg19 used bwa-mem (version 0.7.15-r1140) (16), 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 (18M reads for islet samples) 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 (17) (https://data.broadinstitute.org/alkesgroup/LDSCORE/baseline_bedfiles.tgz). These included coding, untranslated regions (UTRs), promoter and intronic regions obtained from UCSC (18); the histone marks monomethylation (H3K4me1) and trimethylation (H3K4me3) of histone H3 at lysine 4 and acetylation of histone H3 at lysine 9 (H3K9ac) (10,19,20) and acetylation of histone H3 at lysine 27 (H3K27ac) (21,22); open chromatin, as reflected by DNase I hypersensitivity sites (DHSs) (20,23); combined chromHMM and Segway predictions(24), which partition the genome based on distinct and recurring patterns of histone marks into seven underlying chromatin states; regions that are conserved in mammals(25,26); super-enhancers, which are large clusters of highly active enhancers(21); and enhancers with balanced bidirectional capped transcripts identified using CAGE in the FANTOM5 panel of samples, (called Enhancer (Andersson)) (27). 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 TCs and regulatory annotations was calculated using the Genomic Association Tester (GAT) tool (28). 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.

Aggregate signal

We generated the ATAC-seq density plot over islet TC midpoints using the Agplus tool (version 1.0) (29). 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 (30) 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 (28). 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 (6), which were generated using ATAC-seq data in two islet samples and DNA binding motif information for 1,778 publicly available TF motifs (31–33).

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 1,778 footprint motifs being tested against each TC-related annotation by performing an FDR correction with the Benjajmini-Yekutieli method (34) using the stats.multitest.multipletests function from the statsmodels library in Python (35). 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 (10) 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). 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 MPRA

1. Selection of CAGE elements

We generated a library of islet CAGE elements to test in the MPRA 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 MPRA 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 MPRA experiments were already performed, 6,810 (94.7%) of the CAGE elements included in MPRA 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 MPRA 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 MPRA 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 'MPRA transcripts' (i.e. mRNA that had been transcribed from the MPRA plasmids). The other 1ug of mRNA was used in a enzyme-negative reaction to determine DNA/plasmid contamination in cDNA, which we did not. 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 libraries

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 constructed the 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 MPRA 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 MPRA barcode sequencing data included the input DNA barcode library along with three cDNA barcode libraries representing three biological replicates. 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) (36). 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 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 MPRA activities from these count-based data, we used the tool MPRAnalyze (version 1.3.1) (https://github.com/YosefLab/MPRAnalyze) (37) 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. MPRA 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 MPRA 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 (31–33), which we have also used previously (6,9). In order to reduce motif redundancy, we performed PWM clustering in our motif database using the matrix-clustering tool from RSAT (38), 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 MPRA 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 (39). 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 MPRA 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) (40) 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 (41), islet eQTL (42), and lymphoblastoid cell line (LCL) eQTL (GTEx v7) (43) and organized it in the format required by fGWAS. For eQTL data, we included summary statistics for all tested SNPs for eGenes that passed 1% FDR and used the '-fine' option.

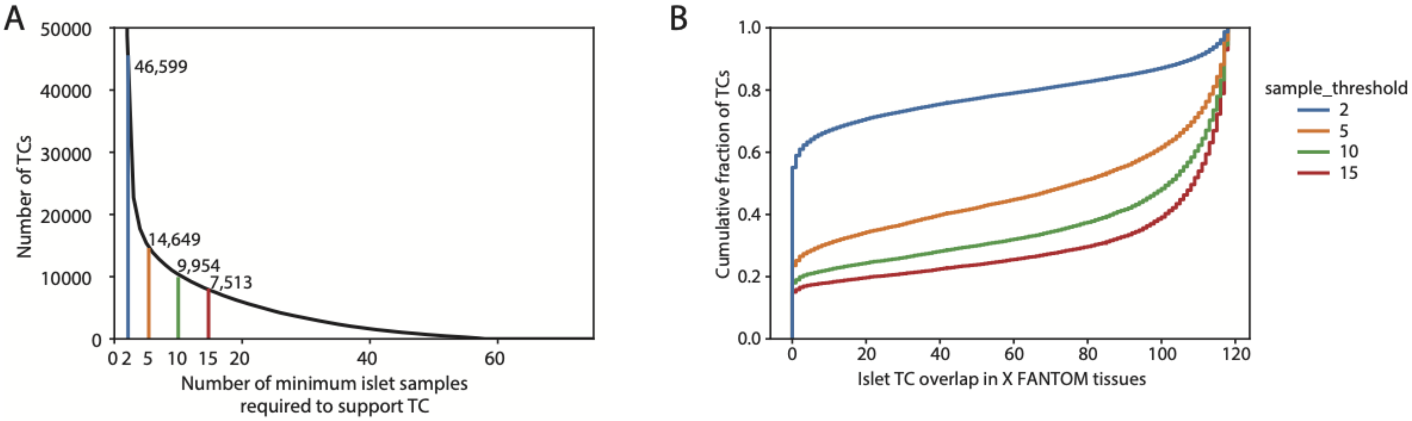
We used fGWAS (default parameters) to calculate enrichment for each annotation and obtained the maximum likelihood enrichment parameter. Annotations were considered as significantly enriched if the log2(parameter estimate) and respective 95% confidence intervals were above zero or significantly depleted 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 addition in fGWAS while including multiple annotations in the model that were individually significantly enriched or depleted (confidence intervals not overlapping 0). We included active TSS, active enhancer, quiescent and polycomb repressed annotations and ATAC-seq peaks with or without or TCs. To compare with T2D genetic credible sets (41), we created 1Mb windows centered on the lead variant at each primary GWAS signal. We then partitioned the rest of the genome into 1Mb windows as well. Specifying these regions using the '-bed' option in fGWAS enabled constraining each primary signal in a single window. The model derived enrichment priors to evaluate both the significance and functional impact of associated variants in GWAS regions, such that variants overlapping more enriched annotations carried extra weight.

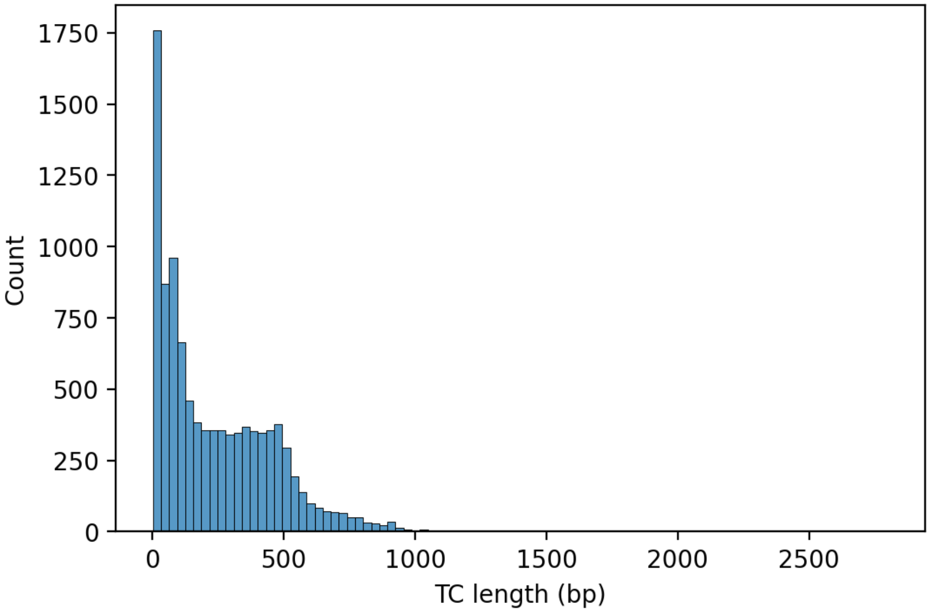
GWAS enrichment using GARFIELD

We tested enrichment of GWAS loci in static and stretch enhancer annotations with GARFIELD (v2) (44). We formatted annotation overlap files as required by the tool; prepared input data at two GWAS thresholds - of 1e-05 and a more stringent 1e-08 by pruning and clumping with default parameters (garfield-prep-chr script). We calculated enrichment in each individual annotation using garfield-test.R with –c option set to 0. We also calculated the effective number of annotations using the garfield-Meff-Padj.R script. We used the effective number of annotations for each trait to obtain Bonferroni corrected significance thresholds for enrichment for each trait.

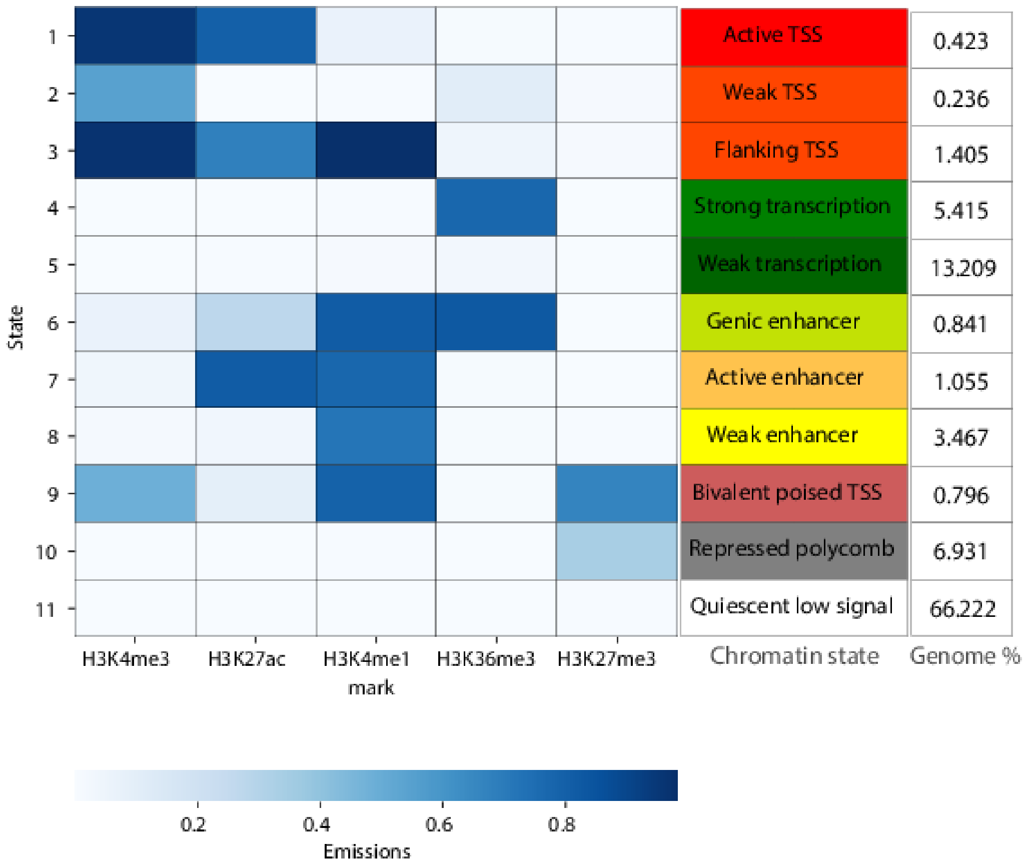
Supplementary Figures



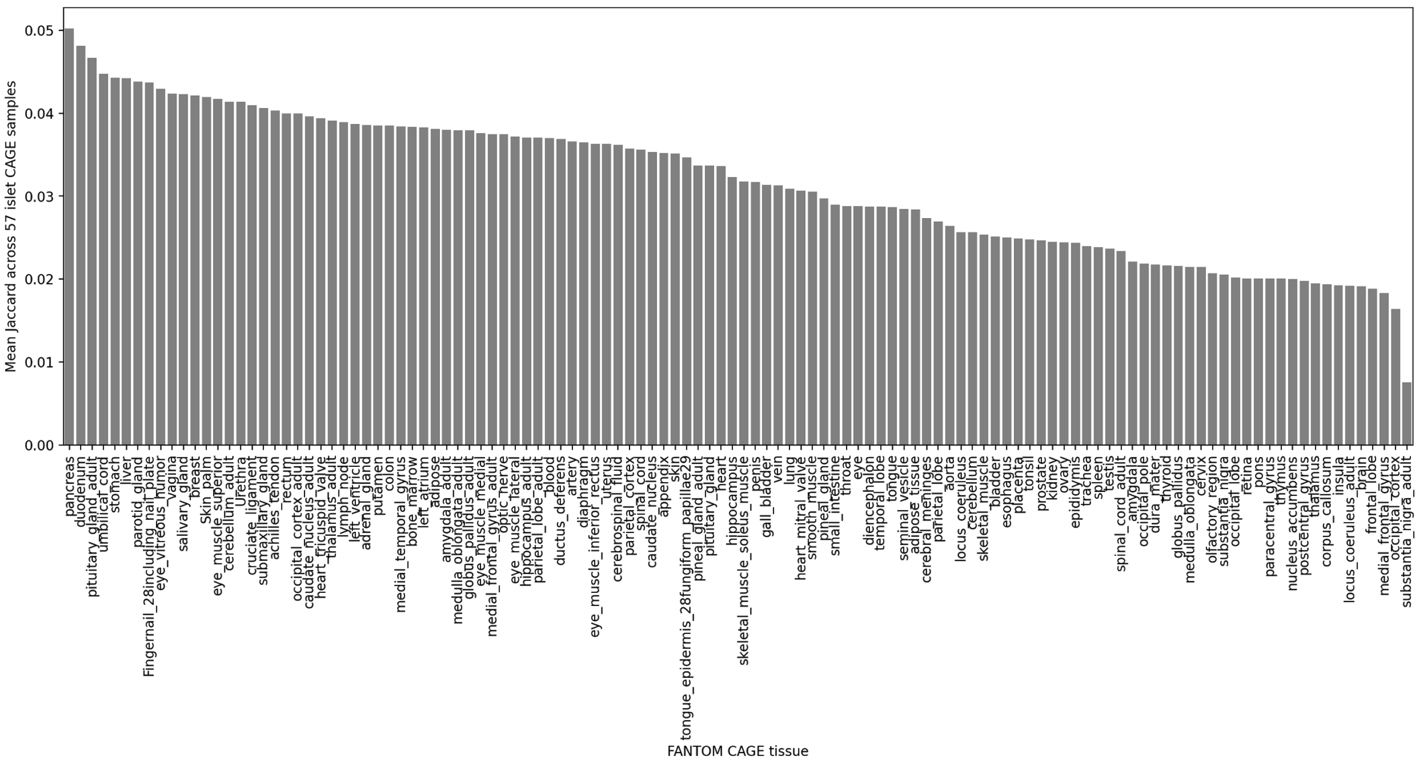
Supplementary Figure 1: Islet TC identification using CAGE data across multiple samples. A: TC segments called using the paraclu method in each of the 57 selected islet samples were merged in a strand specific manner. Shown here are the number of merged segments that are supported by x or more samples. We selected a sample threshold of 10 (dashed line) to define the consensus set of islet TCs (Supplementary table 2A). We also share the merged-TCs at a more lenient threshold of 5 in the Supplementary table 2B. B: Cumulative overlap of islet TC segments with TCs in FANTOM tissues at four example sample-thresholds.



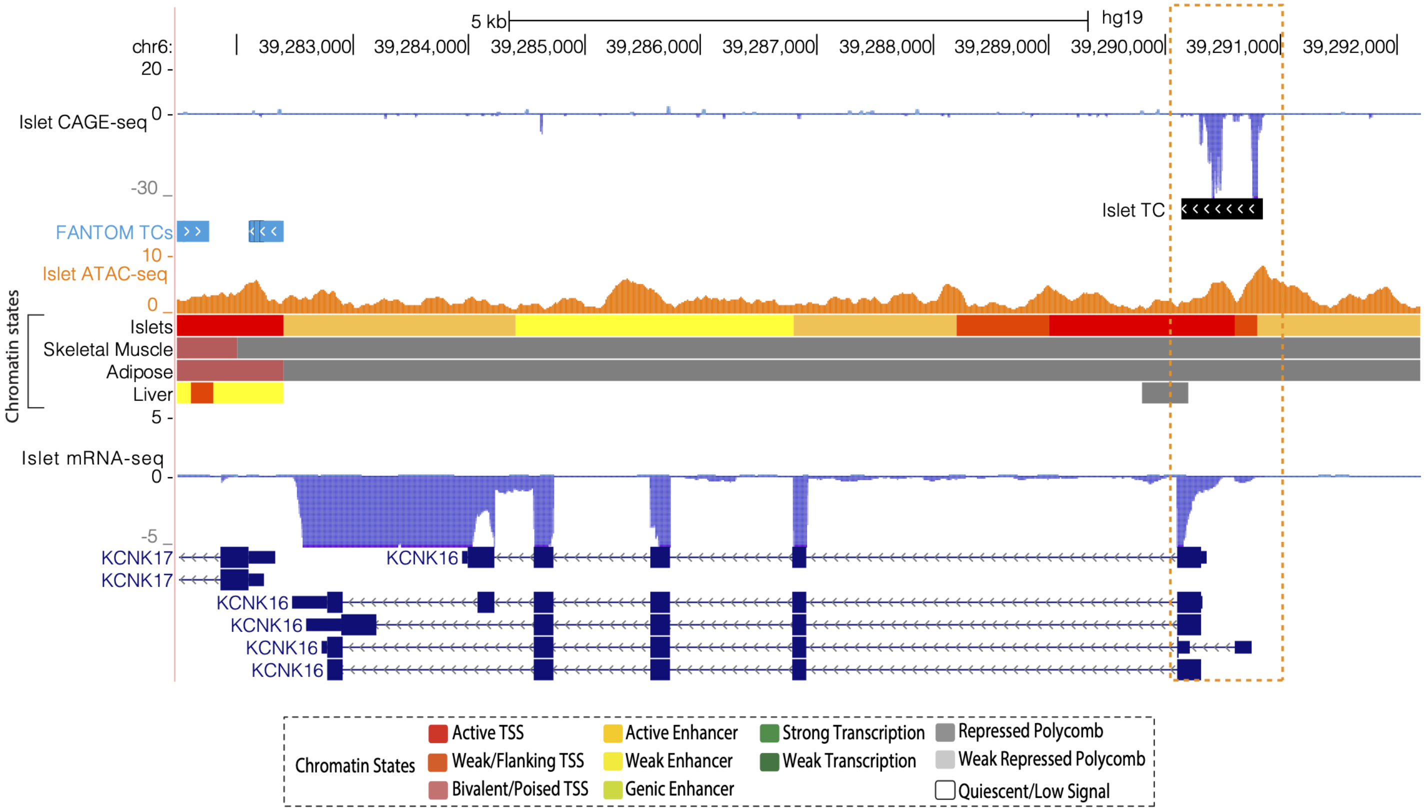
Supplementary Figure 2: Histogram of consensus islet TC segment lengths.



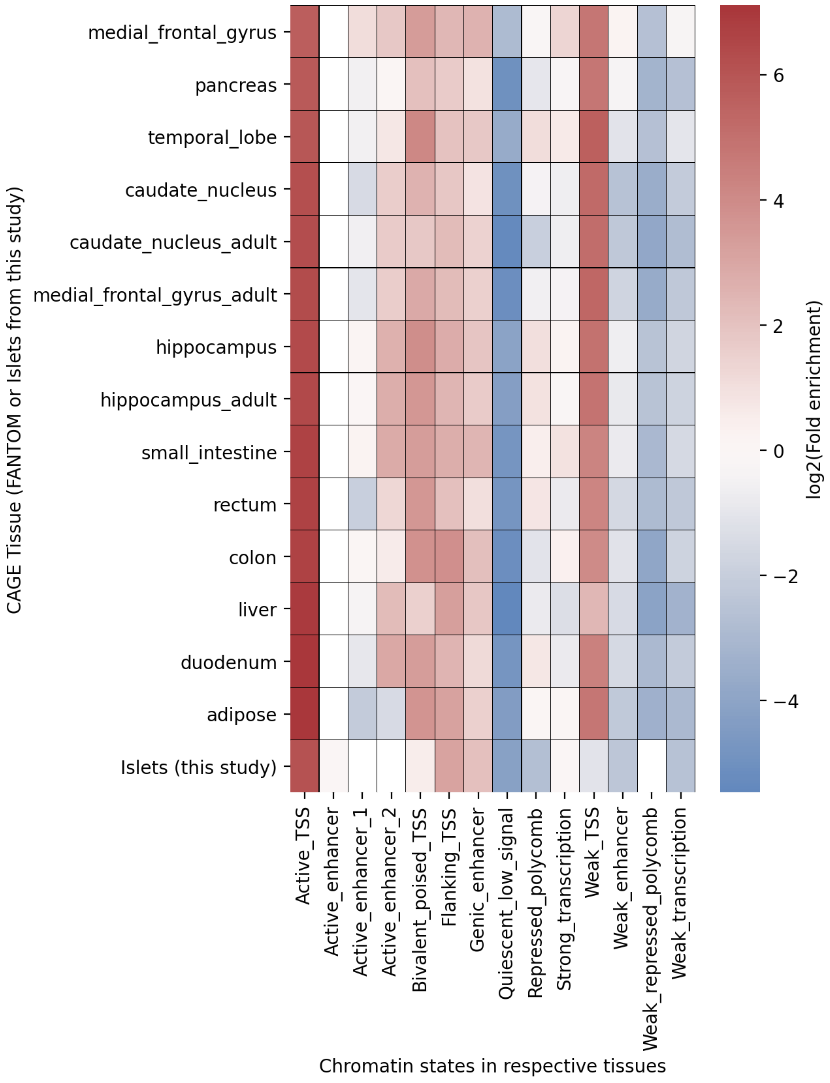
Supplementary Figure 3: 11 chromatin state model generated from publicly available ChIP-seq data (Supplementary table 4) for five histone marks for four cell types (Islets, Skeletal Muscle, Adipose and Liver, see Methods). Shown are the emission probabilities of each of the five histone marks, chromatin state annotation and the percent genome coverage of each state



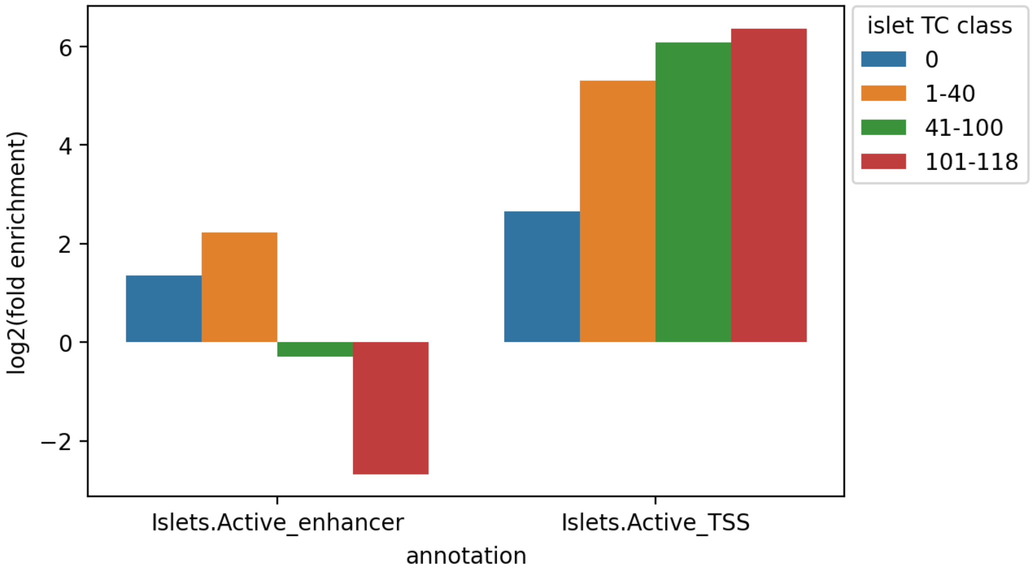
Supplementary Figure 4: Mean Jaccard statistic across 57 islet samples for islet TC overlap with TCs in FANTOM tissues.



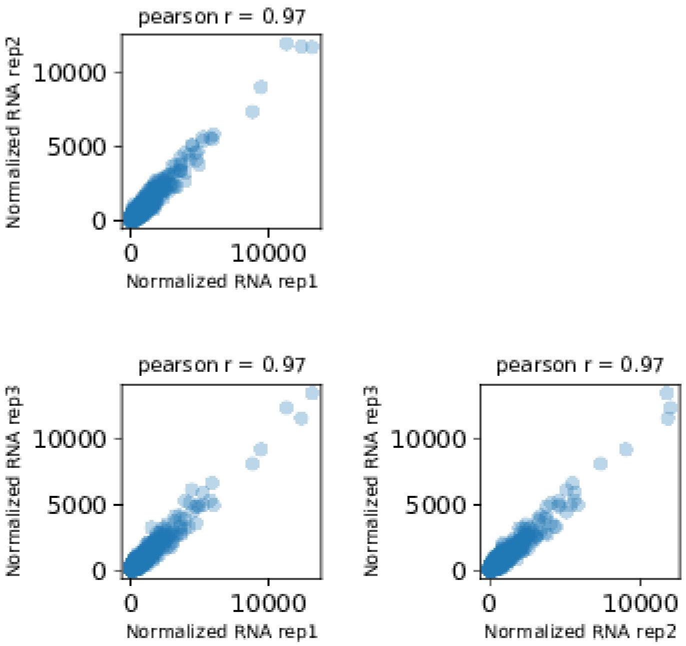
Supplementary Figure 5: *KCNK16* gene locus. Shown are tracks for islet TCs along with TCs called across 118 FANTOM tissues.



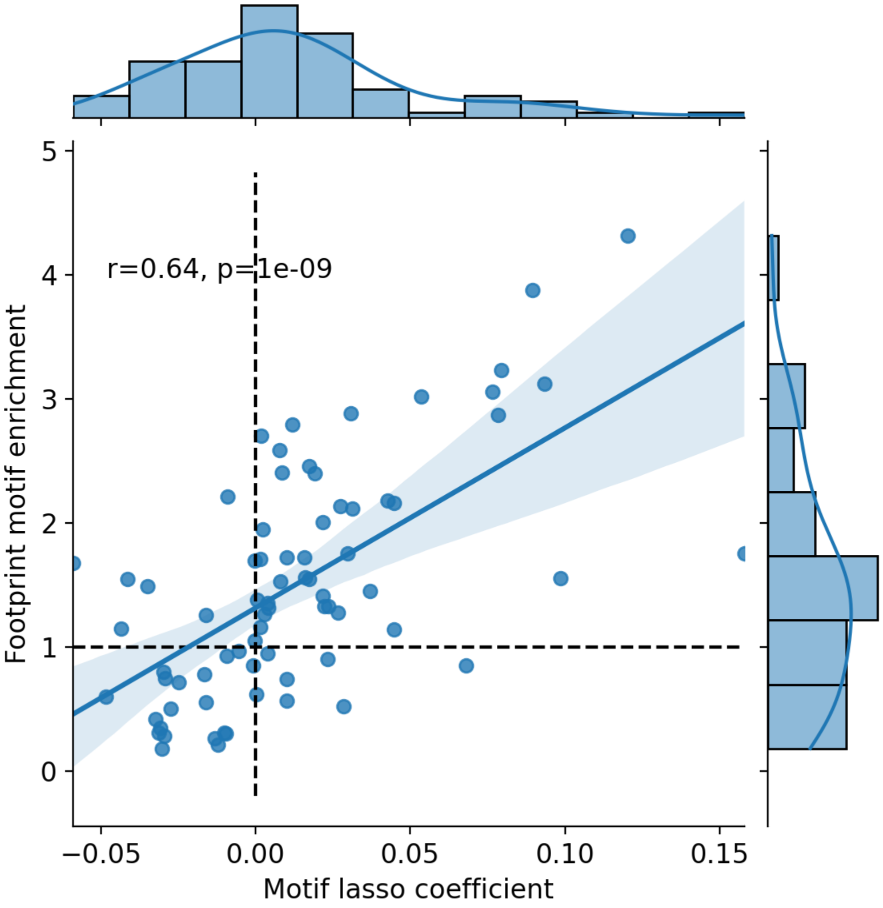
Supplementary Figure 6: Enrichment of TCs identified in various FANTOM tissues that also had publicly available chromatin data (10) to overlap chromatin states in the respective tissue . Also included are islet TCs identified in this study and their enrichment in islet chromatin states.



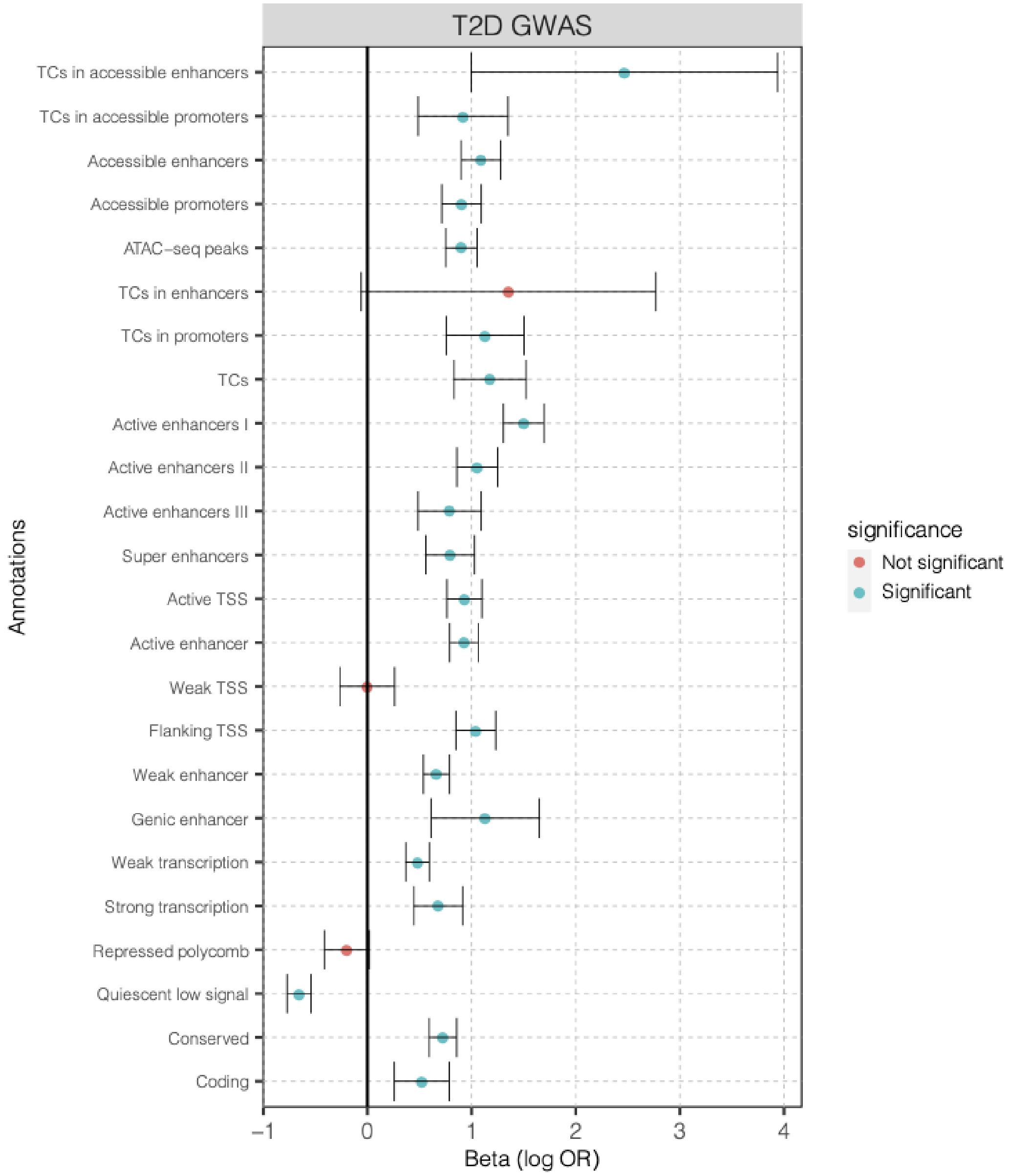
Supplementary Figure 7: Enrichment of islet TCs in the four classes (defined by the number of FANTOM tissues in which the TCs overlap) to overlap islet active enhancer and active TSS chromatin states. The enrichment was computed using GAT.



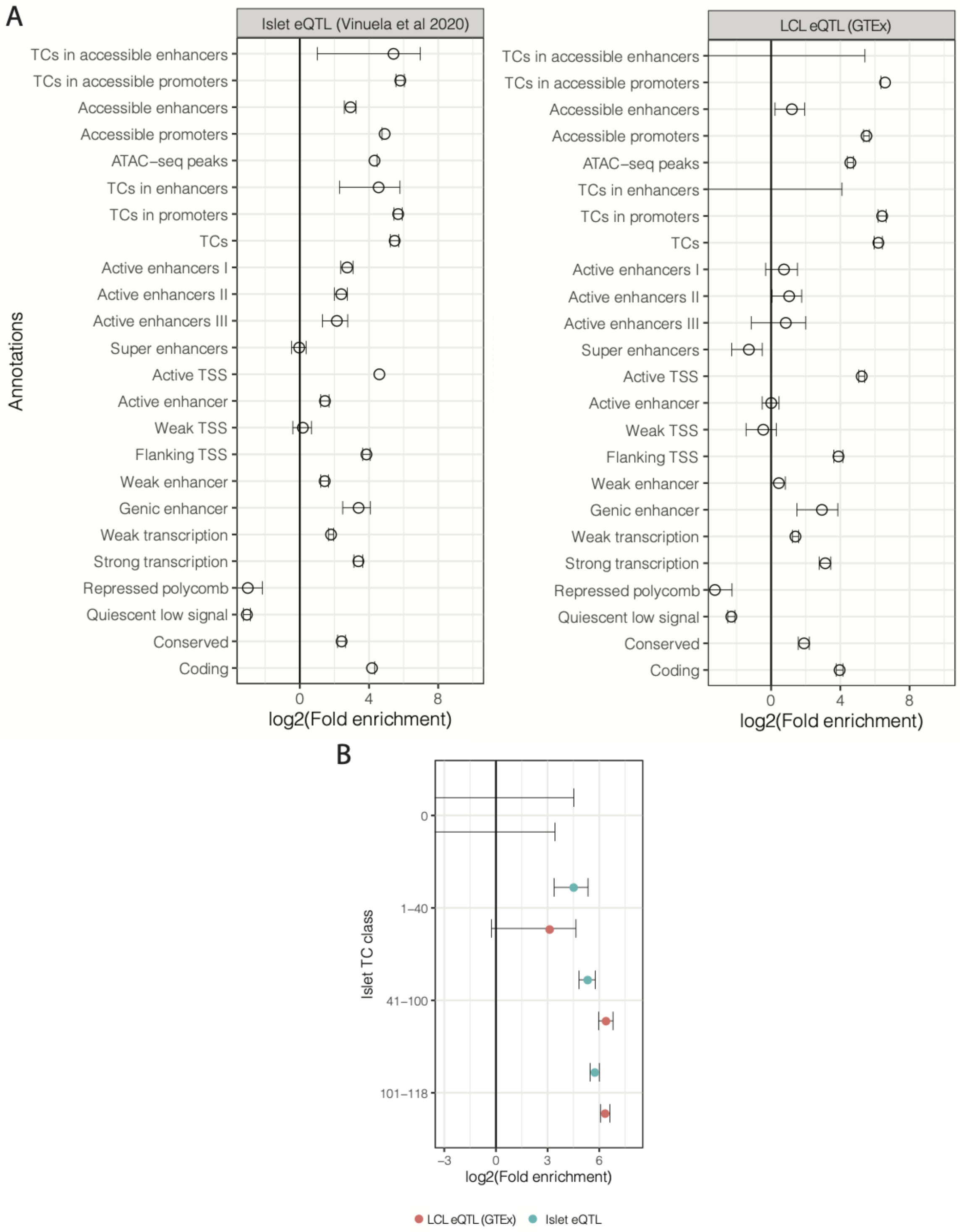
Supplementary Figure 8. Correlation between replicates for normalized total RNA counts for CAGE inserts.



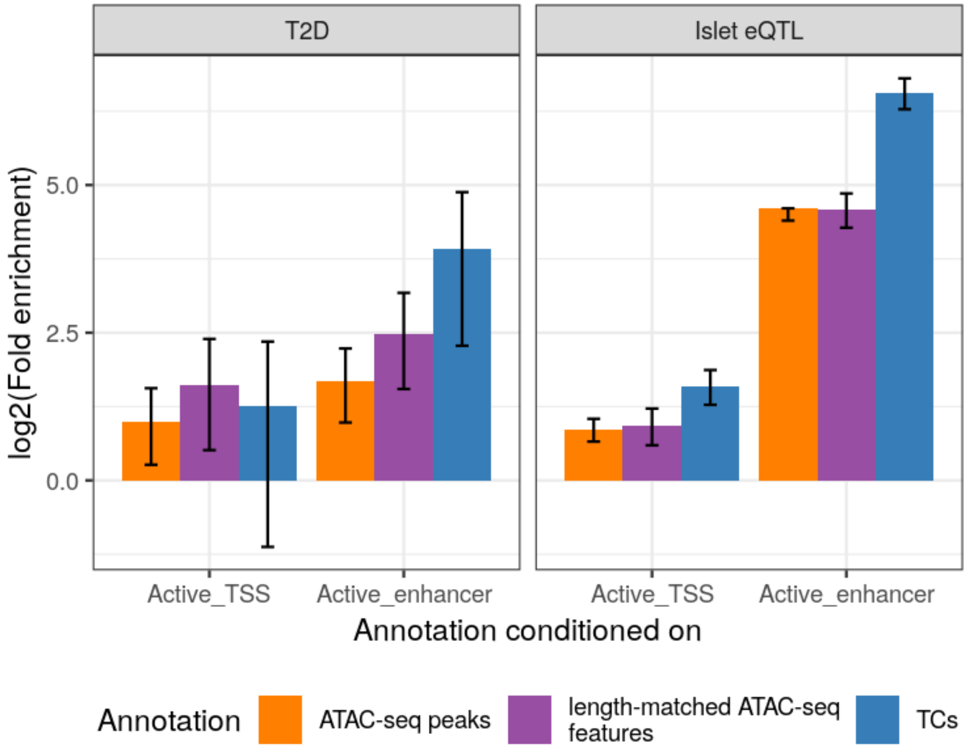
Supplementary Figure 9: TF motif LASSO coefficients vs the fold enrichment for the footprint-motifs to overlap islet TCs



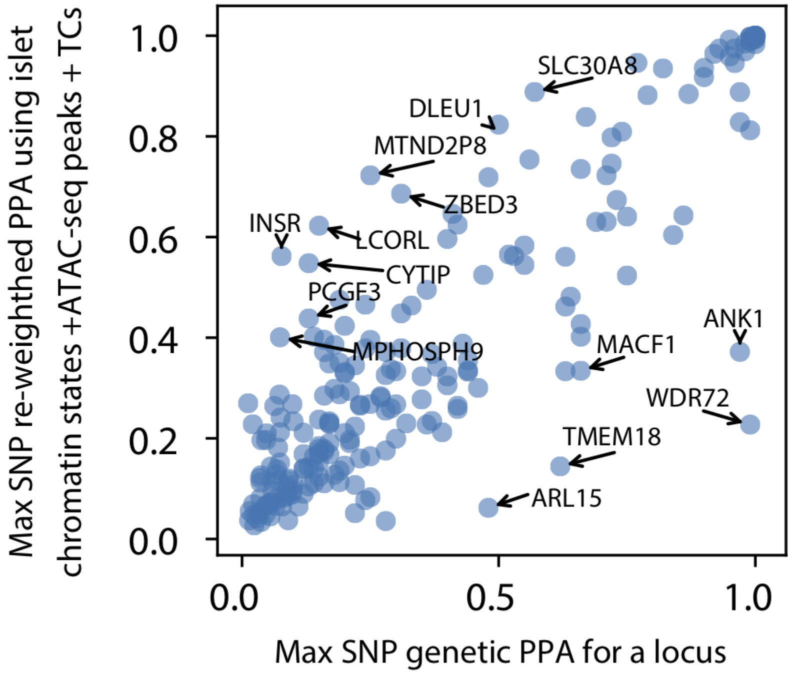
Supplementary Figure 10: Enrichment of T2D GWAS loci in various annotations using GARFIELD (44).



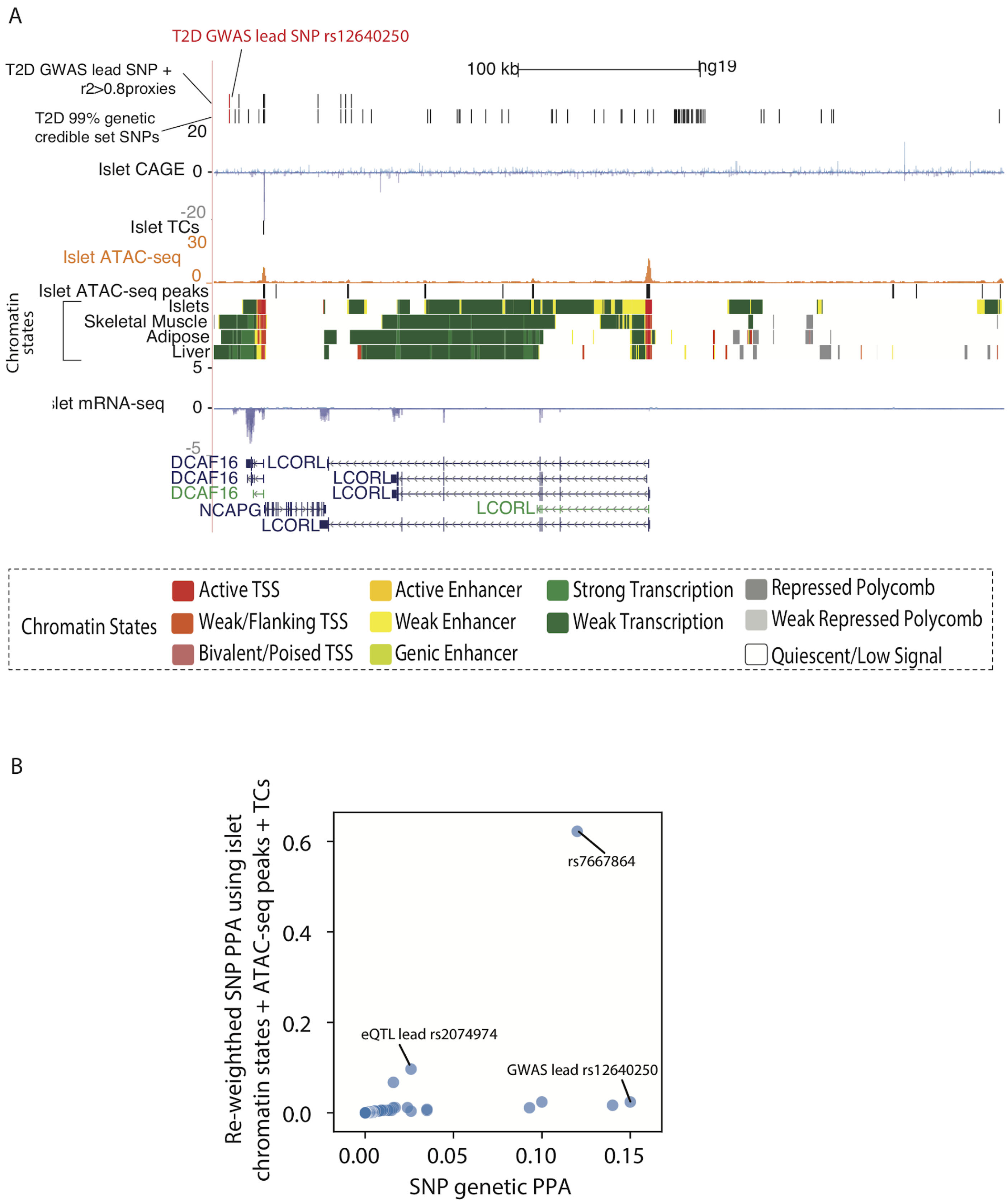
Supplementary Figure 11: eQTL enrichment in TCs. A. Enrichment of islet eQTL compared with LCL eQTL in annotations computed using fGWAS. B: Enrichment of islet or LCL eQTL in TCs in the four specificity classes based on FANTOM tissue CAGE TC overlap.



Supplementary Figure 12: fGWAS conditional enrichment analysis testing the contribution of islet TC (TC) or ATAC-seq peak (full ATAC-seq broadpeak: orange, or length-matched ATAC-seq peak features generated after extending the ATAC-seq peak summit 86 bp on each side: purple) annotations for enrichment to overlap T2D GWAS or islet eQTL loci after conditioning on histone-only based annotations such as active TSS and active enhancer chromatin states (x axis) in islets.



Supplementary Figure 13: The maximal SNP PPA from genetic fine-mapping for a locus vs the maximal SNP PPA after functional re-weighting using islet chromatin states, ATAC-seq peaks and TC annotations.


A
T2D GWAS lead SNP rs12640250
T2D GWAS lead SNP + r2>0.8proxies
T2D 99% genetic credible set SNPs
100 kb
hg19
20
Islet CAGE 0
-20
Islet TCs
30
Islet ATAC-seq
0
Islet ATAC-seq peaks
Chromatin states
Islets
Skeletal Muscle
Adipose
Liver
5
Islet mRNA-seq 0
-5
DCAF16
LCORL
DCAF16
LCORL
DCAF16
LCORL
NCAPG
LCORL
LCORL
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
B
Re-weighted SNP PPA using islet chromatin states + ATAC-seq peaks + TCs
SNP genetic PPA
0.0
0.2
0.4
0.6
0.00
0.05
0.10
0.15
rs7667864
eQTL lead rs2074974
GWAS lead rs12640250

Supplementary figure 14: A: LCORL GWAS locus showing all SNPs in the 99% credible set after genetic fine-mapping. B: Genetic fine-mapping PPA and functionally re-weighted PPA for SNPs at the LCORL locus.

List of supplementary tables:

1. Supplementary_table_1.islet_demographic_data.xlsx - Demographic data for islet samples
2. Supplementary_table_2.tcs.xlsx - Islet consensus TC coordinates (hg19) at two samples thresholds - 10-sample (selected in the manuscript) and a more lenient 5-sample (for reference).
3. supplementary_table_3.gencode.Islet_tcs.xlsx - closest identified islet TCs to known gene TSSs (Gencode V19).
4. Supplementary_table_4_chipseq_references.xlsx - References for histone modification ChIP-seq data used to generate the 11 chromatin state model
5. Supplementary_table_5.results_GAT_formatted.xlsx - Enrichment of islet TCs to overlap various genomic annotations
6. Supplementary_table_6.results_flanking500.xlsx - Enrichment of TF footprint-motifs to overlap 500bp upstream or downstream of TCs
7. Supplementary_table_7.results_states.xlsx - Enrichment of TF footprint-motifs to overlap TCs in accessible enhancers and TCs in accessible promoters.
8. Supplementary_table_8.cage_element_starr-seq.xlsx - quantified MPRA enhancer activities for CAGE elements
9. Supplementary_table_9.lasso.xlsx - Lasso regression coefficients for top 30 TF motifs
10. Supplementary_table_10_fgwas_enrichment.xlsx - Enrichment of islet annotations to overlap T2D GWAS or islet eQTL using fGWAS.
11. Supplementary_table_11_fgwas_reweighting.xlsx - Functionally reweighting T2D GWAS results

References

1. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinforma Oxf Engl. 2013 Jan 1;29(1):15–21.
2. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.
3. Hartley SW, Mullikin JC. QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. BMC Bioinformatics. 2015 Jul 19;16:224.
4. Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A. A code for transcription initiation in mammalian genomes. Genome Res. 2008 Jan 1;18(1):1–12.
5. The FANTOM Consortium, Forrest ARR, Kawaji H, Rehli M, Baillie JK, Hoon MJL de, et al. A promoter-level mammalian expression atlas. Nature. 2014 Mar;507(7493):462.
6. Varshney A, Scott LJ, Welch RP, Erdos MR, Chines PS, Narisu N, et al. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. Proc Natl Acad Sci. 2017 Feb 28;114(9):2301–6.
7. Thurner M, Bunt M van de, Torres JM, Mahajan A, Nylander V, Bennett AJ, et al. Integration of human pancreatic islet genomic data refines regulatory mechanisms at Type 2 Diabetes susceptibility loci. eLife.

2018 Feb 7;7:e31977.

8. Miguel-Escalada I, Bonàs-Guarch S, Cebola I, Ponsa-Cobas J, Mendieta-Esteban J, Atla G, et al. Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. Nat Genet. 2019 Jul;51(7):1137.
9. Scott LJ, Erdos MR, Huyghe JR, Welch RP, Beck AT, Wolford BN, et al. The genetic regulatory signature of type 2 diabetes in human skeletal muscle. Nat Commun. 2016 Jun 29;7:ncomms11764.
10. The Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb;518(7539):317–30.
11. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011 May 2;17(1):10–2.
12. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012 Sep 1;22(9):1813–31.
13. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010 Aug;28(8):817–25.
14. Ernst J, Kellis M. ChromHMM: automating chromatin state discovery and characterization. Nat Methods. 2012 Feb 28;9(3):215–6.
15. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011 May;473(7345):43–9.
16. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio [Internet]. 2013 Mar 16 [cited 2018 Jan 16]; Available from: http://arxiv.org/abs/1303.3997
17. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015 Nov;47(11):1228–35.
18. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996–1006.
19. The ENCODE project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature. 2012 Sep 6;489(7414):57–74.
20. Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet. 2013 Feb;45(2):124–30.
21. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova A a, et al. Super-Enhancers in the Control of Cell Identity and Disease. Cell. 2013 Nov;155(4):934–47.
22. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014 Jul 24;511(7510):421–7.
23. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet. 2014 Nov 6;95(5):535–52.
24. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 2013 Jan;41(2):827–41.
25. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011 Oct 12;478(7370):476–82.
26. Ward LD, Kellis M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science. 2012 Sep 28;337(6102):1675–8.
27. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014 Mar;507(7493):455.
28. Heger A, Webber C, Goodson M, Ponting CP, Lunter G. GAT: a simulation framework for testing the association of genomic intervals. Bioinformatics. 2013 Aug 15;29(16):2046–8.
29. Maehara K, Ohkawa Y. agplus: a rapid and flexible tool for aggregation plots. Bioinformatics. 2015 Sep 15;31(18):3046–7.

30. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinforma Oxf Engl. 2015 Jan 15;31(2):166–9.
31. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-Binding Specificities of Human Transcription Factors. Cell. 2013 Jan 17;152(1):327–39.
32. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. 2014 Mar;42(5):2976–87.
33. Mathelier A, Fornes O, Arenillas DJ, Chen C, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2016 Jan 4;44(D1):D110–5.
34. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001 Aug;29(4):1165–88.
35. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. Proc 9th Python Sci Conf. 2010;5.
36. Zorita E, Cuscó P, Filion GJ. Starcode: sequence clustering based on all-pairs search. Bioinformatics. 2015 Jun 15;31(12):1913–9.
37. Ashuach T, Fischer DS, Kreimer A, Ahituv N, Theis FJ, Yosef N. MPRAnalyze: statistical framework for massively parallel reporter assays. Genome Biol. 2019 Sep 2;20(1):183.
38. Castro-Mondragon JA, Jaeger S, Thieffry D, Thomas-Chollier M, van Helden J. RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. Nucleic Acids Res. 2017 Jul 27;45(13):e119–e119.
39. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011 Apr 1;27(7):1017–8.
40. Pickrell JK. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. Am J Hum Genet. 2014 Apr 3;94(4):559–73.
41. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet. 2018 Nov;50(11):1505.
42. Viñuela A, Varshney A, van de Bunt M, Prasad RB, Asplund O, Bennett A, et al. Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. Nat Commun. 2020 Sep 30;11(1):4912.
43. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017 Oct;550(7675):204.
44. Iotchkova V, Ritchie GRS, Geihs M, Morganella S, Min JL, Walter K, et al. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. Nat Genet. 2019 Jan 28;1.