

A Multi-omic Integrative Scheme Characterizes Tissues of Action at Loci Associated with Type 2 Diabetes

Jason M. Torres,^{1,4} Moustafa Abdalla,^{1,5} Anthony Payne,¹ Juan Fernandez-Tajes,^{1,6} Matthias Thurner,^{1,2} Vibe Nylander,^{2,7} Anna L. Gloyn,^{1,2,3} Anubha Mahajan,^{1,8,*} and Mark I. McCarthy^{1,2,8,*}

Summary

Resolving the molecular processes that mediate genetic risk remains a challenge because most disease-associated variants are non-coding and functional characterization of these signals requires knowledge of the specific tissues and cell-types in which they operate. To address this challenge, we developed a framework for integrating tissue-specific gene expression and epigenomic maps to obtain “tissue-of-action” (TOA) scores for each association signal by systematically partitioning posterior probabilities from Bayesian fine-mapping. We applied this scheme to credible set variants for 380 association signals from a recent GWAS meta-analysis of type 2 diabetes (T2D) in Europeans. The resulting tissue profiles underscored a predominant role for pancreatic islets and, to a lesser extent, adipose and liver, particularly among signals with greater fine-mapping resolution. We incorporated resulting TOA scores into a rule-based classifier and validated the tissue assignments through comparison with data from *cis*-eQTL enrichment, functional fine-mapping, RNA co-expression, and patterns of physiological association. In addition to implicating signals with a single TOA, we found evidence for signals with shared effects in multiple tissues as well as distinct tissue profiles between independent signals within heterogeneous loci. Lastly, we demonstrated that TOA scores can be directly coupled with eQTL colocalization to further resolve effector transcripts at T2D signals. This framework guides mechanistic inference by directing functional validation studies to the most relevant tissues and can gain power as fine-mapping resolution and cell-specific annotations become richer. This method is generalizable to all complex traits with relevant annotation data and is made available as an R package.

Introduction

The scale of genetic studies of type 2 diabetes (T2D [MIM: 125853]) has dramatically expanded in recent years to encompass hundreds of thousands of individuals and tens of millions of variants, culminating in the discovery of over 400 independent genetic associations that influence disease susceptibility.^{1–4} However, as with other complex traits, the majority of T2D-associated variants are non-coding and are presumed to mediate risk by affecting genetic regulatory mechanisms.⁵ Characterization of the processes mediating genetic risk requires definition of the regulatory elements perturbed by these variants, along with the downstream consequences on gene expression and molecular pathways. Such regulatory insights have been typically gleaned through genome-wide approaches that integrate genetic data with information from expression quantitative trait loci (eQTL) analyses, chromatin accessibility and interaction mapping, and functional screening.^{6–11}

A major challenge to these approaches is that the molecular processes that underpin disease risk are often

tissue specific. Although the methods mentioned above can inform a genome-wide view of the tissues most prominently involved in disease (e.g., through patterns of genome-wide enrichment), they do not necessarily identify the most relevant tissue at any given association signal. For example, although several studies have shown strong enrichment of T2D-associated SNPs among regulatory elements in pancreatic islet tissue, there are clearly some signals that exert their impact on disease risk in peripheral tissues such as adipose, skeletal muscle, and liver.^{12–15} Basing functional interpretation on the wrong tissue for a given variant (e.g., relying on islet data for a signal that operates in the liver) is likely to give rise to misleading inference and misdirected efforts at subsequent experimental characterization. Furthermore, as more detailed maps of regulatory elements and functional data in tissues and cell types relevant to disease become available, the need to formulate principled strategies for integrating these features across datasets becomes more important because the ever-expanding scope of epigenomic and transcriptomic reference data can otherwise complicate variant interpretation.

¹The Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, OX3 7BN, UK; ²Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, OX3 7LE, UK; ³Division of Endocrinology, Department of Pediatrics, Stanford School of Medicine, Stanford, CA 94305, USA

⁴Present address: Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, Big Data Institute, University of Oxford, Oxford, OX3 7LF, UK

⁵Present address: Harvard Medical School, Harvard University, Boston, MA 02115, USA

⁶Present address: Genetic and Molecular Epidemiology, Department of Clinical Sciences, Lund University, Malmö, SE

⁷Present address: Danish Medicines Council, 2100 Copenhagen East, DK

⁸Present address: Human Genetics, Genentech, 1 DNA Way, South San Francisco, CA 94080, USA

*Correspondence: anubha@well.ox.ac.uk (A.M.), mark.mccarthy@drf.ox.ac.uk (M.I.M.)

<https://doi.org/10.1016/j.ajhg.2020.10.009>

© 2020 The Authors. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).



To address the challenge of determining most likely “tissues-of-action” at loci associated with complex traits such as T2D, we developed a framework for jointly integrating genetic fine-mapping, gene expression, and epigenome maps across multiple disease-relevant tissues. As an illustration, we show how this scheme enabled a scalable approach for comparing the relative contributions of the key tissues involved in T2D pathogenesis (i.e., those controlling insulin secretion and action) by allowing us to delineate probabilistic tissue scores at individual genetic signals (deemed “tissue-of-action” or TOA scores). We explored the utility of this approach by applying it to a set of fine-mapped genetic associations from a recent large-scale meta-analysis of T2D and assessed the extent to which assigned tissues from a score-based classifier were corroborated by orthogonal datasets. We present results from these analyses along with new insights gleaned from specific loci that show, collectively, that this systematic approach to integrating disparate sources of information effectively resolves relevant tissues at genome-wide association study (GWAS) loci.

Material and Methods

Genetic Data

Genome-wide association summary statistics from a meta-analysis of T2D GWASs corresponding to 32 studies of European ancestry (74,124 affected individuals and 824,006 controls),⁴ conducted by the DIAMANTE consortium, are available on the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) consortium website.

To conduct annotation enrichment analyses (see below), we used GWAS summary statistics from the inverse-variance weighted fixed-effects meta-analysis of T2D unadjusted for BMI that was corrected for residual inflation (accounting for structure between studies) with genomic control.⁴ Of the 403 conditionally independent GWAS signals reported in Mahajan et al., 2018b, 380 signals were amenable to fine-mapping after excluding rare variants (e.g., minor allele frequency (MAF) < 0.25%) and a signal mapping to the major histocompatibility complex (MHC) locus.⁴ Furthermore, 41 of the 403 signals showed heterogeneity in effect estimates between BMI-adjusted and BMI-unadjusted analyses. Fine-mapping of these signals incorporated summary statistics from the appropriate GWAS meta-analysis (i.e., summary statistics from the BMI-unadjusted analysis were used to fine-map signals that were only significant in the BMI-unadjusted analysis).⁴ The 99% genetic credible sets that corresponded to each signal and comprised SNPs that were each assigned a posterior probability of association (PPA)—summarizing the causal evidence for each SNP^{16,17}—were also downloaded from the DIAGRAM website.

Gene Expression Data

Gene expression data for 53 tissues—including liver (n = 175), skeletal muscle (n = 564), and subcutaneous adipose tissue (n = 442)—were downloaded from the Genotype-Tissue Expression Project (GTEx) Portal website. Data correspond to GTEx version 7 (dbGaP accession phs000424.v7.p2) and represent RNA sequencing reads mapped to GENCODE (v19) genes.¹⁸

Gene expression data for pancreatic islets (n = 114) were accessed from a previous study⁵ that involved sequencing stranded

and unstranded RNA library preparations at the Oxford Genomics Centre. We used this set of islet samples to calculate expression specificity scores and perform co-expression analysis (see below and in [Gene Co-expression](#)). An additional set of 60 islet samples available to us in-house was also used for eQTL mapping and enrichment analysis. All 174 islet samples were included in a subsequent analysis¹⁹ performed by the Integrated Network for Systemic Analysis of Pancreatic Islet RNA Expression (InsPIRE) consortium. RNA-sequencing reads of all islet samples were also mapped to gene annotations in GENCODE (v19), in line with GTEx accessed data, with Spliced Transcripts Alignment to a Reference (STAR; v020201) and quantified with featureCounts (v1.50.0-p2).

Gene read counts for each tissue were transcript per million (TPM) normalized to correct for differences in gene length and library depth across samples. The tissue specificity of TPM-normalized gene expression was measured with expression specificity scores (ESSs) obtained with the following formula:

$$\epsilon_{g,t} = \frac{\text{med}(\text{expression}_{g,t})}{\sum_{x \in T} \text{med}(\text{expression}_{g,x})}$$

where $\epsilon_{g,t}$ is the ESS for gene g in tissue t , and T is the set of evaluated tissues.

Partitioning Chromatin States

Chromatin state maps from a previous study²⁰ based on a 13-state ChromHMM²¹ model trained from ChIP-seq input for histone modifications (H3K27ac, H3K27me3, H3K36me3, H3K4me1, and H3K4me3) were downloaded from the Parker lab website. Chromatin state maps for liver, pancreatic islet, skeletal muscle, and adipose tissue (nuclei from crude preps of abdominal fat deposits²²) were used for the present study. Partitioned chromatin state maps used for generating TOA scores (see [Deriving TOA Scores](#)) were obtained in the R statistical environment (v3.6.0) with the Genomic Ranges (v1.36.1) library. For each chromatin state annotation, we used the disjoint function (Genomic Ranges) to delineate non-overlapping segments across each of the four tissues. These segments were then compared with the annotation sets corresponding to each tissue to determine segments that were (1) tissue specific, (2) shared across all tissues, or (3) shared in a combination of two or more (but not all) tissues.

Annotation Enrichment Analysis

To obtain fold enrichment values to use as annotation weights, we performed genome-wide enrichment analysis by using the program fgwas²³ (v0.3.6). For this analysis, we used summary statistics (i.e., Z scores, p values) from the DIAMANTE European BMI-unadjusted meta-analysis of T2D GWASs.⁴ Enrichment of T2D-associated SNPs was assessed for coding sequence (CDS) and 13 chromatin state annotations mapped in human islet, liver, skeletal muscle, and adipose tissue from the Varshney et al. study.²⁰ To estimate log₂-fold enrichment values, we used the `-cc` flag (specifying GWAS input from a case-control study) and applied default distance parameters (i.e., genome partitioned “blocks” of 5,000 SNPs). Weights were obtained by exponentiating the mean log₂-fold enrichment values for each tissue-level annotation.

Deriving TOA Scores

In order to obtain TOA scores for each of the 380 conditionally independent genetic association signals, we partitioned the corresponding PPA values of the 99% genetic credible set SNPs. For

each SNP j in the 99% credible set, we obtain a vector $s_{j,a}$ for each annotation a among the set of coding sequence and chromatin state annotations in set A . Each element in $s_{j,a}$ corresponds to a tissue t in the set T comprising all evaluated tissues and is given by the equation,

$$s_{j,a,t} = \frac{w_{a,t}}{\sum_{i \in T} \mathbb{1}(j, a, i)} \mathbb{1}(j, a, t),$$

where $w_{a,t}$ is the weight of annotation a in tissue t and $\mathbb{1}$ is a SNP-mapping function defined as

$$\mathbb{1}(j, a, t) := \begin{cases} 1 & \text{if SNP } j \text{ overlaps chromatin state annotation} \\ & a \text{ in tissue } t \\ \varepsilon_{g,t} & \text{if SNP } j \text{ overlaps coding sequence annotation,} \\ & a \text{ for gene } g \\ 0 & \text{otherwise} \end{cases},$$

where $\varepsilon_{g,t}$ is the ESS value for gene g in tissue t . Note that this function serves as an indicator function for binary annotations (e.g., chromatin states), whereas in the special case of coding SNPs, continuous values on the interval $[0,1]$ (i.e., ESS values) are used to indicate the relative expression levels of the corresponding gene and can be used to inform tissue-level relevance for each coding SNP. If the SNP j does not map to annotation a in any tissue $t \in T$, the value of $s_{j,a,t}$ is set to 0. The vector s_j is thus given by the following equation:

$$s_j = \sum_{a \in A} s_{j,a},$$

where the elements in s_j correspond to each tissue $t \in T$ and can be interpreted as tissue-specific annotation weights obtained from a linear combination of partitioned genome-wide fold enrichment values for each tissue-level annotation. The vector τ_c that comprises TOA scores for each tissue $t \in T$ and corresponds to 99% genetic credible set c is given by the following equation:

$$\tau_c = \sum_{j \in J} \frac{P_j s_j}{\sum_{t \in T} S_{j,t}},$$

where J is the set of SNPs in the 99% genetic credible set c . Lastly, an unclassified score U_c is defined for each 99% genetic credible set c :

$$U_c = \sum_{j \in J} P_j - \sum_{i=1}^n \tau_{c,i},$$

where $\tau_{c,i}$ is the TOA score of tissue i for credible set c and n is the number of evaluated tissues. This term indicates the cumulative PPA in c that is attributable to credible SNPs that do not map to any of the evaluated tissue-level annotations.

To evaluate the robustness of TOA score-based estimates of overall tissue contributions to T2D risk against the effect of GWAS association strength, we constructed weighted TOA scores:

$$\omega_c = \tau_c \frac{|\beta|}{SE},$$

where β and SE are the effect size and standard error for the conditionally independent SNP upon which the 99% credible set c was mapped.

Profiling Tissue Specificity

The sum of squared distances (SSD) between TOA scores in τ_c for each $c \in C$ (where C is the set of 99% genetic credible sets) was

used as a measure of tissue specificity. To gauge the relationship between fine-mapping resolution and tissue specificity, we used univariate linear models to estimate β coefficients corresponding to the regression of the SSD on either the maximum 99% genetic credible set PPA or the \log_{10} number of SNPs in the 99% genetic credible sets. Signals were designated as “shared” if the difference between the top two TOA scores was ≤ 0.10 . “Shared” signals were then tiered on the basis of fine-mapping resolution: (1) signals corresponded to 99% genetic credible sets comprised of a single credible SNP; (2) signals corresponded to 99% genetic credible sets where the maximum PPA ≥ 0.50 (i.e., where a single SNP explained most of the cumulative PPA); or (3) signals corresponded to 99% genetic credible sets where the maximum PPA < 0.50 . The relationship between the SSD and fine-mapping resolution (i.e., maximum credible set PPA and number of credible SNPs) was visualized with the scatterpie library (v0.1.4) in the R statistical environment (v3.6.0).

Rule-Based Classifier

A rule-based classifier for assigning each genetic signal (i.e., 99% genetic credible set) to a tissue was derived by assigning each genetic signal c to a tissue t if the corresponding TOA score in τ_c had the maximum value and exceeded a specified threshold. Sets of tissue-assigned signals were constructed for each stringency threshold within the set (0.0, 0.2, 0.5, 0.8). The classifier also allowed for a “shared” designation with the criteria described in the previous section (i.e., difference between the top two TOA scores was ≤ 0.10).

eQTL Mapping and Tissue-Specific eQTL Enrichment

eQTLs for human liver ($n = 153$), skeletal muscle ($n = 491$), and subcutaneous adipose tissue ($n = 385$) were accessed from the GTEx Portal website and corresponded to GTEx version 7 (dbGaP accession phs000424.v7.p2). For human islet tissue, we used 174 samples (described in [Gene Expression Data](#)) and performed eQTL mapping by using FastQTL (v2.0) with a nominal pass with the `-normal` flag (to fit TPM-normalized read counts to a normal distribution). Gender and the first 15 PEER factors²⁴ were used as covariates. For each tissue, q values were calculated from nominal p values and a false discovery rate threshold of ≤ 0.05 was applied to identify significant eQTLs.

To obtain sets of tissue-specific eQTLs, we first took the union of all eQTLs for tissues in set T , given by the following equation:

$$M = \bigcup_{t \in T} S_t,$$

where S_t is the set of eQTLs in tissue t . We defined the set of tissue-specific eQTLs for each tissue as the list of significant eQTLs that were significant in only that tissue.

Enrichment analysis was performed by taking the set of signals assigned to each tissue $t \in T$ at each stringency threshold. Each tissue-assigned signal (i.e., 99% genetic credible set) was then mapped to the corresponding GWAS index SNP reported in Mahajan et al., 2018b,⁴ yielding a set of index SNPs for each tissue t .

For each tissue t , fold enrichments were estimated by taking the observed number of tissue-specific eQTLs among the set of tissue-assigned signals for tissue t divided by the mean number of overlapping signals across the 1,000 permuted sets of matched SNPs corresponding to the set of signals (i.e., mapped index SNPs)

assigned to tissue t . Empirical p values were calculated by the following equation:

$$p_{emp} = \frac{n_{null \geq obs} + 1}{N + 1},$$

where $n_{null \geq obs}$ is the number of instances where the number of overlapping tissue-specific eQTLs among a null set of matched SNPs was greater than or equal to the number observed among the set of tissue-assigned signals and N is the total number of permutations.

Functional Fine-Mapping

A set of comparative *functional* fine-mapping analyses were performed with the program fgwas (v0.3.6) and the summary statistics from the GWAS meta-analysis for T2D unadjusted for BMI⁴ and three annotation schemes: (1) “null” analysis without any genomic annotations; (2) “multi-tissue” combined analysis using 13-state chromatin state maps for islet, liver, skeletal muscle, and adipose tissue from Varshney et al.²⁰ (described in [Partitioning Chromatin States](#)); and (3) “deep islet” analysis based on 15-state chromatin segmentation maps for human islet from Thurner et al.,²⁵ and notably, these states were based on a richer set of input features assayed in islets that included ATAC-seq and whole-genome bisulfite sequencing, in addition to histone ChIP-seq.

For both the multi-tissue and deep islet analyses, we used fgwas to obtain a “full model” by first seeding a model with the single annotation that yielded the greatest model likelihood in a single annotation analysis. This model was extended by iteratively adding annotations—in descending order on the basis of their model likelihoods—until the incorporation of additional annotations no longer increased the model likelihood of the joint model. The “full” model resulting from this procedure was then reduced by iteratively dropping annotations that yielded an increased cross-validated likelihood upon their exclusion from the joint model. The “best joint model” was obtained when this process no longer improved the cross-validated likelihood. The annotations remaining in the “best joint model” were then carried forward for functional fine-mapping.

In the next step, a locus-partitioned analysis was performed with the set of annotations from the “best joint model” for the multi-tissue and deep islet analyses or no annotations for the null analysis. The default behavior of fgwas involves partitioning the genome into “blocks” of 5,000 SNPs and assuming no more than one causal variant per block. To account for allelic heterogeneity at loci with conditionally independent signals and to facilitate a comparison with the 99% genetic credible sets (which were constructed with conditionally deconvoluted credible sets), the genome was partitioned into 1 Mb windows centered about each index variant (specified using the `-bed` command) and fgwas was run with the appropriate set of input annotations for each of the three analytic schemes. Windows involving multiple independent signals required separate fgwas runs, each corresponding to the appropriate set of approximate conditioned summary statistics (i.e., conditioning on the effect of one or more additional signals at a locus).⁴ The resulting PPA values for each SNP in each partitioned “block” was used to construct 99% functional credible sets by ranking SNP by PPA in descending order and retaining those that yielded a cumulative PPA ≥ 0.99 .

To compare the differences in fine-mapping resolution between the multi-tissue and deep islet schemes, at each signal, we obtained the difference between maximum 99% functional credible

set PPA for each scheme with that resulting from the null analysis as a baseline. These differentials over the null were then compared between the multi-tissue and deep islet schemes, and significance was assessed with the Wilcoxon rank-sum test. Comparative tests were performed for each set of tissue-assigned signals across the four stringency thresholds.

Gene Co-expression

Genes with TPM counts < 0.1 in $>50\%$ of samples per tissue were excluded, and the remaining genes were ranked on the basis of their mean expression across all tissues. For each set of tissue-assigned genetic signals, at each specified classifier threshold, a set of genes was determined on the basis of nearest proximity to the index SNP for each signal. Signals that corresponded to 99% genetic credible sets where coding variants accounted for a cumulative PPA ≥ 0.1 were excluded from the analysis. A “background” set of genes was then obtained by including all genes with rank values ± 150 about the rank values of each gene in the filtered set. Null sets of genes were then delineated by sampling genes from the background set that had rank values within 100 of those for each gene in the gene set. We repeated this last step to generate 1,000 sets of null genes. To assess expression similarity in each of the 54 tissues, the rank sum of the genes in the set was recorded and compared with the mean rank sum across the 1,000 sets of null genes separately for each tissue. An empirical p value was determined with the following equation:

$$p_{emp} = \frac{n_{null \leq obs} + 1}{N + 1},$$

where $n_{null \leq obs}$ is the number of instances when the rank sum of genes in a null set was less than or equal to the observed rank sum in a given tissue and N is the number of permutations. To gauge the magnitude of similarity of gene expression levels, an enrichment factor was defined by taking the mean rank sum across the null sets divided by the observed rank sum. This procedure was repeated for sets of the second and third nearest genes to each index SNP corresponding to tissue-assigned signals across classifier thresholds.

Gene co-expression was assessed through a correlation-based test wherein, for each set of proximal genes corresponding to tissue-assigned signals described above, pairwise Spearman correlations of gene expression in each of the four T2D-relevant tissues were calculated. The observed mean squared ρ (msr) for each set was compared against a null distribution ascertained from 10,000 random samples of proximal genes with respect to the 380 T2D signals.

Physiological Cluster Enrichment

A set of T2D-associated SNPs that were clustered into physiology groups were obtained from a recent study.³ As previously described, summary statistics (Z scores) for a range T2D-relevant metabolic traits (e.g., anthropometric, lipid, and glycemic) were used to cluster 94 coding and non-coding SNPs associated with T2D via “fuzzy” C-means clustering of Euclidean measures.³ An additional, and partially overlapping, set of 94 T2D-associated SNPs was also accessed and was previously clustered into physiology groups via an input set of sample size-adjusted Z scores corresponding to 47 T2D-related traits and nonnegative matrix factorization (bNMF) clustering.²⁶ Because not all of the physiologically clustered SNPs were present among the set of index SNPs corresponding to the 380 fine-mapped genetic association

signals, pairwise linkage disequilibrium (LD) was measured between all SNPs in these sets with the LDproxy tool on the LD Link website and all European populations from the 1000 Genomes Project (phase 3) as a reference. Physiologically clustered SNPs were assigned to fine-mapping index SNPs on the basis of maximum pairwise LD where $r^2 > 0.3$. From this approach, 82/94 SNPs and 63/94 SNPs from the two sets of physiologically clustered signals (from Mahajan et al., 2018a³ and Udler et al.,²⁶ respectively) were mapped to fine-mapped signals in Mahajan et al., 2018b.⁴ For each set of tissue-assigned signals with n signals, assigned at each classifier threshold, null SNP sets were generated by randomly sampling n signals from the set of 380 fine-mapped signals. A null distribution was obtained by generating 10,000 null sets and recording the overlap of null signals with each of the physiologically-clustered signals. An empirical p value was obtained with the following equation:

$$p_{emp} = \frac{n_{null \geq obs} + 1}{N + 1},$$

where $n_{null \geq obs}$ is the number of instances where the observed overlap between a null set and a reference set of physiologically assigned signals was greater than or equal to the observed value for the query set of tissue-assigned signals and N is the total number of null sets (i.e., 10,000). An enrichment factor was obtained by taking the observed overlap divided by the mean of the null overlap values.

Enrichment for Trait-Associated SNPs from GWASs

GWAS summary statistics for all available traits and diseases were downloaded from the NHGRI-EBI GWAS catalog (v1.0; accessed August 23, 2019). Coordinates for all trait-associated SNPs in the catalog were mapped to genome build GRCh38. GRCh38 coordinates for index SNPs corresponding to each of the 99% genetic credible sets were obtained from the Ensembl website by querying with reference SNP ID number. We determined proxy SNPs for each SNP in the set of index SNPs corresponding to the 99% genetic credible sets by using the `–show-tags` function in PLINK (v1.90b3) to identify SNP proxies with LD $r^2 \geq 0.8$ among a reference panel of European individuals from the 1000 Genomes Project (phase 3). VCF files for SNPs from the 1000 Genomes Project mapped to genome build GRCh38 were downloaded from the project website. For each set of tissue-assigned signals, enrichment was assessed across each of the 3,616 diseases or traits in the GWAS catalog. The observed number of SNPs overlapping the set of index and proxy SNPs corresponding to the tissue-assigned signals and the set of trait-associated SNPs for a given GWAS was recorded. To obviate bias due to local LD, multiple SNPs (i.e., index and proxies) corresponding to a single signal that was shared with the set of GWAS SNPs were recorded as a single overlap for that signal. A null distribution of SNP overlaps was obtained through 10,000 rounds of random sampling from the set of index SNPs corresponding to each of the 380 fine-mapped credible sets. An empirical p value was obtained with the following formula:

$$p_{emp} = \frac{n_{null \geq obs} + 1}{N + 1},$$

where $n_{null \geq obs}$ is the number of instances where the number of SNP overlaps between a null and GWAS SNP set exceeded the observed overlap for the set of tissue-assigned signals. The magnitude of enrichment was measured by the number of

observed overlaps divided by the mean of the overlaps across the null sets.

Results

An Integrative Approach for Obtaining TOA Scores at Trait-Associated Loci

We set out to quantify, in the form of TOA scores, the contribution of disease-relevant tissues to each genetic association signal from a recent GWAS meta-analysis of T2D⁴ by integrating genetic, genomic, and transcriptomic data. To do this, we developed a scheme that derived, for each GWAS signal, a measure of overlap with tissue-specific regulatory annotations. We then combined these by using weights obtained from both genetic fine-mapping and genome-wide enrichment of tissue- and annotation-specific annotations (Figure 1). The rationale for this approach acknowledged that our confidence in the identity of causal SNPs varies considerably across GWAS association signals, and evidence supporting causality for each candidate SNP should be explicitly taken into consideration. Moreover, as we aimed to compare evidence supporting the involvement of candidate tissues at each genetic signal, we focused our analysis on sets of epigenomic annotations of a common data type that could be systematically referenced when profiling trait-associated SNPs.

To obtain tissue scores at each genetic signal, we first delineated a set of annotation vectors on the basis of the physical position of each SNP in the corresponding 99% genetic credible set (from Bayesian fine-mapping) with respect to the panel of tissue-specific chromatin states (Figure 1). We used chromatin states from a recent study²⁰ to form a reference set of epigenomic annotations focusing on tissues involved in insulin secretion (pancreatic islets) and insulin-response (skeletal muscle, adipose, and liver) that play central roles in the pathophysiology of T2D. There is support for the role of these tissues from patterns of overall genome-wide enrichment of tissue-specific regulatory features and from the known effects at the subset of T2D association signals for which causal mechanisms have been established.^{14,15,20,25,27} For non-coding SNPs, binary values were used to encode genome mapping (i.e., whether or not a SNP maps to a regulatory region in a given tissue as shown in step 1A in Figure 1). For the minority of credible set SNPs that map to coding sequence, quantification focused on measures of tissue-specific RNA expression for the genes concerned to further inform the relative importance of the evaluated tissues (see Methods) (Figure 1, step 1B).

Next, we combined and scaled the annotation vectors to yield a vector of tissue scores that were used to partition the PPA of each credible SNP (Figure 1, step 2). To facilitate this partitioning and to account for the relative importance of relevant tissues with respect to overall T2D pathogenesis, we first estimated genome-wide enrichment of T2D-associated SNPs across a set of tissue-specific genomic

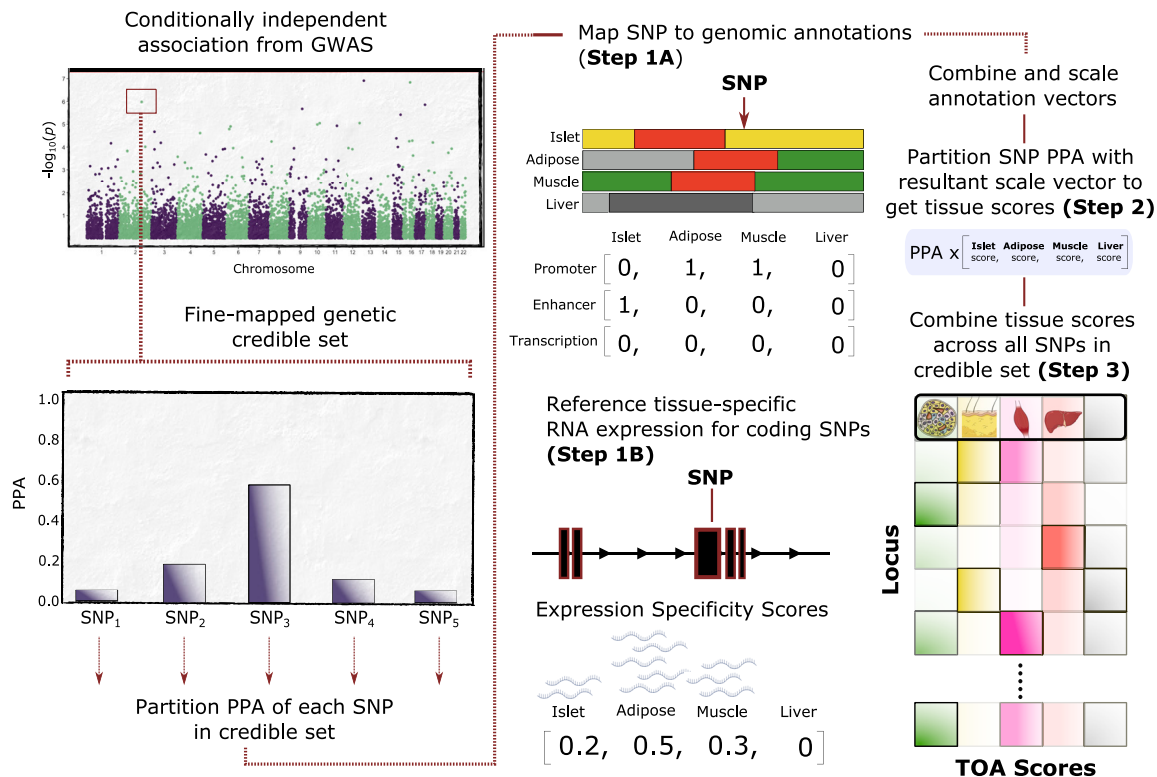


Figure 1. Systematic Approach for Obtaining Tissue-of-Action Scores

Fine-mapping of conditionally-independent GWAS signals results in a set of credible variants, each with a posterior probability of association (PPA). The illustrated example shows a signal with five SNPs in its credible set with SNP₃ as the variant with the maximum PPA. Each credible SNP is then mapped to a panel of chromatin state annotations across four disease-relevant tissues to obtain a set of annotation vectors (step 1A). An additional annotation vector for SNPs mapping to coding sequence (CDS) is obtained from expression specificity scores (ESSs) calculated from gene expression levels across the four tissues (step 1B). The set of annotation vectors for each SNP are then summed and scaled, yielding a vector used to partition the PPA value (step 2). The resultant vectors for each SNP in a genetic credible set are then summed and scaled to yield a tissue-of-action (TOA) score for each tissue at the GWAS signal corresponding to the credible set (step 3). Any residual PPA values from SNPs not mapping to any of the evaluated tissue annotations are allocated to an “unclassified” score (gray column in matrix).

annotations. We used the enrichment values as weights to adjust the relative tissue contributions of SNPs mapping to distinct functional annotations or to functional annotations shared in more than one tissue (see [Methods](#)) ([Figures S1A–S1C](#)). This allowed us, for example, to upweight the islet contribution, relative to that for skeletal muscle, for SNPs mapping to enhancers shared between these tissues to account for the different genome-wide enrichment priors observed for these tissues.

Across all tissues, we found that the active transcription start site (TSS) annotation, distinguished by a strong ChIP-seq signal for H3K27ac and H3K4me1 histone modifications, was the most consistently enriched feature (\log_2 fold enrichment from 2.46 to 2.79) ([Figures S1A](#) and [S1B](#)). However, the most highly enriched single annotation detected involved type 1 active enhancers in human islets (as characterized by H3K27ac and H3K4me3) (\log_2 fold enrichment (FE) = 2.84, 95% CI = 1.48–3.62). Coding sequence was also highly enriched for T2D-associated variants (\log_2 FE = 2.59, 95% CI = 2.08–3.01) ([Figure S1B](#)).

In the final step, the tissue-partitioned PPA values were combined across all SNPs in the credible set to yield a set

of TOA scores for each association signal, which preserves the information captured by the fine-mapping ([Figure 1](#), step 3). PPA values corresponding to SNPs not mapping to active regulatory annotations in any of the four evaluated tissues (e.g., repressed or quiescent regions) were allocated to an “unclassified” score (see [Methods](#)). The resulting set of TOA scores for each genetic signal captures the strength of genetic, genomic, and transcriptomic evidence that the signal acts through each of the evaluated tissues. Using this framework, we calculated TOA scores for each of the 380 fine-mapped T2D signals ([Table S1](#)).

TOA Scores Support a Key Role for Strong Enhancers in Human Islets

By combining TOA scores across all 380 signals, we estimated the relative contribution of each tissue to the overall genetic risk of T2D reflected across fine-mapped loci. Islet accounted for the largest share of the cumulative TOA score (29%) with markedly lower contributions from liver, adipose, and skeletal muscle ([Figure 2A](#), inset). Across the 380 loci, 80% of the cumulative TOA score was attributable

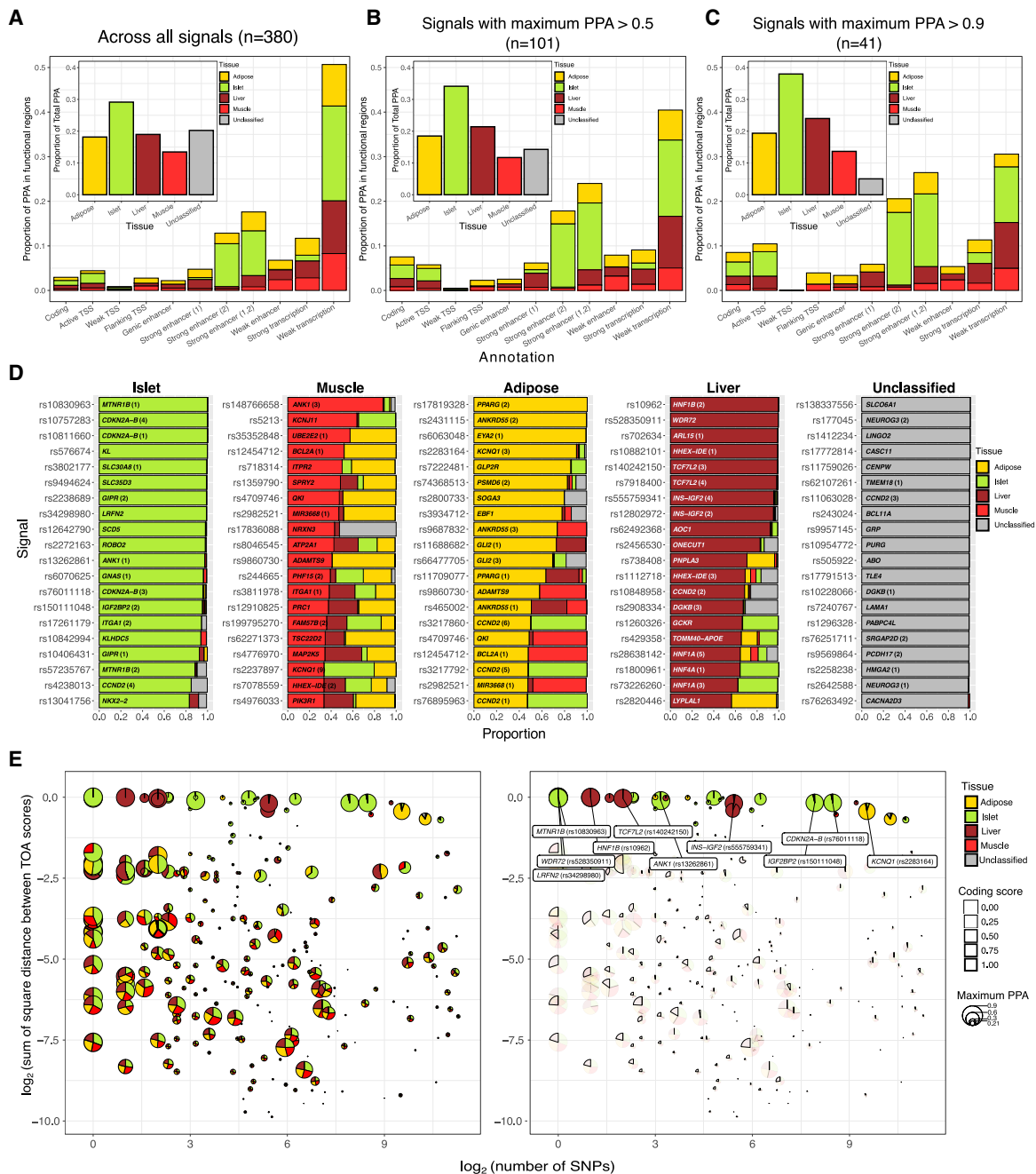


Figure 2. The Profile of Tissue-of-Action Scores across T2D Signals

(A–C) The proportion of total PPA summed across all 380 signals is shown for each tissue (inset). The proportion of total PPA is also shown for each annotation group (outset). Proportions are also exhibited for the subset of signals with maximum credible set PPA > 0.5 in (B) and for the subset of signals with maximum PPA > 0.9 in (C).

(D) The profile of TOA scores is shown for the top 20 signals ranked for each tissue. The locus name and the accession number for the index SNP is indicated for each signal. Signals at loci with multiple conditionally independent signals are indicated by parenthetical numbers (i.e., one is primary signal, two is secondary signal, etc.). Signals in this plot have maximum credible set PPA > 0.5.

(E) Relationship between fine-mapping resolution and TOA score diversity. \log_2 of the number of credible SNPs for each fine-mapped signal is shown on the x axis and the \log_2 value of the sum of square differences between TOA scores for each signal is shown on the y axis (i.e., higher values on the y axis correspond to greater tissue “specificity”). The profiles of TOA scores are indicated within pie charts where the diameter of each circle corresponds to the maximum PPA for the credible set. The line thickness for each circle indicates a coding score for each credible set (i.e., the proportion of cumulative PPA attributable to coding variants). The left panel shows all credible sets with unclassified scores < 0.10 ($n = 259$), and the right panel highlights the subset of “tissue-specific” signals with TOA scores ≥ 0.8 . The ten “tissue-specific” signals with the highest maximum credible set PPA are labeled in the right panel.

to SNPs mapping to coding regions or to active chromatin states in these four tissues (Figure 2A). Within this fraction, SNPs mapping to weakly transcribed regions accounted for the largest share (51%) relative to those mapping to coding and other regulatory annotations (Figure 2A). Overall, weakly transcribed regions account for 23% of the genome (ranging from 22% in skeletal muscle to 26% in islet) and are generally located near other more active annotations (Figure S1D).

Crucially, credible sets vary markedly in their fine-mapping resolution (median credible set size, 42 SNPs; range, 1–3,997 SNPs; median maximum PPA value, 0.24; range 0.01–1.0). We reasoned that the estimates for weakly transcribed regions (and for annotations to tissues outside the four most relevant to diabetes) were most likely inflated by incomplete fine-mapping: less resolved credible sets involving multiple SNPs are likely to map to disparate annotations across tissues. When we evaluated the 101 signals with maximum PPA > 0.5, the TOA score proportions attributed to weak transcription and unclassified proportions decreased to 40% and 14%, respectively (Figure 2B). These proportions further decreased among the 41 signals with maximum PPA > 0.9 (31% and 5% respectively) (Figure 2C). In contrast, the relative contribution of SNPs mapping to strong enhancers increased with greater fine-mapping resolution (from 18% to 26%) (Figures 2A–2C). In particular, the contribution for strong enhancers in islets was disproportionately high among the most finely-mapped signals and underscores a prominent role for these regulatory regions in T2D risk (Figure 2C).

Although the relative TOA score proportions varied with fine-mapping resolution, the contribution from islet was consistently greater than that for liver, adipose, or muscle (by a factor of ≈ 1.5) (Figures 2A–2C, inset). Notably, for credible SNP mapping to strong enhancers, the relative TOA proportions were considerably higher for islets (57%–63%) than for adipose (18%–24%), liver (14%), and skeletal muscle (5%–6%). Increasing fine-mapping resolution tracked with increasing evidence that causal variants were disproportionately concentrated in islet strong enhancers (Figures 2A–2C, outset). When we additionally weighted TOA scores by the adjusted GWAS effect size for each signal (see Methods), the overall islet contribution increased further, albeit slightly, from 29% to 31% across all signals (Figures S2D–S2F). Overall, the profile of TOA scores (particularly across more signals with greater fine-mapping resolution) recapitulates the epigenomic architecture of T2D derived from earlier studies,^{12,13,28} which have indicated that regulatory annotations in islets—and strong enhancers in particular—are important (Figures 2B and 2C).

Distinct TOA Profiles Indicate Pleiotropic Effects in Multiple Tissues

The prime motivation for generating TOA scores was to identify the tissues that most likely mediate disease risk at each genetic signal. We first sought to identify signals where

only a single tissue was likely relevant to disease risk. We found that $\approx 10\%$ (39/380) of signals had profiles where the TOA score for one of the four tissues exceeded a threshold of 0.8, consistent with predominant action in a single tissue: 21 of these involved primary or unique signals at their respective loci, whereas the remaining 18 arose from secondary signals at loci with multiple independent signals (Table S1). Among the primary signals, 14 mapped to islet (including signals at *MTNR1B* [MIM: 600804], *SLC30A8* [MIM: 611145], *CDKN2A/B* [MIM: 600161, 600431] loci), five to liver (e.g., *AOC1* [MIM: 104610], *WDR72* [MIM: 613214]), and two to adipose (*EYA2* [MIM: 601654], *GLP2R* [MIM: 603659]) (Figure 2E and Table S1). No primary signal met this criterion for skeletal muscle: the signal with the highest TOA score for skeletal muscle (0.88) corresponded to a secondary signal (rs148766658) at the *ANK1* (MIM: 612641) locus (Figure 2D). The proportion of signals with TOA profiles consistent with a single TOA increased with greater fine-mapping resolution (17/101 or $\approx 16\%$ of signals with maximum PPA ≥ 0.5) (Table S1).

Aside from these 39 signals, calculated TOA scores for most T2D signals revealed substantial contributions from multiple tissues. We reasoned that this apparent “tissue sharing” could have arisen for two main reasons. The first involves a highly resolved signal from genetic fine-mapping at which the causal variant maps to a single regulatory element active in multiple tissues. The second occurs when a lower resolution signal encompasses many credible set variants that map to distinct regulatory elements with different patterns of tissue specificity. There was some evidence in favor of the latter: maximum credible set PPA values positively correlated with the SSD between TOA scores (i.e., more refined credible sets corresponded to higher measures of tissue specificity) (adjusted $R^2 = 0.04$, p value = 9.8×10^{-5} , Figure S3). However, the magnitude of the effect of fine-mapping resolution on tissue specificity was small (the beta coefficient for the regression of SSD on maximum PPA was 0.17). We conclude that differences in fine-mapping resolution alone do not account for the extent of “tissue-sharing” observed across T2D signals, implying that many signals involved regulatory elements shared across tissues.

To explore this further, we considered signals likely to involve shared effects across tissues on the basis that the difference between the two highest TOA scores was < 0.10 (Table S2). The resulting set of “shared” signals conspicuously spanned the range of mapping resolution, as indicated by the number of credible SNPs and maximum PPA for each signal (Figure 2E). There were eight signals that were fine-mapped to a single credible SNP (i.e., maximum PPA > 0.99) and most clearly demonstrated tissue-shared regulation. This included the primary, non-coding signal at the *PROX1* (MIM: 601546) locus (rs340874) with effects in both islet (TOA = 0.50) and liver (TOA = 0.49): the index SNP at this signal (PPA = 1.0) mapped to a common active TSS in these tissues (Figure S4A, Table S2). This set also included primary signals at the *RREB1*

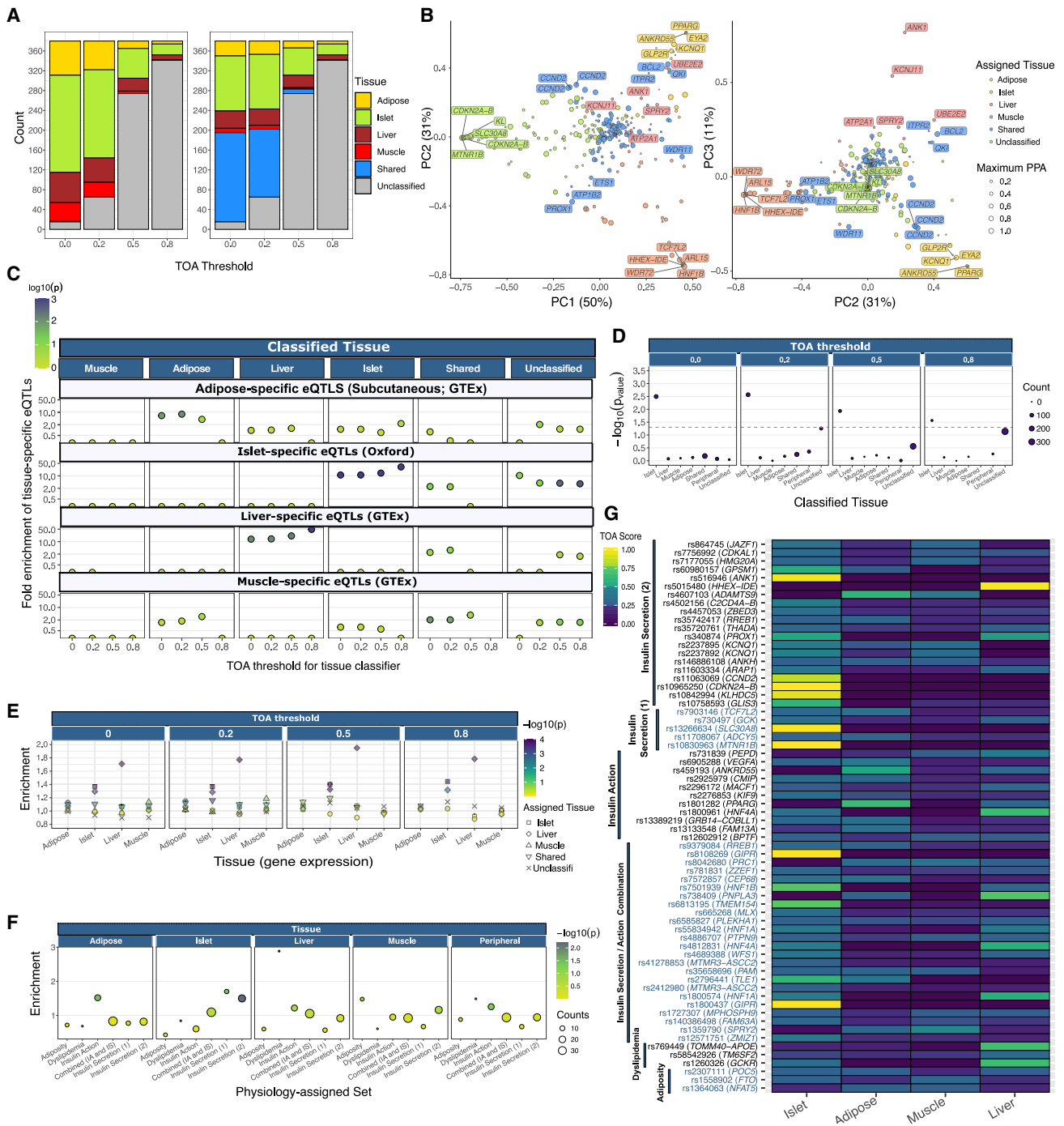


Figure 3. Enrichment of Tissue-Specific Epigenomic and Physiological Features among Classified Signals

(A) Number of signals assigned to each tissue by the classifier for each of the four TOA score thresholds: 0.0, 0.2, 0.5, and 0.8 (left panel). Signal counts are shown across thresholds with a classifier that assigns signals with two or more TOA scores within 0.1 of each other as “shared” signals (right panel).

(B) PCA plots of the decomposition of the TOA score matrix comprising the 306 signals with “unclassified” scores ≤ 0.5 . Each point corresponds to a signal where the size indicates the maximum credible set PPA and the color indicates the assigned tissue at the TOA score threshold ≥ 0.2 via the classifier that included a “shared” designation.

(C) Selective enrichment of tissue-specific eQTLs among credible sets for signals assigned to subcutaneous adipose, islet, liver, and skeletal muscle tissue. Color indicates significance of enrichment ($-\log_{10}$ p value) from permutation tests.

(D) Selective improvement in fine-mapping resolution at islet-assigned signals when richer islet chromatin states are deployed. Comparison of *functional* fine-mapping resolution via a panel of chromatin state annotations based on histone ChIP-seq across the four T2D relevant tissues versus chromatin states based on islet ChIP-seq, ATAC-seq, and DNA methylation (WGBS). p values correspond to Wilcoxon rank-sum test.

(E) Expression similarity of nearest genes annotated to sets of tissue-assigned signals across stringency thresholds. Shape indicates the tissue to which the set of signals were assigned and color shows significance of enrichment ($-\log_{10}$ p value) from permutation tests

(legend continued on next page)

(MIM: 602209) (rs9379084; islet TOA = 0.31; adipose TOA = 0.27; muscle TOA = 0.22), *CCND2* (MIM: 123833) (rs76895963; islet TOA = 0.53; adipose TOA = 0.47), and *BCL2* (MIM: 151430) (rs12454712; muscle TOA = 0.52; adipose TOA = 0.48) loci (Figure S4A, Table S2). There were an additional 33 signals with apparent tissue-sharing where the fine-mapping resolution was somewhat less precise (maximum PPA \geq 0.5). These included the primary signal at the *TCF7L2* (MIM: 602228) locus (rs7903146; adipose TOA = 0.37; islet TOA = 0.31) and secondary signals at *HNF4A* (MIM: 600281) (rs191830490 [liver TOA = 0.40, islet TOA = 0.31] and rs76811102 [islet TOA = 0.32, muscle TOA = 0.25, liver TOA = 0.24]) (Figure S4B, Table S2). Among the total of 101 signals at which the fine-mapping resolution was such as to identify a lead SNP with PPA exceeding 0.5, 41% had evidence that they might involve regulatory effects in two or more tissues.

A Rule-Based Classifier for Assigning Fine-Mapped Signals to Tissues

Because TOA scores appeared to distinguish specific from shared signals (Figures 2D and 2E), we implemented a rule-based classifier that assigns signals to tissues according to their TOA scores across a range of stringencies. A GWAS signal was assigned to a tissue if that tissue had the highest TOA score and exceeded a specified TOA threshold (ranging from permissive thresholds of zero and 0.2 to more stringent thresholds of 0.5 and 0.8). Consistent with the observation that islet accounted for most of the cumulative PPA across loci (Figures 2A–2C), more signals were assigned to islet than to liver, muscle, or adipose tissue across all TOA thresholds. For example, at a TOA threshold of 0.2, 178 signals (47%) were classified as islet, whereas a total of 137 signals (36%) were assigned to insulin-responsive peripheral tissues (58 adipose, 49 liver, 30 muscle) (Figure 3A, left panel). Given the extent of tissue sharing observed across signals, we adapted the classifier scheme to allow for a shared category (defined as above): at the same TOA threshold, this yielded 110 islet, 33 liver, 27 adipose, and 8 muscle signals, plus 137 shared signals (Figure 3A, right panel). These proportional differences between islet, muscle, adipose, and liver were maintained across TOA thresholds (Figure 2D). For example, the distribution of the 39 signals classified at the 0.8 threshold included 22, 10, 6, and 1 signals classified as islet, liver, adipose, and muscle, respectively (Figure 2D).

Principal-component analysis of these data revealed that most variation in TOA scores (50%) distinguished islet signals from those assigned by the classifier to insulin-responsive peripheral tissues, consistent with the distinct functions of these tissues in regulating glucose homeostasis (Figure 3B). The distinction between liver and adipose

signals accounted for a further 31% of variation. Signals classified as shared mapped between the clusters of tissue-assigned signals (Figure 3B). For example, three of the six conditionally independent signals at the *CCND2* locus (including the primary signal at rs76895963; PPA = 1.0) classified as “shared” and mapped equidistant between adipose and islet clusters (Figure 3B, Table S1). Other clear examples include the primary signals at the *PROX1* and *BCL2* loci described above that exhibit profiles with sharing between islet and liver and muscle and adipose, respectively (Figure 3B).

Despite incorporating data from the four tissues most relevant to T2D pathogenesis, a considerable number of signals remained unclassified across stringency thresholds (e.g., 65 signals at the 0.2 threshold), reflecting the appreciable proportion of cumulative PPA at these signals attributable to credible set SNPs that did not map to active regulatory regions in any of these tissues. This can, in part, be explained by the poorer fine-map resolution of these signals compared to classified signals (median credible set size, 57 versus 36 SNPs; median maximum PPA, 0.20 versus 0.25). However, it is possible that some of the unclassified signals involve tissues or cell types not explicitly included in our analysis. Indeed, signals that remained unclassified at the TOA score \geq 0.2 threshold were more likely to map to regions that were actively repressed or quiescent (i.e., low signal) in the four evaluated tissues (Table S3).

Given that a subset of T2D signals are driven by adiposity and presumed to act through central mechanisms,⁴ one obvious omission from the tissues considered in our primary analysis was brain (or, more specifically, hypothalamus). For example, T2D-associated variants at the obesity-associated *MC4R* locus (encoding the melanocortin 4 receptor [MIM: 155541]) were assigned as unclassified in our analyses.^{4,29–32} However, using chromatin state maps from multiple brain regions, we found a deficit, rather than an excess, of PPA enrichment among active enhancers (0.032 versus 0.147; p value = 7.5×10^{-5}) and promoters (0.007 versus 0.043; p value = 0.0054) for unclassified signals (as compared to classified) (Table S3). The data available did not, however, include chromatin state maps for the hypothalamus. Overall, it is to be expected that classification of currently unclassified signals will improve with increased fine-mapping resolution and the availability of detailed chromatin annotations from additional tissue and cell types.

Tissue-Assigned Signals Are Validated by Orthogonal Tissue-Specific Features

We sought to validate the performance of the classifier by evaluating how assignments from the TOA classifier

(F) Selective TOA score enrichment within relevant sets of physiology-assigned signals. Size corresponds to the number-assigned signals in each physiology group and color shows significance of enrichment ($-\log_{10}$ p value) from permutation tests.

(G) Tile plot of TOA scores for physiologically assigned signals. Signals are ordered by physiology group and the corresponding GWAS locus is shown.

matched tissue-specific information from three orthogonal sources: tissue-specific eQTL enrichment, “functional” fine-mapping, and proximity-based gene coexpression analysis of non-coding signals. For these evaluations, we used the version of the classifier that allows for a shared designation.

To determine whether tissue-assigned signals were matched to tissue-specific eQTLs, we assembled *cis*-eQTLs for liver, skeletal muscle, subcutaneous adipose tissue (all GTEx v7) and human islets⁶ and defined sets of tissue-specific eQTLs (see [Methods](#)). The set of signals assigned by the TOA classifier to islets was significantly, and selectively, enriched for islet-specific eQTLs across all TOA thresholds (ranging from 11-fold to 31-fold enrichment [p values < 0.001]) as compared to matched sets of SNPs (see [Methods](#)) ([Figure 3C](#)). Similarly, the set of signals assigned by the TOA classifier to liver showed marked, selective enrichment for liver-specific eQTLs across TOA thresholds ([Figure 3C](#)). Overall, the more confidently assigned genetic signals retained at more stringent TOA thresholds tended to have larger point effect estimates, though the reduced number of signals meeting the more stringent thresholds led to wider confidence intervals and some reduction in the statistical significance of the enrichments. Relatively few signals were assigned to adipose and skeletal muscle at higher thresholds ([Figure 3A](#)); nonetheless, adipose-assigned signals were the most enriched for adipose-specific eQTLs at lower stringency (e.g., 6-fold enrichment, p value = 0.009, at the 0.2 threshold [[Figure 3C](#)]). In contrast, although sets of signals classified as shared showed some enrichment for tissue-specific eQTLs at less stringent thresholds, these enrichments were generally lower than those for signals assigned to the corresponding tissues ([Figure 3C](#)). These data indicate that the tissue assignments made by the classifier are consistent with the information from *cis*-eQTL analyses in corresponding tissues.

The second validation analysis was motivated by the use of high-resolution epigenomic maps to improve genetic fine-mapping. For the present study, we had derived TOA scores by using chromatin states based solely on ChIP-seq data:²⁰ this was a conscious decision designed to minimize technical differences in the depth of annotation available between tissues given that chromatin accessibility and DNA methylation data were not as widely available. However, we had previously shown that islet enhancer chromatin states obtained from a segmentation analysis that incorporated information from DNA methylation, ATAC-seq, and histone ChIP-seq data yielded higher enrichment of T2D-associated SNPs than enhancer states delineated from ChIP-seq data alone.²⁵ We reasoned that accurate assignment of islet signals by the TOA classifier would identify signals that would considerably improve in fine-mapping resolution with the inclusion of more fine-grained islet functional information. To test this hypothesis, we performed a comparative “functional” fine-mapping analysis (see [Methods](#)) using this richer set of islet annotations²⁵ and found that the mean maximum cred-

ible set PPA significantly increased for islet-assigned signals relative to the corresponding value from a joint analysis based on ChIP-seq data alone (e.g., mean PPA increase = 0.064; p value = 0.0027 at the 0.2 threshold) ([Figure 3D](#)). This was true across all TOA thresholds. In contrast, credible sets for signals assigned to insulin-responsive peripheral tissues showed no improvement in fine-mapping resolution with the richer islet annotations ([Figure 3D](#)). These data indicate that the tissue assignments made by the TOA classifier are consistent with the information from more detailed functional annotations in relevant tissues.

The third validation approach involved assessing genes for expression similarity and overlapping coexpression.³³ Although the genes lying closest to the lead regulatory variants at GWAS signals are not guaranteed to be the causal transcript, the set of “nearest genes” is, nonetheless, likely to be enriched for the genes responsible for mediating such associations.³⁴ As such, we reasoned that performance of the classifier would be reflected in the extent to which genes near non-coding signals were expressed in the corresponding tissue as compared to more distal genes. We assigned a single (nearest) gene to each tissue-classified signal and found that the set of genes nearest to islet-assigned signals showed the most pronounced similarity in expression levels in human islet tissue across all TOA thresholds (e.g., p value = 0.0003 at threshold 0.8) ([Figure 3E](#)) and across an expanded set of tissues, including 53 tissues from the GTEx Project ([Figure S5A](#)). This expression signal was lost for the sets of second- and third-nearest genes ([Figures S5B](#) and [S5C](#)). Similar results were observed for liver, muscle, and adipose ([Figure 3E](#)). In contrast to the sets of nearest genes annotated to signals assigned to specific tissues, gene sets annotated to signals classified as either “shared” or “unclassified” did not show pronounced similarity in expression levels in any of the evaluated tissues ([Figure 3E](#), [Figure S5A](#)). We next evaluated co-expression by measuring pairwise Spearman correlations of gene expression within each of the four T2D-relevant tissues. At the 0.2 threshold, we found that genes proximal to islet-assigned signals were significantly correlated in human islet (p value = 0.0037), whereas genes proximal to “shared” signals were significantly correlated across all tissues (the extent of correlation is greater in adipose, liver, and skeletal muscle) ([Figure S6](#)). These data indicate that the tissue assignments made by the classifier are consistent with the information from co-expression analyses in corresponding tissues. Collectively, the data from these three analyses further supports the validity of the TOA scores generated by our approach.

Tissue-Assigned Signals Are Supported by Physiological Clustering

It is possible to assign T2D risk alleles with respect to physiological impact on the basis of patterns of genetic association with related quantitative traits such as fasting glucose and insulin levels, circulating lipid levels, and

anthropometric traits.^{2,3,26,35,36} At the same time, those same physiological processes map to specific tissues (e.g., insulin secretion from pancreatic islets). We asked, therefore, whether the tissue assignment of signals by the TOA classifier (based on tissue-specific molecular data) was consistent with the assignments made on the basis of whole-body physiology. We focused on a set of 82 T2D-associated variants that had previously been partitioned via a “fuzzy” clustering algorithm³ to six physiological clusters and were in LD with lead variants from the set of 380 fine-mapped credible sets (see [Methods](#)).

We first asked whether these signals assigned to these six physiological clusters differed with respect to their TOA score distributions. Variants assigned to the two “insulin secretion” clusters (characterized by associations with reduced fasting glucose and HOMA-B levels but differing with respect to effects on proinsulin and HDL cholesterol levels) had higher islet TOA scores than variants in the other physiological clusters (enrichment = 1.5, 1.7 [$p = 0.006, 0.03$] for the type 2 and type 1 insulin secretion cluster, respectively) ([Figures 3F and 3G](#)). Variants assigned to the “insulin action” and “dyslipidemia” clusters corresponded to signals with significantly higher adipose (1.5-fold, $p = 0.034$) and liver scores (2.9-fold, $p = 0.009$), respectively ([Figures 3F and 3G](#)). Reciprocally, sets of TOA classifier tissue-assigned signals were significantly enriched for SNPs from relevant physiology sets ([Figure S7A](#)). Similar results were obtained from a different (but overlapping) set of physiological clusters derived with an alternative clustering scheme²⁶ ([Figures S7B and S7C](#)).

These patterns were confirmed by evaluating enrichment across all phenotypes present in the NHGRI-EBI GWAS catalog. For example, T2D signals assigned to adipose by the TOA classifier were enriched for variants associated with traits relevant to fat distribution (e.g., waist-to-hip ratio adjusted for BMI, 3.5-fold, p value < 0.0001), whereas signals assigned to liver and islet were enriched for SNPs associated with total cholesterol levels (3.3-fold, p value = 0.0011) and acute insulin response (2.3-fold, p value = 0.009), respectively ([Figure S8](#)). Collectively, these results indicate that tissue assignments based on TOA scores derived from molecular data are consistent with inference based on *in vivo* physiology.

Epigenomic Clustering Implicates Multiple Tissues at Loci with Independent Signals

The 380 fine-mapped genetic credible sets map to 239 loci, 84 of which harbored multiple conditionally independent signals.⁴ As disparate signals within the same locus cannot be assumed, purely on the basis of genomic adjacency, to influence disease risk through the same downstream mechanism, we asked how often the classifier assigned independent signals at a locus to different tissues. We focused on the 0.2 threshold because this allowed us to assign signals to each of the four T2D-relevant tissues while still being widely validated by the approaches described

above ([Figure 3](#)). There were 60 loci where at least two signals were assigned to a tissue or designated as “shared” ([Figure S9](#)), but we focused on 19 loci where two or more independent signals received tissue-specific assignments (rather than “shared”). Of these, there were nine loci where constituent signals were given identical tissue assignments. These included *PPARG* (MIM: 601487) and *EYA2* (all signals designated as adipose) and seven others—including *MTNR1B* and *GIPR* (MIM: 137241)—at which all signals were assigned to islet ([Figure 4A](#)).

This left ten loci where there was divergent assignment of signals. One of the clearest examples involves the *HNF1B* (MIM: 189907) locus where three signals (each comprising non-coding variants) varied markedly in their TOA scores from islet and liver ([Figure 4B](#)). The lead signal, at rs10908278, was assigned to islet because the credible variants with the highest PPAs (0.72 and 0.13) both mapped to the same strong islet-specific enhancer ([Figure 4C](#)). In contrast, the rs10962 signal was assigned to liver as the likely causal variant (PPA = 0.98) mapped to a strongly transcribed region specific to liver. The remaining signal, at rs2189301, was classified as “shared” because the principal credible set variants (both with PPA = 0.49) mapped to a transcribed region in both islet and liver, and the latter showed a stronger epigenomic signature for transcription ([Figure 4C](#)).

Large-scale GWAS meta-analysis in Europeans has uncovered multiple signals at the *ANK1* locus. One of these, at rs13262861, colocalizes with an eQTL for *NKX6.3* (MIM: 610772) expression in pancreatic islets.⁴ Using the TOA classifier, we found that this signal (rs13262861; PPA = 0.97) was designated as islet given overlap with a strong islet enhancer. On the other hand, an independent signal at rs148766658 (43 kb from rs13262861) was categorized as a muscle signal because credible set SNPs (maximum PPA = 0.25) mapped to strong enhancer and transcribed chromatin states in skeletal muscle ([Figures S10A and S10B](#)). These data suggest that this “locus” is really a composite of overlapping associations, with entirely distinct effector transcripts and TOAs. Notably, a recent GWAS meta-analysis of T2D in 433,530 East Asians has uncovered independent signals in this region that distinctly colocalize with either an eQTL for *NKX6-3* in islet or an eQTL for *ANK1* expression in skeletal muscle and subcutaneous adipose tissues.³⁷ Although there is incomplete LD between the specific *ANK1* variants detected in the European and East Asian meta-analyses (between the secondary signals in particular), our results are consistent with the presence of distinct signals near *ANK1* with disparate tissue effects. This example highlights the growing limitations of attributing shared functional relationships to nearby genetic signals solely on the basis of their proximity. Instances such as this, where proximal signals represent functionally distinct mechanisms, indicate that such assumptions can be misleading and are likely to become less tenable as the density of GWAS hits increases.

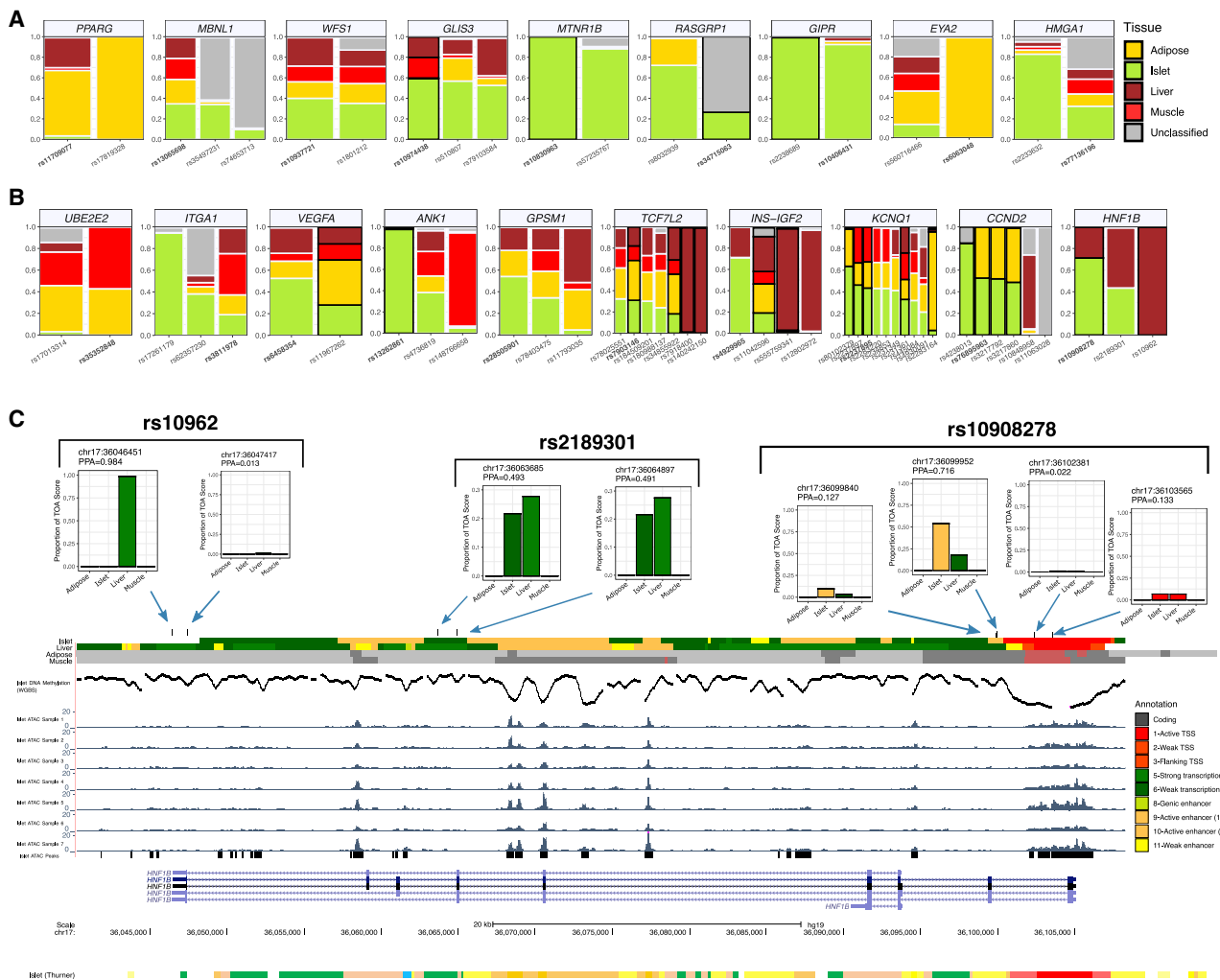


Figure 4. Multiple Tissues Implicated by Epigenomic Scores at Heterogenous Loci

(A) Profile of TOA scores for the nine loci with all signals receiving identical, “non-shared” tissue assignments at the 0.2 stringency threshold.

(B) Profile of TOA scores for the ten loci with all signals receiving distinct, “non-shared” tissue assignments at the 0.2 threshold.

(C) Epigenomic profile of PPA values attributable to each credible SNP of the primary signal at the *HNF1B* locus. For each credible SNP, the PPA value attributable to each tissue annotation is shown along with its position on chromosome 17 (genome build hg19). Chromatin state maps for islet, adipose, muscle, and liver tissue from Varshney et al.²⁰ are shown along with ATAC-seq tracks for seven representative islet samples, called ATAC-seq peaks from a set of islet ATAC samples ($n = 17$), and DNA methylation (whole genome bisulfite sequencing) in human islets from Thurner et al.²⁵

Among the ten loci displaying evidence for “tissue heterogeneity” across signals was *TCF7L2*. Of the seven independent signals at *TCF7L2* revealed by conditional fine-mapping, two (at rs7918400 and rs140242150) were assigned solely to liver (Figure 4B). The remaining five signals revealed contributions from both islet and adipose (Figure 4B). This group includes the lead signal at *TCF7L2* (lead SNP, rs7903146), which remains the strongest common variant T2D association in Europeans. This signal was classified as “shared,” with similar TOA scores from islet (0.31) and adipose (0.37). Crucially, this signal did not fine-map exclusively to rs7903146 (PPA = 0.59; MAF = 0.26) in Europeans: the 99% credible set included two additional SNPs⁴ (rs35198068 and rs34872471). Whereas rs35198068 had a PPA value of

only 0.05, rs34872471 had a PPA value of 0.36 and is in near perfect LD ($r^2 = 0.99$) with rs7903146 in Europeans.³⁸ Notably, rs7903146 has a pronounced signature in both islet and adipose due to its mapping to an epigenetically active region in these tissues (a strong enhancer with, at least in islet, high chromatin accessibility and low DNA methylation). On the other hand, rs34872471 mapped to a strong enhancer only in adipose (Figure S10C). The net effect, based on this information, is a “shared” designation. In truth, either there is a single causal variant at this locus (rs7903146, or potentially rs34872471), and once resolved, this signal can be correctly assigned to the relevant tissue, or both SNPs are directly contributing to T2D risk through distinct mechanisms in islet and adipose tissue.

TOA Scores Advance Resolution of Effector Transcripts

Given the TOA score classifier was able to discriminate sets of genetic signals that were supported by orthogonal validation features, we next considered the value of TOA scores to clarify regulatory mechanisms and enhance the identification of downstream effector transcripts at T2D-associated loci. One widely used approach for promoting candidate causal genes at GWAS loci involves identifying *cis*-eQTL signals that colocalize with trait-associated SNPs.^{39,40} However, *cis*-eQTL signals show appreciable tissue specificity, raising the possibility of misleading inference if analyses are conducted in a tissue irrelevant to the signal of interest.^{41,42} For example, a *cis*-eQTL specific to liver is likely to be more informative for a T2D signal assigned to liver than one assigned to islet.

We explored the utility of incorporating TOA scores for T2D-relevant tissues into a previous colocalization analysis.⁴ To do so, we evaluated eQTL colocalization results for the 101 T2D GWAS signals that had lead SNPs with maximum PPA ≥ 0.5 . A total of 378 eQTL colocalizations (eCaviar colocalization posterior probability [CLPP] ≥ 0.01) were detected across 53 signals with a median of four colocalizations (implicating four distinct pairs of tissues and eGenes) per signal (Table S4). At some loci, the number of colocalizations detected was substantial: at the *CLUAP* (MIM: 616787) locus, for example, the lead T2D SNP (rs3751837, PPA = 0.90) was the source of 64 *cis*-eQTL colocalizations involving 15 eGenes across 37 tissues (Table S4).

Restricting colocalization results to those SNP-gene pairs arising from the tissue assignments provided by the TOA classifier (at a threshold of 0.2) reduced the number of colocalizations to 133 at 32 signals, a 65% reduction overall, and a 36% reduction (from 209 at 49 signals) if considering only the subset of colocalizations that involved the four T2D-relevant tissues (Table S5). This reduced set of TOA-filtered colocalizations retained many of the T2D effector transcripts previously reported in the literature, including those benefiting from additional chromatin conformation data.^{8,9} For example, the primary signal at the *CDC123* (MIM: 617708) locus (rs11257655; PPA = 1.0) was classified as an islet signal (TOA = 0.40) and has been previously reported to colocalize with an eQTL for *CAMK1D* (MIM: 607957) expression in human islets.^{6,19} The regulatory element harboring this variant was recently shown, via promoter capture HiC, to physically interact with the *CAMK1D* promoter in human islet cells.⁹ Similarly, the designation of islet signals at the *MTNR1B* (rs10830963; PPA = 1.0; TOA = 1.0) and *IGF2BP2* (rs150111048; PPA = 0.94; TOA = 0.96) loci was consistent with colocalized eQTLs implicating *MTNR1B* and *IGF2BP2* as effector genes at these loci influencing T2D risk through effects on human islet function.^{6,8}

At other signals, the integration of TOA scores with eQTL colocalization data allowed us to further resolve signals that featured multiple candidate eGenes in T2D-relevant tissues. For example, the lead SNP at the *CCND2* locus

(rs76895963; PPA = 1) has 16 eQTL colocalizations, involving three eGenes across 11 tissues. Of these, only two involved any of the four T2D-relevant tissues, implicating *CCND2* expression in subcutaneous adipose (CLPP = 1.0) and skeletal muscle (CLPP = 1.0). From a TOA perspective, this signal was classified as “shared” with high TOA scores for both islet (0.53) and adipose (0.47). This suggests that of the two colocalized eQTLs, the eQTL affecting *CCND2* expression in adipose tissue is likely to be more important to T2D pathophysiology. *CCND2* encodes cyclin D2, a signaling protein involved in cell cycle regulation and cell division. Consistent with our inference, *CCND2* was previously shown to be differentially expressed between insulin-sensitive and insulin-resistant individuals in subcutaneous adipose tissue but not in skeletal muscle.⁴³

At the *CLUAP1* locus, referred to above, the lead signal (rs3751837) was classified as “shared” with comparable TOA scores across each of the four T2D-relevant tissues (0.22–0.29). Restricting to these four tissues reduced the overall number of colocalizations (across genes and tissues) from 64 to 16. Of the remaining colocalized eQTLs, the highest colocalization posterior probability (CLPP = 0.41) corresponded to an eQTL where the T2D-risk allele associates with increased expression of *TRAP1* in subcutaneous adipose (Table S5). This variant is also associated with *TRAP1* expression in skeletal muscle. *TRAP1* encodes TNF receptor-associated protein 1, a chaperone protein that expresses ATPase activity and functions as a negative regulator of mitochondrial respiration, modulating the metabolic balance between oxidative phosphorylation and aerobic glycolysis.⁴⁴ Although *TRAP1* has not been directly implicated in T2D risk, a proteomic analysis has previously found *TRAP1* protein levels to be differentially abundant in cultured myotubes from T2D patients versus normal glucose tolerant donors.⁴⁵ Further experimental validation will be required to resolve the effector transcript(s) at this and other T2D-associated loci. However, collectively these results, demonstrate that TOA scores can be systematically incorporated into integrative analyses to prioritize effector transcripts, particularly when there are multiple candidate genes in multiple relevant tissues.

Discussion

We have developed a principled and extensible approach for integrative multi-omic analysis to advance the resolution of genetic mechanisms at disease-associated loci by elucidating relevant TOAs. Existing approaches in this space have focused on characterizing the contributions of tissue- and cell-type-specific regulatory features to the overall genetic architecture of the complex trait of interest (e.g., through genome-wide enrichment or heritability partitioning). However, to ensure that functional follow-up is directed to appropriate cellular systems, it is also critical to understand tissue- and cell-type-specific effects at each

individual signal. In line with previous work, our analyses support a prominent role for pancreatic islets in the pathogenesis of T2D, but these results also emphasize the extent to which risk-associated variants may involve shared effects across multiple tissues. Some of this tissue “sharing” was the result of incomplete resolution of causal variants at less-well fine-mapped signals. However, we also found multiple examples of fine-mapped signals that overlapped regulatory elements active in multiple tissues (pointing to pleiotropic effects across tissues) as well as of loci where independent signals manifested diverse TOA profiles.

A salient exemplar of these scenarios for tissue “sharing” is the *TCF7L2* locus that plays a distinguished, but as yet mechanistically unresolved, role in T2D pathogenesis and is complicated by pronounced allelic heterogeneity. The TOA for the lead signal at rs7903146 has been the subject of recent debate: early studies emphasized consequences focused on islet dysfunction, whereas recent data have supported a role in adipose tissue.^{28,46} Evidence from murine studies has supported an important role for *Tcf7l2* in pancreatic β -cell proliferation, insulin secretion, and glucose homeostasis.^{47–50} In human studies, variation at rs7903146 has been associated with chromatin accessibility and *TCF7L2* gene expression in islets.^{19,28} However, *TCF7L2* activation also regulates Wnt signaling during adipogenesis, and *in vivo* deactivation of *TCF7L2* protein in mature adipocytes results in hepatic insulin resistance and systemic glucose intolerance.⁴⁶ *TCF7L2* expression was also found to be downregulated in human subjects with impaired glucose tolerance and adipocyte insulin resistance.⁴⁶ Our TOA analysis of this signal yielded a profile that is consistent with shared effects in both pancreatic islets and adipocytes that jointly contribute to T2D pathogenesis. In addition, two independent signals at this locus (rs7918400 and rs140242150) had profiles that suggest a primary mechanism of action in liver, a possibility supported by *in vivo* studies linking liver-specific perturbations of *Tcf7l2* expression in adult mice to altered hepatic glucose production and glucose production.^{51,52} Overall these data lend credence to the idea that the impact of genetic variation at this locus on T2D risk is mediated through several parallel mechanisms operating via multiple tissues. This may explain why it has such a comparatively large effect on T2D risk in humans.

Given the important role that skeletal muscle plays in insulin action (i.e., postprandial insulin response), the paucity of signals assigned to this tissue is conspicuous. However, it is worth pointing out that skeletal muscle was actually well represented among “shared” signals, particularly among signals that had sizable TOA contributions from adipose tissue. Therefore, rather than discounting the relevance of skeletal muscle to T2D, our results are consistent with a genetic architecture wherein disease-associated variants that impact skeletal muscle are also likely to have effects in other tissues rather than eliciting effects specific to skeletal muscle. A practical implication

of this observation is that the inclusion of tissues that share, at least partially, physiological activity (e.g., insulin action) and related molecular “machinery” (e.g., membrane receptors, secondary messengers, etc.) is that they are more likely to be classified as “shared” on the basis of TOA scores. However, the inclusion of higher resolution tissue and cellular annotations, as discussed below, can be leveraged to refine tissue assignments and potentially reduce the number of “shared” signals.

In this study, we have incorporated gene-level expression data and publicly available chromatin states based on histone ChIP-seq to determine TOAs at loci associated with T2D. This scheme yielded tissue designations that were supported by validation analyses (e.g., functional fine-mapping and physiological clustering) and are consistent with previously elucidated effector mechanisms at specific loci. However, such tissue designations, though informative, constitute a first step and will undoubtedly become more refined with the increasing availability and incorporation of higher resolution datasets. In particular, our approach will benefit from more extensive genetic fine-mapping that will accompany large-scale discovery efforts involving greater samples, denser imputation reference panels, and the inclusion of more diverse populations representing underrepresented genetic ancestries.

The performance of our approach will also improve with regulome maps delineated from chromatin segmentation or hierarchical clustering analyses based on an expanded set of input features (e.g., histone post-translational modification [PTM] and transcription factor ChIP-seq, DNA methylation, chromatin accessibility). This allows more of the genome to be assigned to a regulatory state. For example, incorporating ATAC-seq and whole-genome bisulfite sequencing, in addition to histone PTM ChIP-seq data, into a chromatin segmentation analysis of human islets reduced the proportion of quiescent regions from 6.6% to 3.1%.^{20,25} Interestingly, islet enhancer annotations characterized by the presence of mediator binding were recently shown to exhibit a notably strong enrichment of islet-specific chromatin interactions;⁹ the inclusion of such input features would help to delineate regulatory annotations that can further differentiate tissue effects. Similarly, the elucidation of tissue-specific effects at coding variants will benefit from long-read RNA sequencing methods that can leverage patterns of isoform expression. In principle, measures of relative protein abundances assayed from proteomic technologies (i.e., mass spectrometry, immunoassays, aptamer-based methods) may also inform relevant tissues with respect to coding variants. However, available proteomic datasets are more sparse than RNA-sequenced datasets because of the comparatively lower proteome coverage achievable in high-throughput.⁵³ Furthermore, discerning molecular features under a spectrum of biological contexts (e.g., hyperglycemia, developmental stages) will provide valuable insight into the specific conditions, within TOAs, that are most relevant to individual genetic signals.

Lastly, incorporating regulatory information ascertained from single-cell approaches (e.g., scRNA-seq and snATAC-seq) will advance the resolution of “cells-of-action” against different physiological backdrops. Indeed, it may be the case that some of the tissue sharing observed in this study is reflecting cell type composition *within* tissues rather than sharing *across* tissues. The inclusion of single-cell regulome maps will help resolve this question. Notably, the inclusion of closely related tissues or tissue subtypes would most likely increase the observed number of “shared” signals as a result of variants mapping to functional elements shared between more similar tissues (e.g., subcutaneous and visceral adipose tissue) than between tissues with more distinct physiological roles (e.g., adipose and islet tissue).

The strategy presented here for integrating multi-omic information can provide valuable insight for prioritizing variants and determining appropriate model systems to employ in experimental validation studies. This scheme may also enhance the construction of process-specific genetic risk scores that can identify and profile individuals with genetic burden that impacts pathophysiological processes impacting specific tissues and organ systems. Lastly, this approach can be deployed more widely across other complex diseases, especially as more tissue- and cell-specific data become available. To support this wider use, we have implemented our method and made it openly available in an R package: Tissue of ACTion scores for Investigating Complex trait-Associated Loci (TACTICAL).

Data and Code Availability

The code scripts used to perform the bioinformatic and statistical analyses described in this study can be accessed from a GitHub directory through the following URL: https://github.com/Jmtorres138/t2d_classification/. The method described in this study has been implemented in an R package titled TACTICAL (Tissue of ACTion scores for Investigating Complex trait-Associated Loci). The package can be installed from GitHub through the following URL: <https://github.com/Jmtorres138/TACTICAL>.

Supplemental Data

Supplemental Data can be found online at <https://doi.org/10.1016/j.ajhg.2020.10.009>.

Acknowledgments

A.L.G. is a Wellcome Trust Senior Fellow in Basic Biomedical Science. M.I.M. was a Wellcome Senior Investigator and an NIHR Senior Investigator. This work was funded in Oxford by the Wellcome Trust (095101 [A.L.G.], 200837 [A.L.G.], 098381 [M.I.M.], 106130 [A.L.G., M.I.M.], 203141 [A.L.G., M.I.M.], 212259 [M.I.M.]), Medical Research Council (MR/L020149/1) [M.I.M., A.L.G.], European Union Horizon 2020 Programme (T2D Systems) [A.L.G.], NIH (U01-DK105535; U01-DK085545) [M.I.M., A.L.G.],

and NIHR (NF-SI-0617-10090) [M.I.M.]. The research was funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) [A.L.G., M.I.M.]. A.P. was supported by the Rhodes Trust, the Natural Sciences and Engineering Research Council of Canada, and the Canadian Centennial Scholarship Fund. This work was also supported by Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health.

Declaration of Interests

M.I.M. has served on advisory panels for Pfizer, NovoNordisk, and Zoe Global; has received honoraria from Merck, Pfizer, NovoNordisk, and Eli Lilly; and has received research funding from Abbvie, AstraZeneca, Boehringer Ingelheim, Eli Lilly, Janssen, Merck, NovoNordisk, Pfizer, Roche, Sanofi Aventis, and Servier Takeda. As of June 2019, M.I.M. is an employee of Genentech and holds stock in Roche. A.M. is now an employee of Genentech and holds stock in Roche.

Received: June 25, 2020

Accepted: October 20, 2020

Published: November 12, 2020

Web Resources

1000 Genomes Project data, <http://ftp.1000genomes.ebi.ac.uk>
Diabetes Epigenome Atlas (chromatin state maps from Varshney et al.²⁰), <https://www.diabetesepigeneome.org/>
DIAGRAM website, <https://www.diagram-consortium.org>
Ensembl gene annotations, <https://www.ensembl.org>
fgwas software, <https://github.com/joepickrell/fgwas>
GTEx Portal website, <https://gtexportal.org>
LD Link, <https://ldlink.nci.nih.gov>
NHGRI-EBI GWAS catalog, <http://www.ebi.ac.uk/gwas>
Online Mendelian Inheritance in Man, <https://www.omim.org/>

References

1. Xue, A., Wu, Y., Zhu, Z., Zhang, F., Kemper, K.E., Zheng, Z., Yengo, L., Lloyd-Jones, L.R., Sidorenko, J., Wu, Y., et al.; eQTLGen Consortium (2018). Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* 9, 2941.
2. Scott, R.A., Scott, L.J., Mägi, R., Marullo, L., Gaulton, K.J., Kaakinen, M., Pervjakova, N., Pers, T.H., Johnson, A.D., Eicher, J.D., et al. (2017). An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* 66, 2888–2902.
3. Mahajan, A., Wessel, J., Willems, S.M., Zhao, W., Robertson, N.R., Chu, A.Y., Gan, W., Kitajima, H., Taliun, D., Rayner, N.W., et al.; ExomeBP Consortium; MAGIC Consortium; and GIANT Consortium (2018a). Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* 50, 559–571.
4. Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A.,

- Grarup, N., et al. (2018b). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* *50*, 1505–1513.
5. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* *101*, 5–22.
 6. van de Bunt, M., Manning Fox, J.E., Dai, X., Barrett, A., Grey, C., Li, L., Bennett, A.J., Johnson, P.R., Rajotte, R.V., Gaulton, K.J., 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.
 7. Khamis, A., Canouil, M., Siddiq, A., Crouch, H., Falchi, M., Bulow, M.V., Ehehalt, F., Marselli, L., Distler, M., Richter, D., et al. (2019). Laser capture microdissection of human pancreatic islets reveals novel eQTLs associated with type 2 diabetes. *Mol. Metab.* *24*, 98–107.
 8. Greenwald, W.W., Chiou, J., Yan, J., Qiu, Y., Dai, N., Wang, A., Nariai, N., Aylward, A., Han, J.Y., Kadakia, N., et al. (2019). Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. *Nat. Commun.* *10*, 2078.
 9. Miguel-Escalada, I., Bonàs-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Atla, G., Javierre, B.M., Rolando, D.M.Y., Farabella, I., Morgan, C.C., et al. (2019). Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nat. Genet.* *51*, 1137–1148.
 10. Khetan, S., Kursawe, R., Youn, A., Lawlor, N., Jillette, A., Marquez, E.J., Ucar, D., and Stitzel, M.L. (2018). Type 2 Diabetes-Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets. *Diabetes* *67*, 2466–2477.
 11. Thomsen, S.K., Ceroni, A., van de Bunt, M., Burrows, C., Barrett, A., Scharfmann, R., Ebner, D., McCarthy, M.I., and Gloyn, A.L. (2016). Systematic Functional Characterization of Candidate Causal Genes for Type 2 Diabetes Risk Variants. *Diabetes* *65*, 3805–3811.
 12. Parker, S.C.J., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., van Bueren, K.L., Chines, P.S., Narisu, N., Black, B.L., 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.
 13. Pasquali, L., Gaulton, K.J., Rodríguez-Seguí, S.A., Mularoni, L., Miguel-Escalada, I., Akerman, Í., Tena, J.J., Morán, I., Gómez-Marín, C., van de Bunt, M., et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* *46*, 136–143.
 14. Scott, L.J., Erdos, M.R., Huyghe, J.R., Welch, R.P., Beck, A.T., Wolford, B.N., Chines, P.S., Didion, J.P., Narisu, N., Stringham, H.M., et al. (2016). The genetic regulatory signature of type 2 diabetes in human skeletal muscle. *Nat. Commun.* *7*, 11764.
 15. Zhong, H., Beaulaurier, J., Lum, P.Y., Molony, C., Yang, X., Macneil, D.J., Weingarh, D.T., Zhang, B., Greenawalt, D., Dobrin, R., et al. (2010). Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet.* *6*, e1000932.
 16. Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* *81*, 208–227.
 17. Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M., Auton, A., Myers, S., Morris, A., et al.; Wellcome Trust Case Control Consortium (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* *44*, 1294–1301.
 18. The GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* *348*, 648–660.
 19. Viñuela, A., Varshney, A., van de Bunt, M., Prasad, R.B., Asplund, O., Bennett, A., Boehnke, M., Brown, A.A., Erdos, M.R., Fadista, J., et al. (2020). Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. *Nat. Commun.* *11*, 4912.
 20. Varshney, A., Scott, L.J., Welch, R.P., Erdos, M.R., Chines, P.S., Narisu, N., Albanus, R.D., Orchard, P., Wolford, B.N., Kursawe, R., et al.; NISC Comparative Sequencing Program (2017). Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc. Natl. Acad. Sci. USA* *114*, 2301–2306.
 21. Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* *12*, 2478–2492.
 22. Kundaje, A., Meuleman, W., Ernst, J., Bilenyk, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. *Nature* *518*, 317–330.
 23. 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.
 24. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* *7*, 500–507.
 25. Thurner, M., van de Bunt, M., Torres, J.M., Mahajan, A., Nylander, V., Bennett, A.J., Gaulton, K.J., Barrett, A., Burrows, C., Bell, C.G., et al. (2018). Integration of human pancreatic islet genomic data refines regulatory mechanisms at Type 2 Diabetes susceptibility loci. *eLife* *7*, e31977.
 26. Udler, M.S., Kim, J., von Grotthuss, M., Bonàs-Guarch, S., Cole, J.B., Chiou, J., Boehnke, M., Laakso, M., Atzmon, G., Glaser, B., et al.; Christopher D. Anderson on behalf of METASTROKE and the ISGC (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.
 27. Franks, P.W., and McCarthy, M.I. (2016). Exposing the exposures responsible for type 2 diabetes and obesity. *Science* *354*, 69–73.
 28. Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D., et al. (2010). A map of open chromatin in human pancreatic islets. *Nat. Genet.* *42*, 255–259.
 29. Fan, W., Boston, B.A., Kesterson, R.A., Hruby, V.J., and Cone, R.D. (1997). Role of melanocortinergic neurons in feeding and the agouti obesity syndrome. *Nature* *385*, 165–168.
 30. Fan, W., Voss-Andreae, A., Cao, W.H., and Morrison, S.F. (2005). Regulation of thermogenesis by the central melanocortin system. *Peptides* *26*, 1800–1813.
 31. Xi, B., Chandak, G.R., Shen, Y., Wang, Q., and Zhou, D. (2012). Association between common polymorphism near the MC4R gene and obesity risk: a systematic review and meta-analysis. *PLoS ONE* *7*, e45731.
 32. Martinelli, C.E., Keogh, J.M., Greenfield, J.R., Henning, E., van der Klaauw, A.A., Blackwood, A., O’Rahilly, S., Roelfsema,

- F, Camacho-Hübner, C., Pijl, H., and Farooqi, I.S. (2011). Obesity due to melanocortin 4 receptor (MC4R) deficiency is associated with increased linear growth and final height, fasting hyperinsulinemia, and incompletely suppressed growth hormone secretion. *J. Clin. Endocrinol. Metab.* *96*, E181–E188.
33. Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Res.* *14*, 1085–1094.
 34. Stacey, D., Fauman, E.B., Ziemek, D., Sun, B.B., Harshfield, E.L., Wood, A.M., Butterworth, A.S., Suhre, K., and Paul, D.S. (2019). ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res.* *47*, e3.
 35. Dimas, A.S., Lagou, V., Barker, A., Knowles, J.W., Mägi, R., Hivert, M.-F., Benazzo, A., Rybin, D., Jackson, A.U., Stringham, H.M., et al. (2014). Impact of Type 2 Diabetes Susceptibility Variants on Quantitative Glycemic Traits Reveals Mechanistic Heterogeneity. *Diabetes* *63*, 2158–2171.
 36. Wood, A.R., Jonsson, A., Jackson, A.U., Wang, N., van Leewen, N., Palmer, N.D., Kobes, S., Deelen, J., Boquete-Vilarino, L., Paananen, J., et al. (2017). A Genome-Wide Association Study of IVGTT-Based Measures of First Phase Insulin Secretion Refines the Underlying Physiology of Type 2 Diabetes Variants. *Diabetes* *66*, 2296–2309.
 37. Spracklen, C.N., Horikoshi, M., Kim, Y.J., Lin, K., Bragg, F., Moon, S., Suzuki, K., Tam, C.H.T., Tabara, Y., Kwak, S.-H., et al. (2020). Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature* *582*, 240–245.
 38. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* *526*, 68–74.
 39. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* *10*, e1004383.
 40. Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* *99*, 1245–1260.
 41. Gamazon, E.R., Segrè, A.V., van de Bunt, M., Wen, X., Xi, H.S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E.M., Aguet, F., et al.; GTEx Consortium (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* *50*, 956–967.
 42. Bettella, F., Brown, A.A., Smeland, O.B., Wang, Y., Witoelar, A., Buil Demur, A.A., Thompson, W.K., Zuber, V., Dale, A.M., Djurovic, S., and Andreassen, O.A. (2018). Cross-tissue eQTL enrichment of associations in schizophrenia. *PLoS ONE* *13*, e0202812.
 43. Elbein, S.C., Kern, P.A., Rasouli, N., Yao-Borengasser, A., Sharma, N.K., and Das, S.K. (2011). Global gene expression profiles of subcutaneous adipose and muscle from glucose-tolerant, insulin-sensitive, and insulin-resistant individuals matched for BMI. *Diabetes* *60*, 1019–1029.
 44. Yoshida, S., Tsutsumi, S., Muhlebach, G., Sourbier, C., Lee, M.-J., Lee, S., Vartholomaïou, E., Tatokoro, M., Beebe, K., Miyajima, N., et al. (2013). Molecular chaperone TRAP1 regulates a metabolic switch between mitochondrial respiration and aerobic glycolysis. *Proc. Natl. Acad. Sci.* *110*, E1604–E1612.
 45. Al-Khalili, L., de Castro Barbosa, T., Östling, J., Massart, J., Katayama, M., Nyström, A.C., Oscarsson, J., and Zierath, J.R. (2014). Profiling of human myotubes reveals an intrinsic proteomic signature associated with type 2 diabetes. *Transl. Proteom.* *2*, 25–38.
 46. Chen, X., Ayala, I., Shannon, C., Fourcaudot, M., Acharya, N.K., Jenkinson, C.P., Heikkinen, S., and Norton, L. (2018). The Diabetes Gene and Wnt Pathway Effector TCF7L2 Regulates Adipocyte Development and Function. *Diabetes* *67*, 554–568.
 47. Shao, W., Xiong, X., Ip, W., Xu, F., Song, Z., Zeng, K., Hernandez, M., Liang, T., Weng, J., Gaisano, H., et al. (2015). The expression of dominant negative TCF7L2 in pancreatic beta cells during the embryonic stage causes impaired glucose homeostasis. *Mol. Metab.* *4*, 344–352.
 48. Mitchell, R.K., Mondragon, A., Chen, L., McGinty, J.A., French, P.M., Ferrer, J., Thorens, B., Hodson, D.J., Rutter, G.A., and Da Silva Xavier, G. (2015). Selective disruption of Tcf7l2 in the pancreatic b cell impairs secretory function and lowers b cell mass. *Hum. Mol. Genet.* *24*, 1390–1399.
 49. da Silva Xavier, G., Mondragon, A., Sun, G., Chen, L., McGinty, J.A., French, P.M., and Rutter, G.A. (2012). Abnormal glucose tolerance and insulin secretion in pancreas-specific Tcf7l2-null mice. *Diabetologia* *55*, 2667–2676.
 50. Shu, L., Matveyenko, A.V., Kerr-Conte, J., Cho, J.H., McIntosh, C.H.S., and Maedler, K. (2009). Decreased TCF7L2 protein levels in type 2 diabetes mellitus correlate with downregulation of GIP- and GLP-1 receptors and impaired beta-cell function. *Hum. Mol. Genet.* *18*, 2388–2399.
 51. Ip, W., Shao, W., Song, Z., Chen, Z., Wheeler, M.B., and Jin, T. (2015). Liver-Specific Expression of Dominant-Negative Transcription Factor 7-Like 2 Causes Progressive Impairment in Glucose Homeostasis. *Diabetes* *64*, 1923–1932.
 52. Boj, S.F., van Es, J.H., Huch, M., Li, V.S.W., José, A., Hatzis, P., Mokry, M., Haegebarth, A., van den Born, M., Chambon, P., et al. (2012). Diabetes risk gene and Wnt effector Tcf7l2/TCF4 controls hepatic response to perinatal and adult metabolic demand. *Cell* *151*, 1595–1607.
 53. Suhre, K., McCarthy, M.I., and Schwenk, J.M. (2020). Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.* Published online August 28, 2020. <https://doi.org/10.1038/s41576-020-0268-2>.

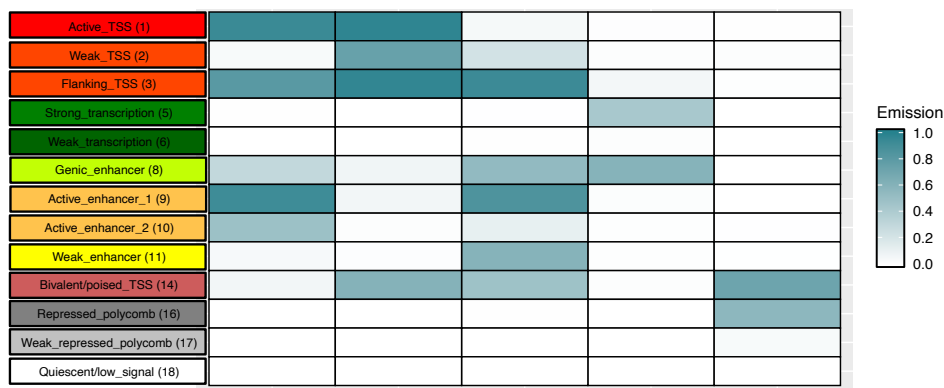
The American Journal of Human Genetics, Volume 107

Supplemental Data

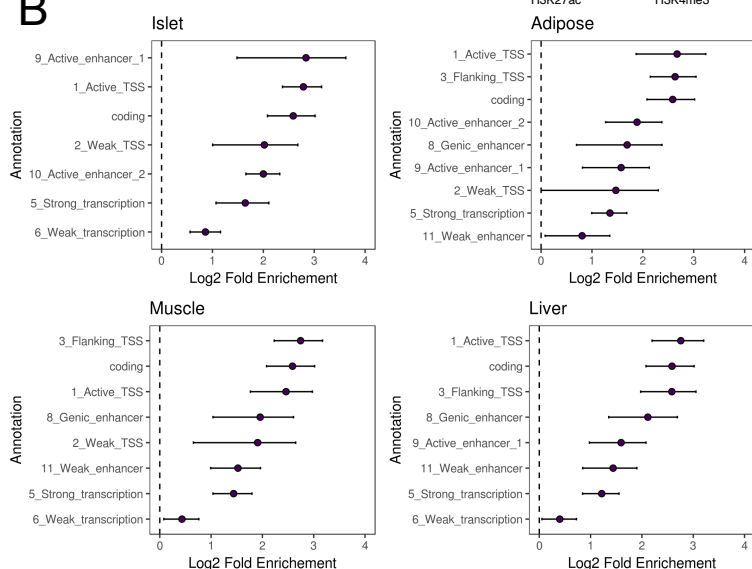
**A Multi-omic Integrative Scheme Characterizes Tissues
of Action at Loci Associated with Type 2 Diabetes**

Jason M. Torres, Moustafa Abdalla, Anthony Payne, Juan Fernandez-Tajes, Matthias Thurner, Vibe Nylander, Anna L. Gloyn, Anubha Mahajan, and Mark I. McCarthy

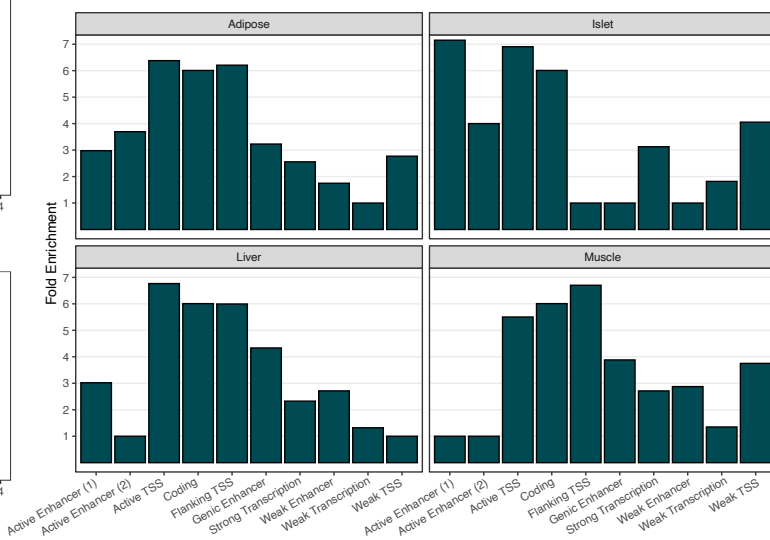
A



B



C



D

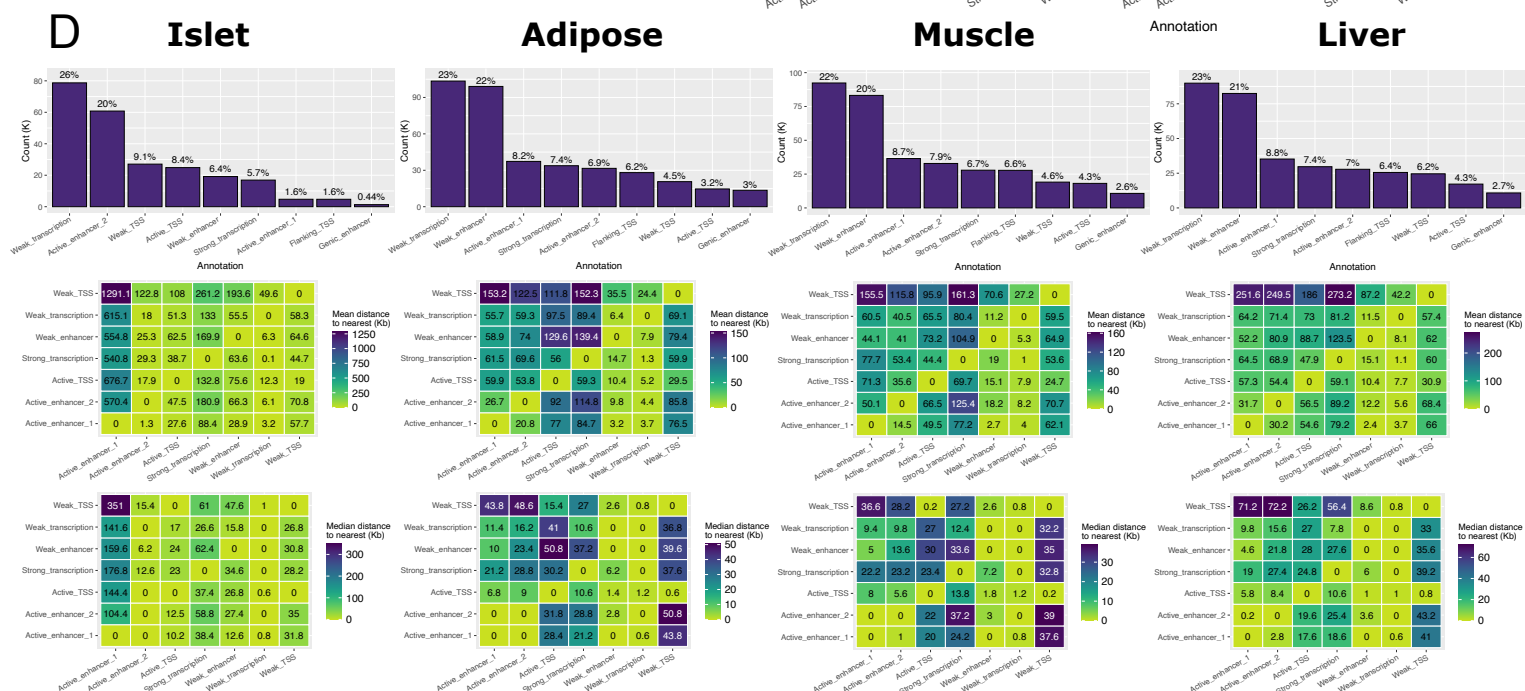


Figure S1. Epigenomic annotations are enriched for variants associated with type 2 diabetes. A)

ChromHMM emission values from Varshney et al. 2017. are indicated for each chromatin state annotation. **B)** Chromatin state annotations were assessed for genome-wide enrichment of T2D-associated SNPs using the program *fgwas*. Log₂ fold enrichment values and 95% confidence bounds are shown for each chromatin state annotation meeting statistical significance. **C)** The mean fold enrichment values for each tissue annotation were used as weights for scaling the relative contribution of each annotation vector in calculating TOA scores. **D)** The count distribution of epigenomic annotations is shown for each tissue along with heat plots showing the pairwise mean and median distances to the nearest annotation in kilobases (Kb).

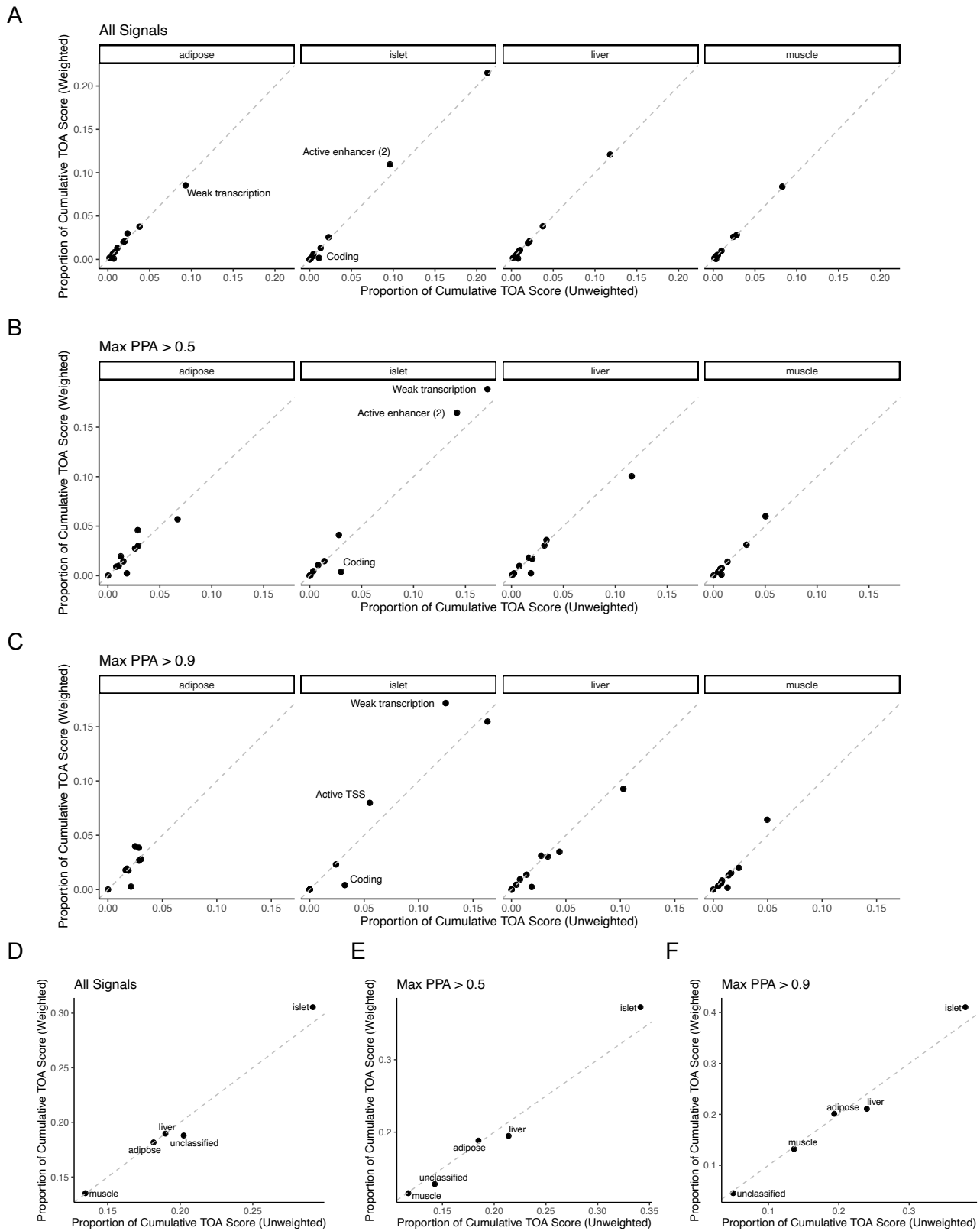


Figure S2. Comparison of GWAS effect size weighted and unweighted TOA scores. TOA scores weighted by $|\beta|/SE$, corresponding to the absolute value of the GWAS effect size divided by standard error of the sentinel SNP corresponding to each signal, are compared with the unweighted TOA scores. The proportion of cumulative score for unweighted (x-axis) and weighted (y-axis) TOA scores are shown for each annotation group in each tissue and is shown for **(A)** all signals, **(B)** all signals with maximum credible set PPA > 0.5, and **(C)** all signals with maximum credible set PPA > 0.9. The three annotation groups with the greatest differences between weighted and unweighted TOA scores, across all tissues, are indicated. The proportions of cumulative TOA scores combined across annotation groups for each tissue are shown for **(D)** all signals, **(E)** all signals with maximum credible set PPA > 0.5, and **(F)** all signals with maximum credible set PPA > 0.9.

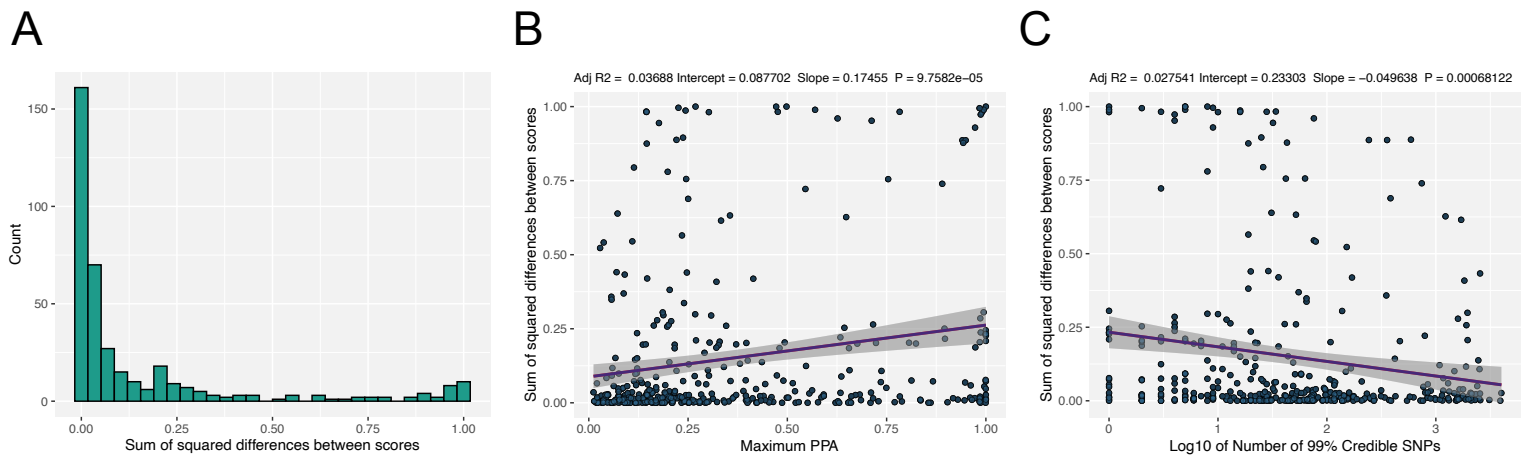


Figure S3. The relationship between credible set resolution and tissue specificity. **A)** Histogram of sum of squared differences (SSD) between tissue-of-action scores for islet, subcutaneous adipose, skeletal muscle, and liver across all 380 fine-mapped signals. **B)** Scatter plot of maximum 99% credible set PPA (x-axis) and SSD values (y-axis) is shown for all signals with regression line. **C)** Scatter plot of base 10 logarithm of number of 99% credible SNPs (x-axis) and SSD values (y-axis) is shown for all signals with regression line.

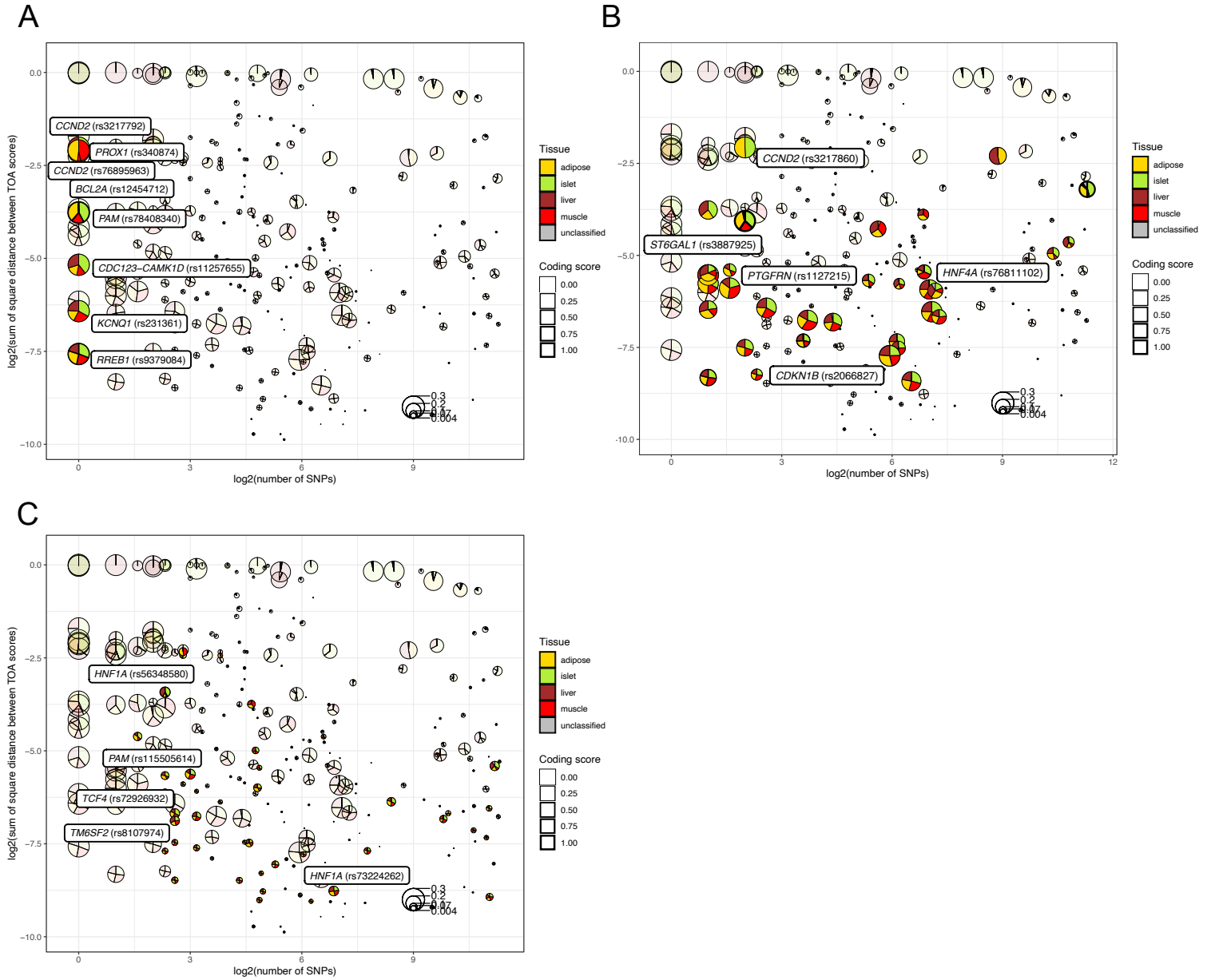


Figure S4. Fine-mapping resolution and TOA score diversity at shared signals. A) Signals that have been fine-mapped (rs6 to a single credible variant. **B)** Signals with more than one credible variant but maximum PPA > 0.5. **C)** Signals with more than one credible variant and maximum PPA \leq 0.5. \log_2 of the number of credible SNPs for each fine-mapped signal is shown on the x-axis and the \log_2 value of the sum of square differences between TOA scores for each signal is shown on the y-axis (i.e. higher values on the y-axis correspond to greater tissue “specificity”). The profile of TOA scores are indicated within pie charts where the diameter of each circle corresponds to the maximum PPA for the credible set. The line thickness for each circle indicates a coding score for each credible set (i.e. the proportion of cumulative PPA attributable to coding variants).

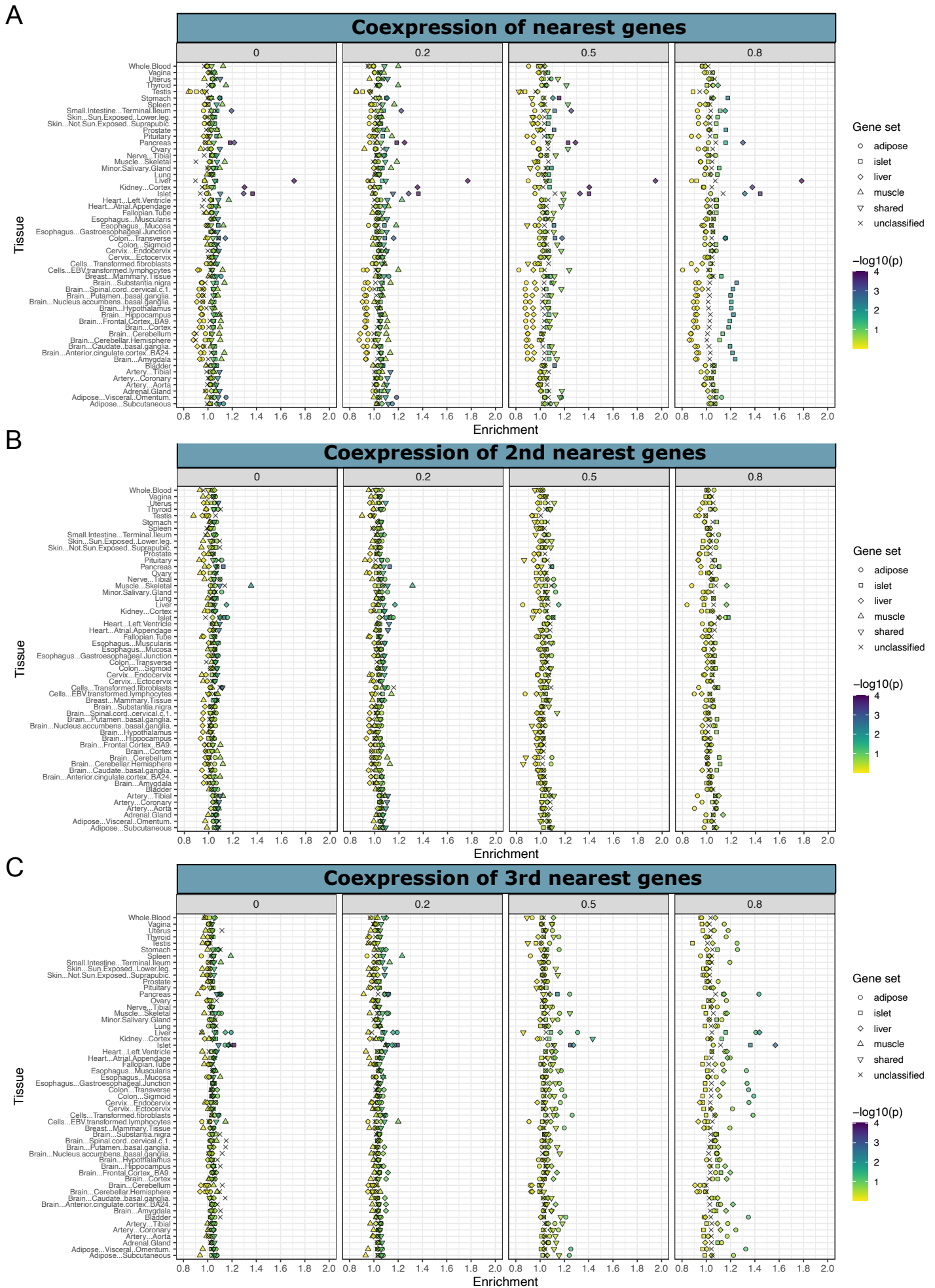


Figure S5. Expression similarity of proximal genes to tissue-assigned signals. A) Significance in similarity of expression levels of nearest genes annotated to sets of tissue-assigned signals across TOA score stringency thresholds. Corresponding values for sets of second nearest and third nearest genes are shown in **B)** and **C)**, respectively. Results are shown for human islet and 53 tissues from the Genotype-Tissue Expression Project (GTEx). Shape indicates the tissue to which the set of signals were assigned.

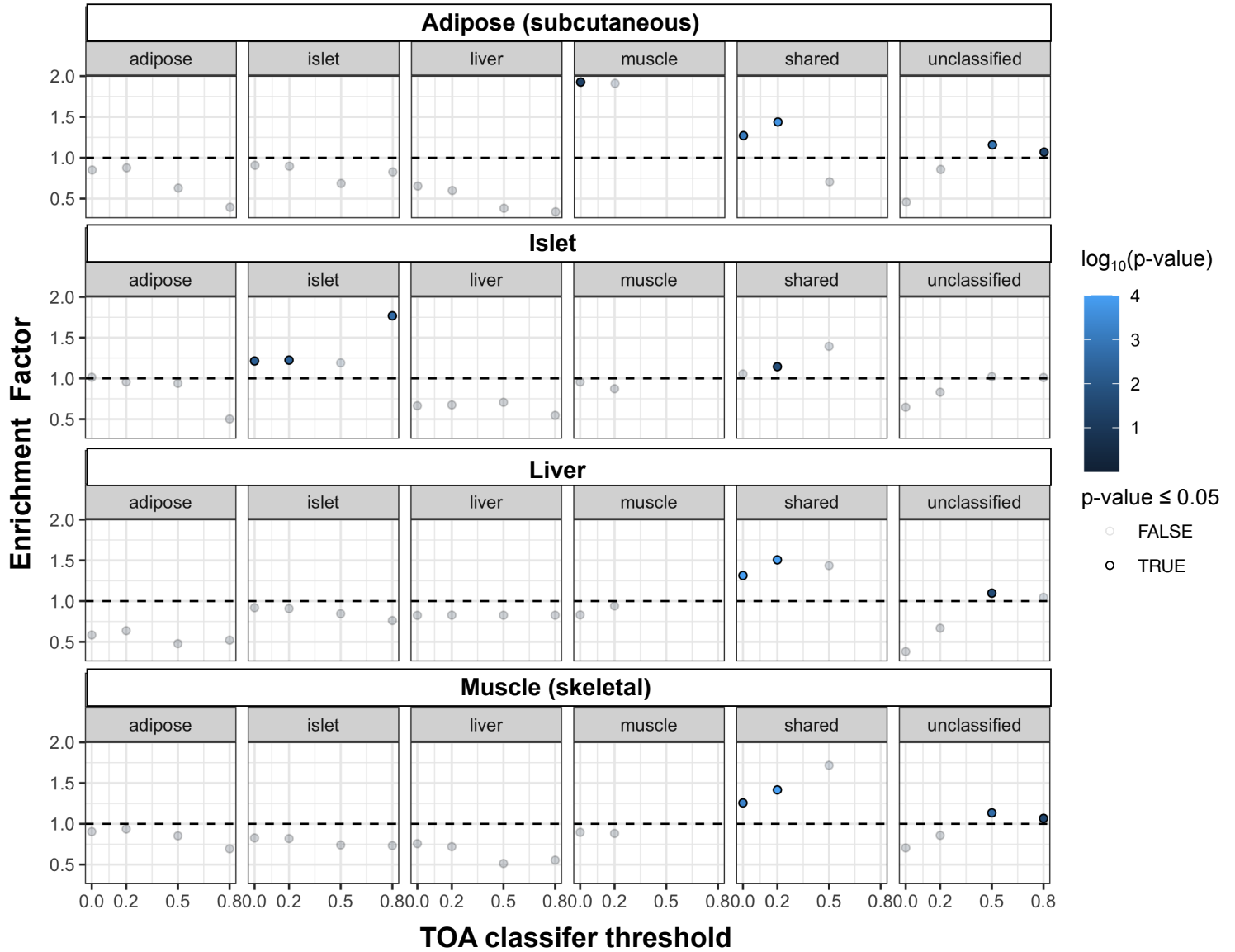


Figure S6. Co-expression of proximal genes to tissue-assigned signals. Enrichment factors from a Spearman correlation-based test of gene co-expression are shown for each set of proximal genes annotated to tissue-assigned signals, at each evaluated stringency threshold. Gene co-expression was evaluated in subcutaneous adipose, islet, liver, and skeletal muscle tissue.

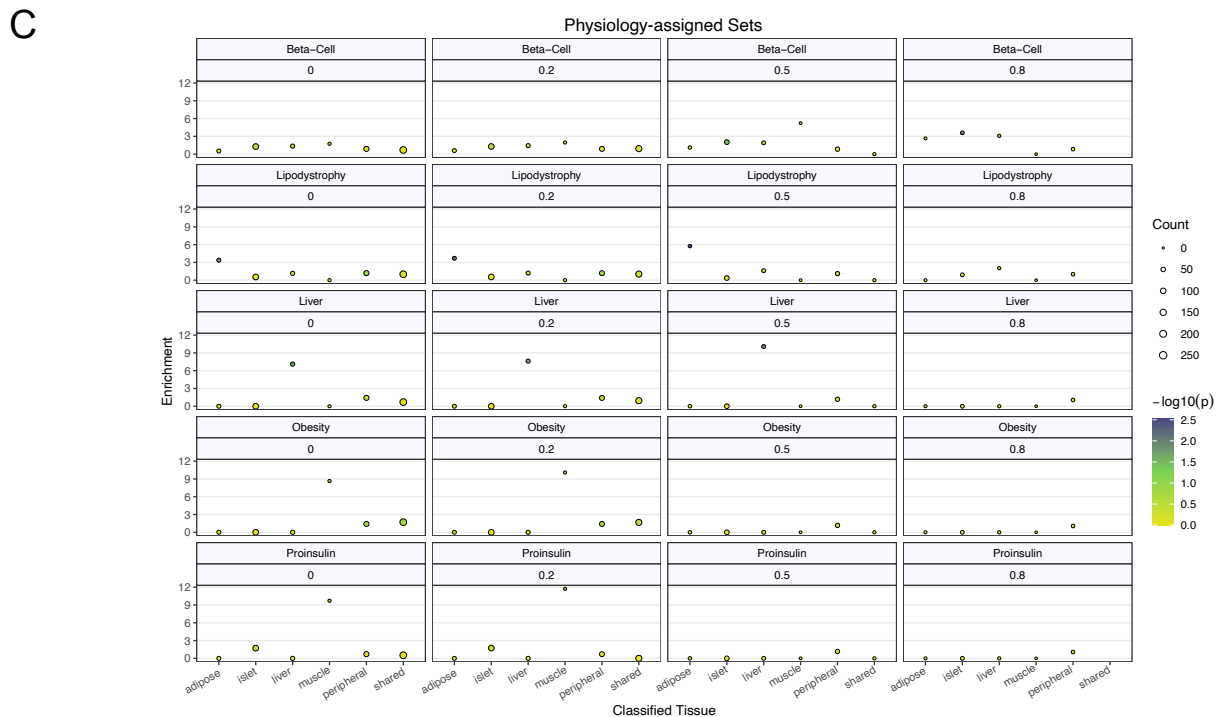
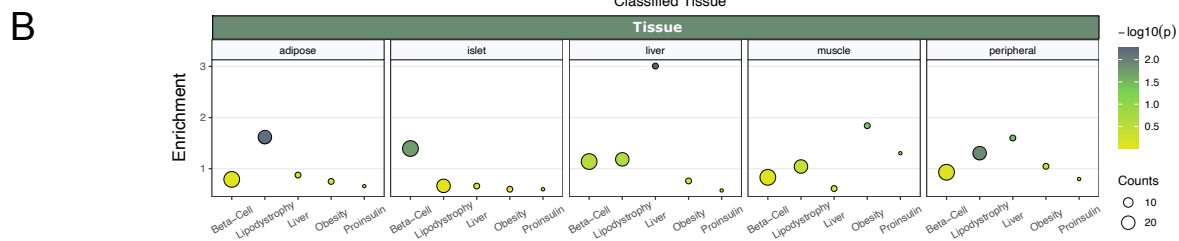
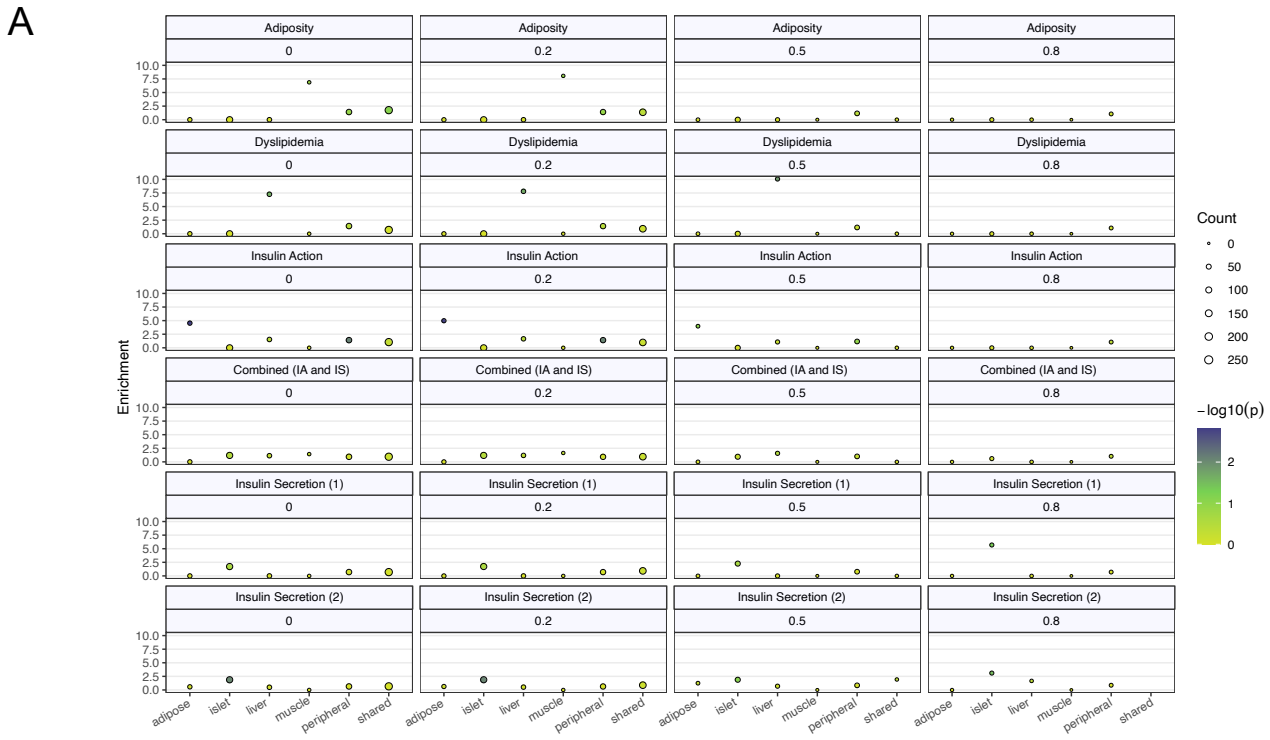
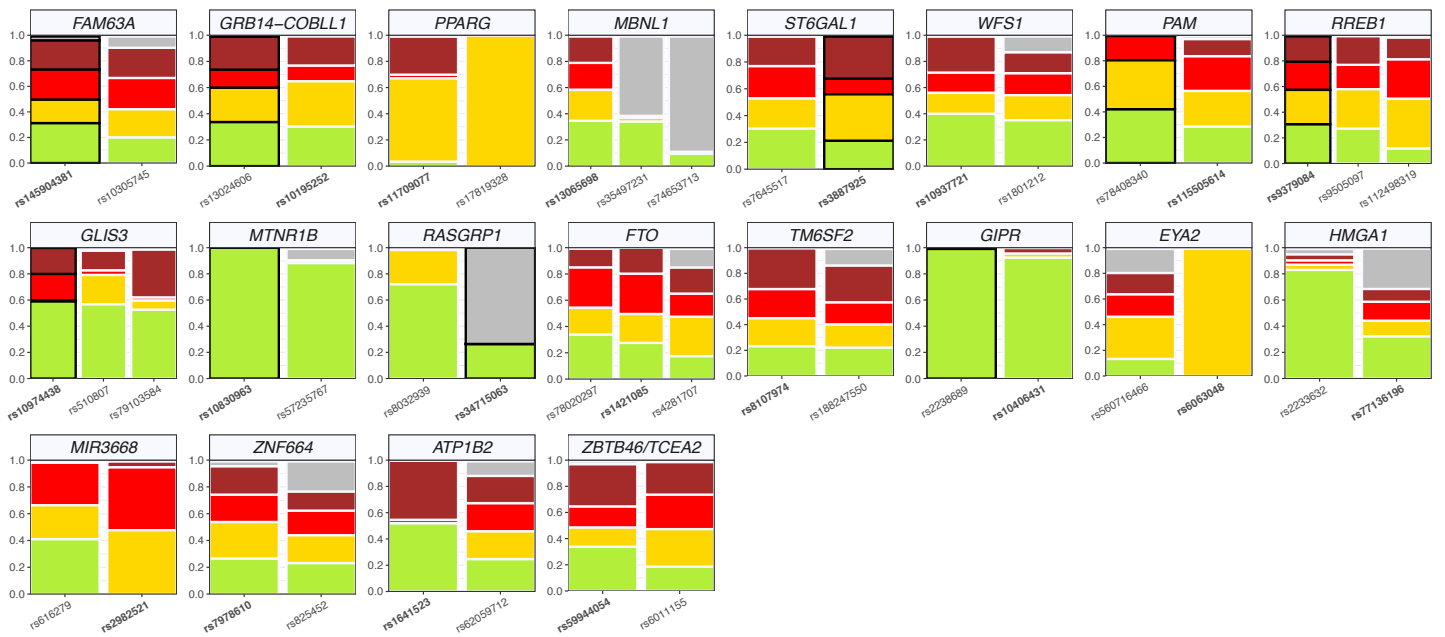


Figure S7. Comparison of sets of epigenomic- and physiology-assigned signals. A) Enrichment of 82 physiology-assigned signals, from six physiology groups, among sets of epigenomic-assigned signals based on TOA scores. Size corresponds to the number of physiology signals among each set of tissue-assigned signals. **B)** Comparison of TOA scores - as enrichments of mean observed value over a sampled mean - among sets of physiology-assigned signals from Udler et al. 2018. The size corresponds to the number assigned signals in each physiology group. **C)** Enrichment of 62 physiology-assigned signals from Udler et al. 2018., from five physiology groups, among sets of epigenomic-assigned signals based on TOA scores. Size corresponds to the number of physiology signals among each set of tissue-assigned signals.



Figure S8. Enrichment of trait-associated SNPs. The results from an enrichment test are shown each trait from the NHGRI-EBI GWAS catalogue where trait associated SNPs were enriched among sets of tissue-assigned signals at a p-value ≤ 0.05 . All results correspond to signals classified at the 0.2 TOA score threshold. Color corresponds to the negative \log_{10} p-value and the text size corresponds to the number of SNPs that overlap lead SNPs (through identity or LD proxy at $r^2 \geq 0.8$) for each tissue-assigned signal.

A



B

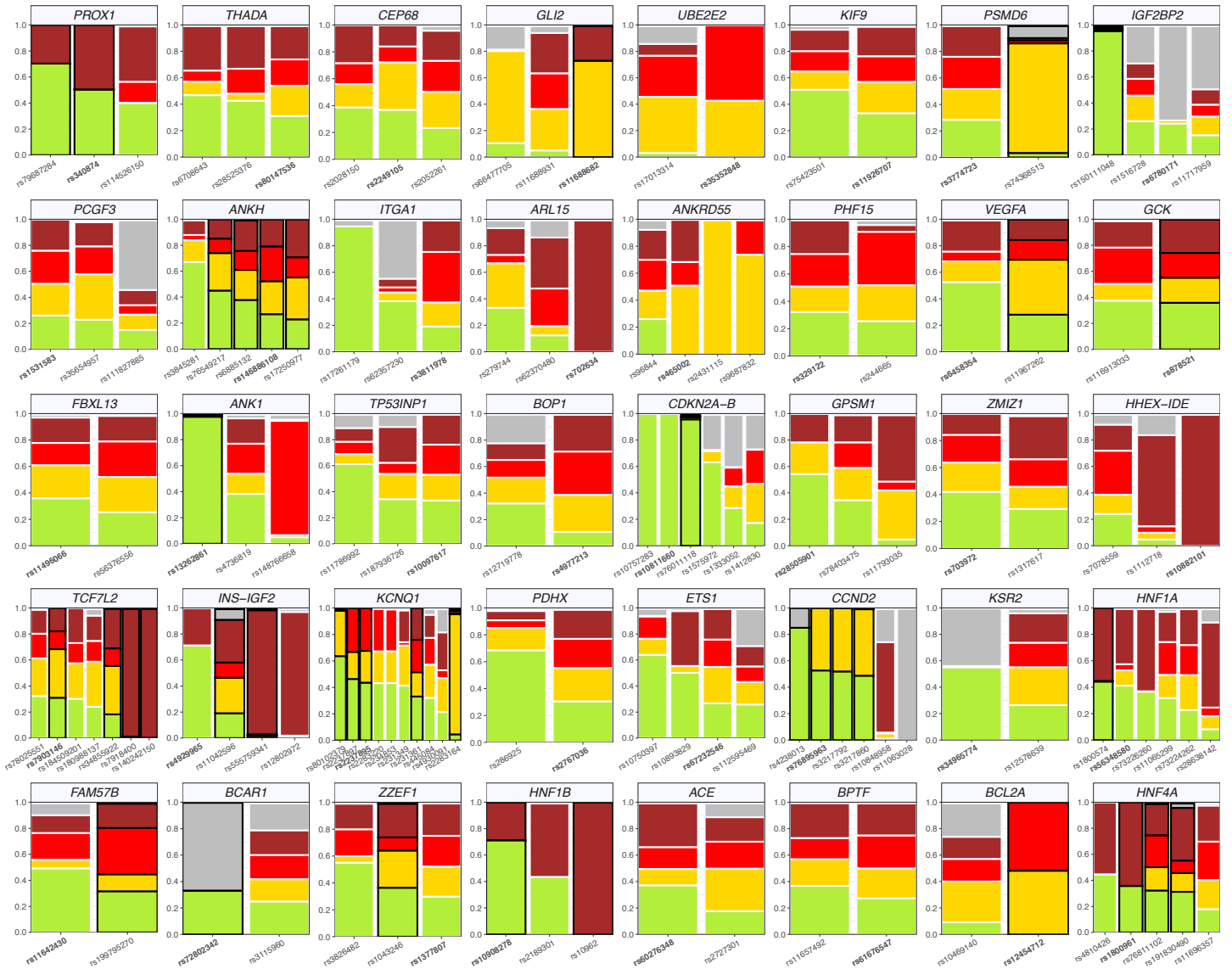


Figure S9. Multiple tissues implicated by epigenomic scores at heterogenous loci. A) Profile of TOA scores for the 20 loci with all signals receiving identical tissue assignments at the 0.2 stringency threshold B) Profile of TOA scores for the 40 loci with all signals receiving distinct tissue assignments at the 0.2 threshold.

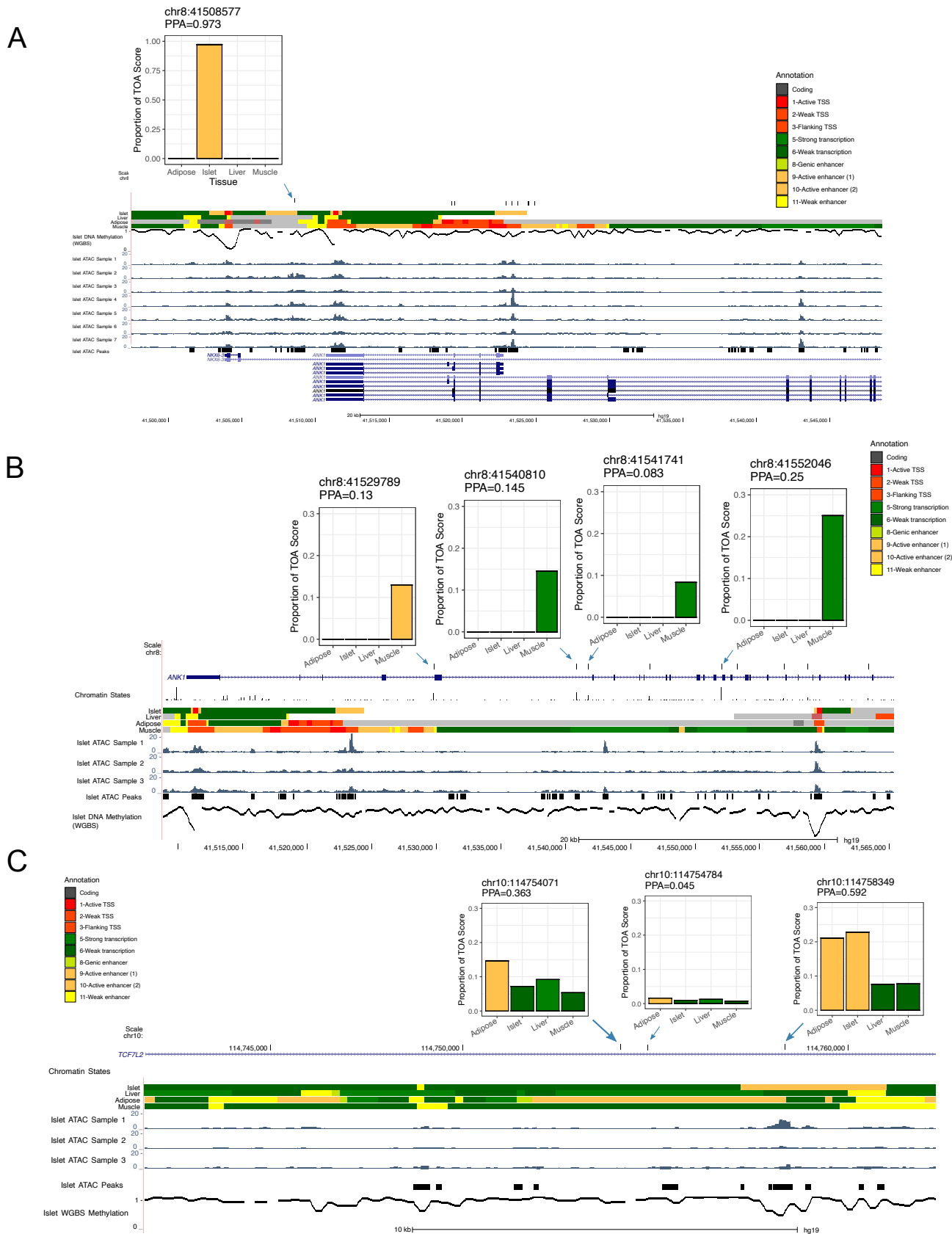


Figure S10. Epigenomic profiles of fine-mapped signals at T2D-associated loci. A) Islet-assigned and primary signal (rs13262861) at the *ANK1* locus. **B)** Muscle-assigned signal at the *ANK1* locus (rs148766658). **C)** Islet-assigned and primary signal (rs7903146) at the *TCF7L2* locus. For each credible SNP, the PPA value attributable to each tissue annotation is shown along with its position on chromosome 10 (genome build hg19). Chromatin state maps for islet, adipose, muscle, and liver tissue from Varshney et al. 2017. are shown along with ATAC-seq tracks for representative islet samples, called ATAC-seq peaks from a set of islet ATAC samples (n=17), and DNA methylation (whole genome bisulfite sequencing) in human islets from Thurner et al. 2018.