Using a modular massively parallel reporter assay to discover context-specific regulatory grammars in type 2 diabetes

Adelaide Tovar[1], Yasuhiro Kyono[1,4], Kirsten Nishino[1], Maya Bose[1], Arushi Varshney[1], Stephen C.J. Parker[1,2,3,5], and Jacob O. Kitzman[1,2,5]

[1]Department of Computational Medicine and Bioinformatics,

[2]Department of Human Genetics,

[3]Department of Biostatistics,

University of Michigan, Ann Arbor, MI 48109,

[4]Current affiliation: University of Illinois at Chicago, Chicago, IL 60607,

[5]Co-corresponding authors.

Address correspondence to: scjp@umich.edu (S.C.J.P.) and kitzmanj@umich.edu (J.O.K.)

Running head: MODULAR MPRA TO DISCOVER REGULATORY GRAMMARS

Keywords: transcription factors, enhancers, GWAS, genetic variation

Abstract

Recent genome-wide association studies have established that most complex disease-associated loci are found in noncoding regions where defining their function is nontrivial. In this study, we leverage a modular massively parallel reporter assay (MPRA) to uncover sequence features linked to context-specific regulatory activity. We screened enhancer activity across a panel of 198-bp fragments spanning over 10k type 2 diabetes- and metabolic trait-associated variants in the 832/13 rat insulinoma cell line, a relevant model of pancreatic beta cells. We explored these fragments' context sensitivity by comparing their activities when placed up- or downstream of a reporter gene, and in combination with either a synthetic housekeeping promoter (SCP1) or a more biologically relevant promoter corresponding to the human insulin gene (*INS*). We identified clear effects of MPRA construct design on measured fragment enhancer activity. Specifically, a subset of fragments (n = 702/11,656) displayed positional bias, evenly distributed across up- and downstream preference. A separate set of fragments exhibited promoter bias (n = 698/11,656), mostly towards the cell-specific *INS* promoter (73.4%). To identify sequence features associated with promoter preference, we used Lasso regression with 562 genomic annotations and discovered that fragments with *INS* promoter-biased activity are enriched for HNF1 motifs. HNF1 family transcription factors are key regulators of glucose metabolism disrupted in maturity onset diabetes of the young (MODY), suggesting genetic convergence between rare coding variants that cause MODY and common T2D-associated regulatory variants. We designed a follow-up MPRA containing HNF1 motif-enriched fragments and observed several instances where deletion or mutation of HNF1 motifs disrupted the *INS* promoter-biased enhancer activity, specifically in the beta cell model but not in a skeletal muscle cell line, another diabetes-relevant cell type. Together, our study suggests that cell-specific regulatory activity is partially influenced by enhancer-promoter compatibility and indicates that careful attention should be paid when designing MPRA libraries to capture context-specific regulatory processes at disease-associated genetic signals.

Introduction

Complex diseases arise from individual and interactive effects of genetic, environmental, and lifestyle factors. Recent genome-wide association studies have revealed that the majority of complex disease-associated loci (~90%) are found in noncoding regions where they are expected to alter gene regulation (Watanabe et al. 2019). Progress to identify causal variants and define their role(s) in disease pathogenesis has been slow due to the lack of a clear variant-to-function framework for noncoding variants. Therefore, recent efforts have focused on defining functional consequences of noncoding variants to describe a unified transcriptional regulatory grammar. Massively parallel reporter assays (MPRAs) have emerged as a useful method for screening large collections of putative regulatory elements, efficiently prioritizing candidate causal variants at disease loci, and learning principles of disease-associated regulatory variation (Khetan et al. 2021; Tewhey et al. 2016; Wang et al. 2018).

A wide array of MPRA designs exist, largely varying based on where fragments of interest are cloned relative to the promoter and reporter gene (summarized in (Klein et al. 2020)). Enhancers are classically defined as acting at a distance in an orientation-independent manner (Banerji et al. 1981), though how distinct they are from promoters remains unclear. In the context of an MPRA construct, Klein et al. reported only a modest impact of fragment position on enhancer activity; however, this study was not conducted with the intent of identifying causal variants at disease-associated loci. Noncoding variants often have subtle effects on expression (Ulirsch et al. 2016; Choi et al. 2020; Tewhey et al. 2016), and we hypothesize that differences in MPRA design could contribute outsized effects, clouding downstream interpretation.

In addition to the fragment position within the MPRA, the other main customizable design component is the promoter. To date, most MPRAs used to study disease-associated genetic variants rely on a standard design featuring one of several housekeeping promoters (e.g., minimal promoter [minP], super core promoter [SCP1]) (Tewhey et al. 2016; McAfee et al.

2022; Melnikov et al. 2012; Patwardhan et al. 2012). There is mixed evidence on the extent to which sequence content determines promoter-enhancer compatibility: some studies suggest that unique sequence features including transcription factor motifs dictate how strongly an interaction will occur (Zabidi et al. 2015; Martinez-Ara et al. 2022) while others indicate that there is minimal specificity in these interactions and that gene regulatory control is dependent on intrinsic enhancer and promoter strength (Hong and Cohen 2022; Bergman et al. 2022). As with studies on fragment position in MPRA constructs, investigations of enhancer-promoter compatibility have been largely separate from studies of allelic effects. We therefore speculated that using a promoter of a physiologically relevant gene may improve our ability to detect disease-associated regulatory perturbations in a tractable and well-matched cell model.

Here, we used a modular MPRA to assess enhancer activity across a panel of 198 bp fragments spanning over 13k type 2 diabetes- and metabolic trait-associated variants. To evaluate their activity across multiple promoter and enhancer contexts, we cloned these fragments up- or downstream of a reporter gene driven by a synthetic housekeeping promoter (SCP1) or the pancreatic beta cell-specific human insulin (*INS*) promoter. We examined enhancer activity bias across this library based on position and linked promoter and discovered a subset of ~500 fragments with greater enhancer activity when paired with the *INS* promoter compared to the SCP1 promoter. Interestingly, these fragments were enriched for pancreas- and neuroendocrine-relevant transcription factor motifs, including the HNF1 family members HNF1A and HNF1B. Using a follow-up MPRA library enriched for HNF1 motifs, we identified several fragments in which HNF1 motifs were necessary for *INS* promoter-biased enhancer activity, specifically in beta cells. Together, our study suggests that enhancer-promoter compatibility contributes to context-specific gene regulation and underscores the importance of thoughtful MPRA design.

Results

Massively parallel reporter design and cloning

To explore regulatory activity of type 2 diabetes (T2D) and related metabolic trait-associated sequences, we designed a library of 79,511 198-bp sequences spanning 13,226 variants (**Figure 1**; **Supplemental Table 1**). In brief, we collected signal index variants for T2D or metabolic trait association signals reported by the DIAMANTE (Mahajan et al. 2022, 2018; Spracklen et al. 2020) and MAGIC (Chen et al. 2021) consortia and performed linkage disequilibrium (LD) expansion to identify variants in high LD ($R^2 > 0.8$). At each associated variant, we selected three pairs of fragment sequences: one with the variant site centered (at position 99 of 198 bp), positioned to the left (at 50 of 198 bp) or to the right (at 149 of 198 bp). We added short 16 bp anchor sequences on both sides of each fragment to enable cloning into the MPRA constructs.

We cloned this fragment library into four different MPRA constructs: upstream or downstream of a portion of the human insulin (*INS*) promoter or the synthetic super core (SCP1) promoter (Juven-Gershon et al. 2006). We simultaneously delivered the resulting plasmid libraries to the 832/13 rat insulinoma cell line via electroporation, then collected cells 24 hours later to isolate RNA. We performed this experiment in triplicate and prepared barcode sequencing libraries from the cDNA and plasmid input to estimate enhancer activity of each fragment.

Regulatory activities measured across fragments

After filtering for fragments represented by greater than two barcodes and with at least 10 input DNA reads, we proceeded with a set of 11,656 fragments for downstream analysis. First, we performed quantification with the R package MPRAnalyze (Ashuach et al. 2019) by comparing the RNA counts for each barcode to DNA counts for the same barcodes. To qualitatively examine the influence of position and promoter on fragment activity, we used principal

component analysis (**Figure 2A**). We identified that the four constructs separate on principal component 2 (13.25 % of variance) based on the position into which fragments were cloned, where both downstream constructs cluster together separately from the two upstream constructs. The two upstream constructs further separate on principal component 1 (15.68% of variance) based on the promoter that was used. We also observed modest differences in activity based on the MPRA construct, where fragments tended to be most active when placed upstream of the promoter, regardless of which promoter was used (**Supplemental Figure 1A**).

Within each of the four different constructs, we calculated the proportion of fragments that were categorized as significantly active by MPRAnalyze (**Supplemental Figure 1B**; FDR < 0.05): 7,724 (of 11,656 total fragments tested, 66.27%) downstream of the *INS* promoter, 5,596 (48.01%) upstream of the *INS* promoter, 6,337 (54.37%) downstream of the SCP1 promoter, and 4,680 (40.15%) downstream of the SCP1 promoter. These proportions are comparable to those calculated when using all fragments without requiring their representation across all four configurations (**Supplemental Table 2**). Of the set of 11,656 fragments, only 860 (7.38%) were significantly active across all four configurations. We also defined sets of fragments that were uniquely active in a single configuration: 1,123 (9.63%) downstream of the *INS* promoter, 526 (4.51%) upstream of the *INS* promoter, 391 (3.35%) downstream of the SCP1 promoter, and 648 (5.56%) upstream of the SCP1 promoter. Taken together, these initial analyses indicate that each MPRA construct influences measured enhancer activity in a unique manner where most fragments were significantly active across only one or two constructs of the four tested.

We systematically quantified effects of each promoter and position on an individual fragment basis using a linear modeling approach in MPRAnalyze. We modeled the activity of each fragment with position and promoter as covariates, then performed a Wald test between each contrast (SCP1 vs. *INS* and upstream vs. downstream) to extract model coefficients. Overall, we identified 698/11,656 fragments (5.99%) with significant promoter choice effects (**Figure 2C**; *FDR* < 0.05) and 703/11,656 fragments (6.03%) with significant

upstream/downstream position effects. Between the two groups, only 74 fragments display both promoter and position effects. While fragments with significant position effects were roughly divided between upstream- and downstream-bias, we observed that 73.35% of fragments with significant promoter effects were biased towards the *INS* promoter (n = 512/698; $p = 4.419 \times 10^{-36}$, binomial test with 50% null expectation). The sequence content of the SCP1 promoter is quite different from the *INS* promoter used here. SCP1 is an 81-bp sequence composed of four highly conserved core promoter motifs (Juven-Gershon et al. 2006): a TATA box, the initiator element (Inr), a motif ten element (MTE), and the downstream promoter element (DPE). SCP1 is known to act as a strong promoter for enhancers in conventional reporter constructs and is the published promoter used for self-transcribing active regulatory region sequencing (STARR-seq) (Arnold et al. 2013). By contrast, we used a 408-bp fragment of the *INS* promoter spanning -364 to +44 bp of the human *INS* gene (German et al. 1995) which contains a battery of pancreas-expressed TF binding sites (*e.g.,* RFX6, NEUROD1, INSM1) (Chandra et al. 2014) and designated glucose-responsive elements (Melloul et al. 2002). Thus, we hypothesized that specific sequence features of the *INS* promoter enable greater regulatory compatibility with a subset of diabetes and related trait associated fragments in this library.

Discovery of sequence features associated with position- or promoter-biased enhancer activity

Next, we were interested in defining specific sequence features associated with promoter- or position-driven activity bias. We used LASSO regression with 10-fold cross validation to select specific features from a set of ~3,000 annotations, including transcription factor motifs, accessible chromatin peaks measured in human pancreatic islets, and ChromHMM-based chromatin state inferences derived from genomic profiling in islets (**Figure 3A**). We used signed Wald statistics from the previous joint analysis as position or promoter bias scores. For each unique fragment sequence, we calculated overlap scores across all annotations which we standardized for use as model features using rank-based inverse

normalization. To account for redundancy in TF motifs, we first performed LASSO using a clustered set of 540 motifs (D'Oliveira Albanus et al. 2021) and compared to results using the full set of annotations.

Only four features showed significant association with promoter bias, with all four enriched in fragments with *INS* promoter-biased activity versus SCP1 (**Figure 3B**). Notably, two of these were motifs for HNF family transcription factors, which play important roles in pancreatic beta cell development, differentiation, and homeostasis, and when disrupted cause maturity onset diabetes of the young (MODY), an early onset monogenic form of diabetes (Nkonge et al. 2020; Zhang et al. 2021). Concordantly, many noncoding variants associated with complex, later-onset diabetes (i.e., T2D) are enriched in HNF1 motifs (Quang et al. 2015). Additionally, the *INS* promoter itself overlaps a putative HNF1 motif; however, there is conflicting evidence as to whether HNF1A/HNF1B directly mediates *INS* transcription (Okita et al. 1999; Párrizas et al. 2001; Emens et al. 1992). Together this regulatory convergence suggests that these two TFs serve as critical regulators of transcriptional activity central to diabetes pathogenesis.

When performing LASSO with the unclustered set of motifs, we additionally found that motifs for NF-κB constituents (RELA/p65, NFKB) and the ubiquitously expressed activator TFs SP1 and SRF were associated with SCP1 promoter-biased fragment enhancer activity (**Supplemental Figure 2A**). SP1 functions by interacting with the transcription factor II D complex ($TF_{II}D$) (Pugh and Tjian 1990), a constituent of the RNA polymerase II preinitiation complex that binds to core sequences found in the SCP1 promoter (Louder et al. 2016). HINFP (also known as MIZF), a transcriptional repressor (Sekimata and Homma 2004; Sekimata et al. 2001), was the strongest coefficient associated with SCP1 promoter-biased activity. In addition to the annotations associated with *INS* promoter-biased activity in the reduced set, we found that several GATA family TFs and the SOX family member SRY motifs were strongly predictive of fragments' bias towards the *INS* promoter, as was the motif for NKX6.1, another key

pancreatic beta cell identity TF. Similar to *HNF1A*/*HNF1B*, coding variants in *NKX6-1* have more recently been accepted as very rare causes of MODY (<1%) (Mohan et al. 2018; Taylor et al. 2013) and T2D-associated variants are also enriched in NKX6.1 motifs (Chiou et al. 2021; Khetan et al. 2018, 2021).

We applied the same approach to define features associated with upstream/downstream position bias (**Figure 3C; Supplemental Figure 2B**). Of the significantly associated features, most (32/46) exhibited bias towards upstream-cloned fragments, which was consistent with the overall higher activity of that reporter configuration (**Supplemental Figure 1A**). Features predictive of upstream bias included overlap with open chromatin peaks in pancreatic islets ("atac_seq"), motifs for the transcriptional activator NF-Y (which binds to CCAAT motifs in promoters) (Dolfini et al. 2012) and several TFs (e.g., CUX1, ISX, POU family members, and PAX family members) known to bind to promoters of target genes important for developmental processes and fate specification (Vadnais et al. 2013). The strongest predictor of upstream-biased activity was presence of the "Active_TSS," an annotation denoting active promoters as defined by ChromHMM (Ernst and Kellis 2012). Conversely, many fewer annotations were associated with downstream-biased enhancer activity, though among these included the "typicalEnhancer" ChromHMM state, which overlaps conventional enhancer marks such as p300 or H3K27ac (Vahedi et al. 2015).

Evaluating effects of HNF1 motif perturbations on regulatory activity

To explore the putative interaction between the HNF1 motif within enhancer regions and the *INS* promoter, we designed and constructed a targeted MPRA library (**Figure 4A**). We selected 47 fragments with significant *INS* promoter-biased activity (FDR < 0.05) containing HNF1 motifs. To generate motif deletion or shuffled fragments, we first identified the most likely motif position using SNP-aware position weight matrix scanning. For motif deletion fragments, we removed the matched sequence and added (*n* motif bp)/2 of surrounding sequence to each

end of the fragment to keep the fragment length constant. For motif shuffled fragments, we performed a dinucleotide shuffle on the matched motif sequence to disrupt the motif while maintaining the same sequence content. In cases where the variant was adjacent to the HNF1 motif (rather than contained within), we synthesized motif deletion and motif shuffled fragments with both alleles. We cloned this set of fragments into an MPRA vector upstream of the *INS* promoter then delivered this library to 832/13 cells (n = 6), as described previously.

Of the 47 fragment sets we synthesized and tested, 28 sets had at least one fragment with significant activity (FDR < 0.05) and were retained for downstream analysis. One such variant was rs1635852, which is found in the first intron of *JAZF1*, a well-characterized transcriptional regulator of the pancreatic beta cell response to metabolic stress (Kobiita et al. 2020). This variant falls in a high-information content position of the HNF1 consensus motif (JASPAR record MA0046.2), where changing from the reference allele (T, risk allele for T2D) to the alternate allele (C, non-risk) is predicted to strengthen the motif (**Figure 4B**). We did not observe a statistically significant allelic difference in activity between the reference and alternate alleles, likely owing to high variance in the alternate allele fragment (Wilcoxon rank-sum test p = 0.31); however, the effect is in the expected direction (Fogarty et al. 2013). We did observe that deleting the HNF1 motif led to a significant decrease in activity (**Figure 4C**; Wilcoxon rank-sum test p = 0.016). Regulatory mechanisms at this variant were previously investigated using reporter and DNA affinity assays (Fogarty et al. 2013). This study reported strong allelic bias with the alternate allele displaying higher enhancer activity than the reference allele, consistent with the direction of effect we observed. Interestingly, this study concluded that regulatory activity at this locus was mediated by differential interaction with another key transcription factor expressed in beta cells, PDX1. Our results point to a possible complementary regulatory mechanism at this locus mediated by HNF1A/B.

We again observed HNF1 motif-dependent activity at rs11819995, a T2D signal variant recently identified in a multi-ancestry GWAS by the DIAMANTE consortium (Mahajan et al.

2022). This variant is located in the first intron of the canonical *ETS1* transcript and overlaps an alternative transcription start site active in pancreatic islets (Varshney et al. 2017; Mahajan et al. 2022). The 14-bp HNF1 motif matched by position weight matrix scanning (*GTTAAACTTTAAAC*) starts 11 bp downstream of the variant in the fragment sequence (**Figure 4D**). Here, we observed that deleting the HNF1 motif significantly decreases the regulatory activity of fragments, in the presence of either the reference (C, non-risk allele for T2D) or alternate allele (T, risk) of this variant (**Figure 4E**; Wilcoxon rank-sum test $p = 0.008$ for each). Additionally, shuffling the HNF1 motif in the reference allele fragment decreased activity in comparison to the original fragment (Wilcoxon rank-sum test $p = 0.008$).

We were also interested in exploring the HNF1 motif-*INS* promoter interaction in light of two related, but distinct questions: (1) does perturbing the HNF1 motif have similar effects when cloned in proximity to a different tissue-specific promoter, and (2) is the interaction between the HNF1 motif and *INS* promoter restricted to a pancreatic beta cell line? To this end, we cloned the same library of original, deleted, or shuffled fragments in one of two new promoter contexts, with either the SCP1 promoter or with the promoter for the skeletal muscle-specific gene *MYBPC2*. We then delivered these two libraries along with the *INS* promoter MPRA library to the LHCN-M2 human skeletal muscle myoblast line (n = 6) after differentiation to myotubes (**Figure 4F**). In this line, *MYBPC2* is abundantly expressed and *INS* is transcriptionally silent.

In the LHCN-M2 cell line, none of the HNF1 motif fragment sets displayed significant enhancer activity ($\log_2$(RNA/DNA) > 0 and FDR < 0.05) when paired with any of the three different promoters. In the case of rs11819995, we observed that all versions of the fragments displayed little to no enhancer in LHCN-M2 cells compared to 832/13 cells, suggesting that these sequences function more strongly as enhancers in beta cells compared to skeletal muscle and that this fragment alone is not sufficient to serve as a promoter (**Figure 4G**). Additionally, we observed that their activity was modulated by promoter context, where the sequences tended to be more active when cloned with the *MYBPC2* promoter compared to the SCP1 or

INS promoter. Finally, we only observed one instance where disrupting the HNF1 motif leads to modulated regulatory activity. Compared to the original fragment, scrambling the HNF1 motif caused a modest increase in enhancer activity when paired with the *INS* promoter (Wilcoxon rank-sum test $p = 0.015$). In all other cases across all three promoter pairings, we no longer observed any dependence upon the HNF1 motif, which is consistent with the lack of HNF1A and HNF1B expression in skeletal muscle. Taken together, this indicates HNF1A/B-orchestrated regulatory activity is beta cell-specific.

Discussion

Massively parallel reporter assays (MPRAs) enable large-scale variant-to-function studies of cis-regulatory elements. However, MPRA studies often differ in key technical aspects including the configuration of the cloning construct, whether the construct is integrated or episomal, and which promoter the construct carries. These differences in sequence and cellular context can influence the activity of the cloned fragments (Klein et al. 2020; Bergman et al. 2022); however, given these screens' scale, it is not routine practice to examine how each of these factors influence individual fragments' activity or how they may modulate the effects of genetic variation.

In this study, we used a modular MPRA construct to explore the effects of sequence context and proximity upon regulatory elements' activity. We found that for a minority of disease-associated genomic sequences, enhancer activity was sensitive to proximity to and identity of the promoter used, independent of overall activity. We found specific genomic annotations that were predictive of these biases and noted that direction of effect was concordant with those interpretations. For instance, fragments annotated as proximal transcription start sites by ChromHMM tended to be more active when cloned upstream, and conversely, those annotated as enhancers were more active when cloned downstream. Of note, we discovered that sequences preferentially active with the human insulin (*INS*) promoter were

enriched in motifs for the important pancreatic beta cell transcription factors including HNF1A/B and NKX6.1. In a follow-up MPRA, we demonstrated the necessity of HNF1A/B motifs for activity of a subset of these fragments and confirmed that the dependence upon those motifs was specific to both the cell type and the paired promoters.

Our observation that some sequences' activity depends upon proximity to a promoter is consistent with many previous reports. Enhancers are known to work at a range of distances, including 1,000s of kb away from their targets. Here, we identified that sequences that were more active downstream (distal to) the promoter and reporter gene tended to be enriched for features of typical enhancers while those that were more active in the upstream position (proximal) contained promoter-like elements including enrichment for active TSSs. It is important to note that the MPRA construct we used in this study to explore positional effects is distinct from the STARR-seq vector in that fragments are cloned downstream of the poly-A signal and thus are not transcribed (Klein et al. 2020). Therefore, we do not suspect that interaction with RNA-binding proteins or accelerated RNA decay influenced the results observed here.

Previous studies using reporter assays have indicated that core promoter sequences display some level of tissue-specificity (Mattioli et al. 2019), where observed activity in an MPRA vector is strongly correlated with endogenous transcriptional activity from the same loci. This has also been observed by others working to design minimal synthetic promoters that can function in a context-dependent manner, including with respect to cell type (Zahm et al. 2023). Together with our work, this suggests that promoter sequences contain a unique syntax that interacts with a cell's regulatory milieu to produce highly tuned gene expression beyond the standard elements common to many promoters (e.g., TATA box, downstream promoter element, etc.). It is important to note that we used a well-delimited promoter for the highly cell type-specific human insulin gene that was previously defined through deletion mapping and

motif analysis (German et al. 1995; Melloul et al. 2002). Selecting suitable promoters for other cell types will require a similar level of careful inspection.

Though not explicitly considered in the work presented here, choice of using an episomal or integrated MPRA vector has also been examined previously (Inoue et al. 2017; Klein et al. 2020; Guzman et al. 2023). Here, we used an integrated construct to compare activity at a specific locus in the LHCN-M2 human skeletal muscle cell line compared to an electroporated episomal construct in the 832/13 rat beta cell line (**Figure 4D-E**), due to technical limitations that make electroporation or transfection of LHCN-M2 cells unfeasible. While we see differences in enhancer activity across the two lines, we believe that this is a biological distinction and not methodological one. However, given the systematic discrepancies measured between the two methods in the studies referenced above, this design choice may also require careful forethought.

The work presented here indicates that selection of a promoter and its proximity to screened elements has a measurable effect on regulatory activity in MPRA constructs. Though it takes careful consideration to select appropriate promoters for a disease of interest, we show that using a promoter for the crucial pancreatic beta cell-specific gene *INS* was sufficient to capture regulatory activity at many relevant genomic regions.

Methods

Cell culture

We obtained INS-1 832/13 rat insulinoma cells from Dr. Christopher Newgard (Sarah W. Stedman Nutrition and Metabolism Center, Duke University, Durham, NC) and LHCN-M2 human skeletal muscle myoblasts from Evercyte. We cultured 832/13 cells in RPMI-1640 containing 10% fetal bovine serum, 0.05 mM β-mercaptoethanol, 1 mM sodium pyruvate, 2 mM L-glutamine, 10 mM HEPES, and 1,000 U/mL penicillin/streptomycin. We maintained LHCN-M2 cells on 0.1% gelatin-coated flasks in 4:1 DMEM (4.5 g/L glucose, glutamine,

bicarbonate):Medium 199 (with bicarbonate) containing 15% FBS, 20 mM HEPES, 30 ng/mL zinc sulfate, 1.4 ug/mL vitamin B12, 55 ng/mL dexamethasone, 2.5 ng/mL recombinant human hepatocyte growth factor, 10 ng/mL basic fibroblast growth factor, and 600 U/mL penicillin/streptomycin.

Large (first) MPRA library design, construction, and delivery

To select regions of interest for the large MPRA library relevant to type 2 diabetes (T2D) and related metabolic traits, we gathered signal variants for T2D reported by the DIAMANTE Consortium (Mahajan et al. 2022) and signal variants for HbA1C, fasting insulin adjusted for BMI, fasting glucose adjusted for BMI, and 2-hr glucose adjusted for BMI reported by the MAGIC Consortium (Chen et al. 2021). Additionally, we collected variants for non-T2D traits including rheumatoid arthritis, systemic lupus erythematosus, and type 1 diabetes. We used UK Biobank imputed data to calculate linkage disequilibrium and fetch proxies with an R^2 greater than 0.8. We removed any variants that were not SNVs (i.e., indels, CNVs) resulting in a set of 13,226 variants. For all variants, we designed 198-bp fragment sequences for three different windows: variant at bp 50 (left), variant at bp 99 (center), and variant at bp 149 (right). In cases where a variant had two alternate alleles, we designed both sets of sequences. We extracted all sequences from the hg19 reference genome. We added 16-bp flanking adapter sequences for PCR amplification and cloning. We obtained a pool of 79,511 full-length 230-bp fragments from Agilent.

We constructed the large MPRA library as described previously (Varshney et al. 2021). We added homology arms via PCR, then cloned amplified fragments into the EcoRV-digested MPRA backbone via HiFi assembly. We transformed this first assembly into electrocompetent 10-beta cells. We post-barcoded by digesting the plasmid library with PmeI then used HiFi assembly to insert 16-bp random barcodes at the PmeI restriction site. We transformed this

reaction into electrocompetent 10-beta cells and prepared final plasmid libraries for transfection using the ZymoPURE Plasmid Maxiprep Kit.

For the first library, we electroporated 50 ug of plasmid into 25 million 832/13 rat insulinoma cells for each biological replicate and lysed cells with TRIzol 24 hours later. After phase separation, we used the Direct-zol RNA Miniprep Kit to isolate total RNA. We enriched for mature mRNA transcripts with fragment(dT) beads then treated 2 ug of mRNA with DNase I to remove plasmid or genomic DNA contamination. We reverse transcribed 1 ug of mRNA using SuperScript III with a custom primer containing a 6-bp unique molecular identifier (UMI) targeting transcripts derived from the MPRA plasmids (**Supplemental Table 3**). We treated cDNA samples with DpnI to eliminate any residual plasmid DNA.

Motif (second) MPRA library design, construction, and delivery

The motif MPRA library was constructed using a different method based on Tewhey, et al. 2016, but the two are functionally the same. We selected fragment sequences from the first library with strong *INS* promoter-biased activity that contained HNF1 motifs. We obtained fragments spanning the same amount of genomic sequence as the previous library, but containing different adapters (IDT). We added barcodes and a promoter cloning scaffold via PCR, then cloned the amplified fragments into a modified pMPRA1 vector (a gift from Tarjei Mikkelsen; Addgene #49349) using Golden Gate assembly with PaqCI. We digested this assembly with SfiI to remove empty backbones then transformed into electrocompetent 10-beta cells. We inserted clonal promoter fragments using Golden Gate assembly and digested with AsiSI to remove promoterless constructs. We transformed this final assembly into electrocompetent 10-beta cells, expanded in 150 mL cultures, and isolated plasmid using the ZymoPURE Plasmid Maxiprep Kit.

For the second library, we electroporated 30 ug of plasmid into 30 million 832/13 cells. We lysed cells 24 hours later with Buffer RLT Plus and isolated total RNA with the Qiagen RNeasy Midi Kit. We reverse transcribed 30 ug of DNAse I-treated total RNA using SuperScript IV. We performed a cDNA amplification PCR and sample indexing with all cDNA made.

For use with the LHCN-M2 skeletal muscle myocyte cell line, we ported the assembled MPRA block (fragment, barcode, promoter, GFP) to a lentiviral transfer vector (a gift from Nadav Ahituv; Addgene #137725) via restriction cloning (Gordon et al. 2020; Inoue et al. 2017). The University of Michigan Viral Vector Core produced infectious lentiviral particles with this transfer vector and third generation lentiviral plasmids in HEK293T cells. Per replicate, we infected 4 x 10^6 LHCN-M2 human skeletal myoblasts with our MPRA library at an MOI of ~10. After infection, we passaged the cells for one week to remove any unincorporated virus or contaminating transfer plasmid, then differentiated the cells for one week. We isolated RNA and gDNA from each replicate using the Qiagen AllPrep DNA/RNA mini kit.

For all samples generated through use of the second set of MPRA libraries, we reverse transcribed RNA into cDNA with a GFP-specific primer containing a 6-bp UMI, then constructed indexed sequencing libraries for both the cDNA and gDNA libraries using Illumina-compatible primers (**Supplemental Table 3**).

Sequencing data collection

Pairing and cDNA/input sequencing libraries generated for the initial set of MPRA studies were sequenced on an Illumina Hiseq while libraries generated for the second set of MPRA studies were sequenced on an Illumina Novaseq 6000. For both sets of studies, we received paired-end 150-bp reads.

To create the barcode-fragment pairing dictionary, we used a custom pipeline. First, we merged the paired-end 150-bp reads using bwa v0.7.17 then extracted barcode and fragment sequences using cutadapt v4.3 (Li and Durbin 2009; Martin 2011). To align the merged fragment reads against our reference FASTA file, we used minimap2 v2.24 (Li 2018). After filtering for mapping quality and sequencing depth, we created a final table with fragment-barcode pairs and removed any duplicate barcodes.

We processed MPRA sequencing data using a custom pipeline that quantifies barcode counts while accounting for possible sequencing errors. We identified reads with the barcode sequence and trimmed constant sequences using cutadapt v4.3. We used UMI-tools v1.1.2 to identify and cluster UMIs, then used starcode-umi to cluster UMI-barcode and perform deduplication (Zorita et al. 2015; Smith et al. 2017). Finally, we used starcode to cluster and count barcodes.

Activity estimation and joint modeling

We used the R package MPRAnalyze to estimate enhancer activity and promoter or position bias (Ashuach et al. 2019). This modeling approach fits two separate models for cDNA and input counts, then relates the two to yield a per-fragment transcription rate. First, we used this method to estimate enhancer activity within each construct type and perform *post hoc* comparisons outside of the analysis framework. Then, to evaluate quantitative effects of promoter and position on fragment activity, we jointly estimated enhancer activity across all four constructs and included promoter and position as covariates.

Owing to the smaller number of fragments included in the final library, we chose to perform enhancer activity and allelic bias testing using DESeq2 with a nested model as described previously (Love et al. 2014; Abell et al. 2021). As previously, we merged cDNA and input barcode counts with our pairing dictionary, requiring an exact match. Prior to model fitting, within

each sample we summed all barcode counts for a given fragment to generate consensus counts. We extracted comparisons between cDNA and input counts to evaluate significant activity differences.

LASSO and elastic net regression

Here we used LASSO regression to model MPRA activity bias as a function of signed z-scores derived from Wald testing. We used a set of annotations and TF motifs described previously (Varshney et al. 2021). Within each fragment we quantified motif occurrences or annotation overlap. To generate overlap scores that could be used in LASSO and compared across motifs, we inverse-normalized the $-\log_{10}$(p-value) for each motif or annotation using R/RNOmni (McCaw et al. 2020). We ran the regression using R/glmnet with default parameters ($\alpha = 1$ corresponding to LASSO regression) and 10-fold cross validation using the z-score as the outcome and motif scores as the predictors (Friedman et al. 2010). glmnet automatically determined the appropriate lambda value by minimizing the mean cross-validated error.

Data and code access

All raw sequencing data associated with this study will be submitted to the NCBI Sequencing Read Archive and corresponding processed data to the Gene Expression Omnibus at the time of submission.

All code used to preprocess and analyze data from the first MPRA library is adapted from a previous publication (Varshney et al. 2021) and is found at https://github.com/ParkerLab/STARR-seq-Analysis-Pipeline. All code used to preprocess and analyze data from the second MPRA library and other auxiliary analyses is deposited at https://github.com/adelaidetovar/modular-t2d-mpra.

Competing Interest Statement

The authors declare they have no conflicts of interest.

Acknowledgements

The authors would like to acknowledge members of the Kitzman and Parker labs (University of Michigan) for their critical feedback. The authors would also like to thank the University of Michigan Viral Vector Core for producing lentivirus and the University of Michigan Advanced Genomics Core for sequencing services.

Funding

The authors acknowledge support from the National Institute of Diabetes and Digestive and Kidney Diseases, grants 1UM1DK126185-01 (S.C.J.P.), R01 DK117960 (S.C.J.P.), Opportunity Pool Funding (A.T., S.C.J.P., J.O.K.), and T32 DK101357 (A.T.); and the Burroughs Wellcome Fund Postdoctoral Diversity Enrichment Fellowship (A.T.).

Contributions

Conceptualization by A.T., Y.K., J.O.K., and S.C.J.P. Fragment design and MPRA library cloning by Y.K., A.T., and K.N. Data processing and analysis by A.T. and Y.K. with contributions from M.B and A.V. A.T. wrote the manuscript with input from and editing by J.O.K. and S.C.J.P.

References

Abell NS, DeGorter MK, Gloudemans M, Greenwald E, Smith KS, He Z, Montgomery SB. 2021. Multiple Causal Variants Underlie Genetic Associations in Humans. *bioRxiv* 2021.05.24.445471. https://www.biorxiv.org/content/10.1101/2021.05.24.445471v2.full (Accessed November 15, 2021).

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**: 1074–1077.

Ashuach T, Fischer DS, Kreimer A, Ahituv N, Theis FJ, Yosef N. 2019. MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol* **20**: 183.

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**: 299–308.

Bergman DT, Jones TR, Liu V, Ray J, Jagoda E, Siraj L, Kang HY, Nasser J, Kane M, Rios A, et al. 2022. Compatibility rules of human enhancer and promoter sequences. *Nature* **607**: 176–184.

Chandra V, Albagli-Curiel O, Hastoy B, Piccand J, Randriamampita C, Vaillant E, Cavé H, Busiah K, Froguel P, Vaxillaire M, et al. 2014. RFX6 regulates insulin secretion by modulating Ca2+ homeostasis in human β cells. *Cell Rep* **9**: 2206–2218.

Chen J, Spracklen CN, Marenne G, Varshney A, Corbin LJ, Luan J, Willems SM, Wu Y, Zhang X, Horikoshi M, et al. 2021. The trans-ancestral genomic architecture of glycemic traits. *Nat Genet* **53**: 840–860.

Chiou J, Zeng C, Cheng Z, Han JY, Schlichting M, Miller M, Mendez R, Huang S, Wang J, Sui Y, et al. 2021. Single-cell chromatin accessibility identifies pancreatic islet cell type- and state-specific regulatory programs of diabetes risk. *Nat Genet* **53**: 455–466.

Choi J, Zhang T, Vu A, Ablain J, Makowski MM, Colli LM, Xu M, Hennessey RC, Yin J, Rothschild H, et al. 2020. Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma. *Nat Commun* **11**: 2718.

Dolfini D, Gatta R, Mantovani R. 2012. NF-Y and the transcriptional activation of CCAAT promoters. *Crit Rev Biochem Mol Biol* **47**: 29–49.

D'Oliveira Albanus R, Kyono Y, Hensley J, Varshney A, Orchard P, Kitzman JO, Parker SCJ. 2021. Chromatin information content landscapes inform transcription factor and DNA interactions. *Nat Commun* **12**: 1307.

Emens LA, Landers DW, Moss LG. 1992. Hepatocyte nuclear factor 1 alpha is expressed in a hamster insulinoma line and transactivates the rat insulin I gene. *Proc Natl Acad Sci U S A* **89**: 7300–7304.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**: 215–216.

Fogarty MP, Panhuis TM, Vadlamudi S, Buchkovich ML, Mohlke KL. 2013. Allele-specific transcriptional activity at type 2 diabetes-associated single nucleotide polymorphisms in regions of pancreatic islet open chromatin at the JAZF1 locus. *Diabetes* **62**: 1756–1762.

Friedman JH, Hastie T, Tibshirani R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**: 1–22.

German M, Ashcroft S, Docherty K, Edlund H, Edlund T, Goodison S, Imura H, Kennedy G, Madsen O, Melloul D. 1995. The insulin gene promoter. A simplified nomenclature. *Diabetes* **44**: 1002–1004.

Gordon MG, Inoue F, Martin B, Schubach M, Agarwal V, Whalen S, Feng S, Zhao J, Ashuach T, Ziffra R, et al. 2020. lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat Protoc* **15**: 2387–2412.

Guzman C, Duttke S, Zhu Y, De Arruda Saldanha C, Downes NL, Benner C, Heinz S. 2023. Combining TSS-MPRA and sensitive TSS profile dissimilarity scoring to study the sequence determinants of transcription initiation. *Nucleic Acids Res*. http://dx.doi.org/10.1093/nar/gkad562.

Hong CKY, Cohen BA. 2022. Genomic environments scale the activities of diverse core promoters. *Genome Res* **32**: 85–96.

Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N, Shendure J. 2017. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res* **27**: 38–52.

Juven-Gershon T, Cheng S, Kadonaga JT. 2006. Rational design of a super core promoter that enhances gene expression. *Nat Methods* **3**: 917–922.

Khetan S, Kales S, Kursawe R, Jillette A, Ulirsch JC, Reilly SK, Ucar D, Tewhey R, Stitzel ML. 2021. Functional characterization of T2D-associated SNP effects on baseline and ER stress-responsive β cell transcriptional activation. *Nat Commun* **12**: 5242.

Khetan S, Kursawe R, Youn A, Lawlor N, Jillette A, Marquez EJ, Ucar D, Stitzel ML. 2018. Type 2 Diabetes-Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets. *Diabetes* **67**: 2466–2477.

Klein JC, Agarwal V, Inoue F, Keith A, Martin B, Kircher M, Ahituv N, Shendure J. 2020. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat Methods* **17**: 1083–1091.

Kobiita A, Godbersen S, Araldi E, Ghoshdastider U, Schmid MW, Spinas G, Moch H, Stoffel M. 2020. The Diabetes Gene JAZF1 Is Essential for the Homeostatic Control of Ribosome Biogenesis and Function in Metabolic Stress. *Cell Rep* **32**: 107846.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Louder RK, He Y, López-Blanco JR, Fang J, Chacón P, Nogales E. 2016. Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. *Nature* **531**: 604–609.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for

RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Mahajan A, Spracklen CN, Zhang W, Ng MCY, Petty LE, Kitajima H, Yu GZ, Rüeger S, Speidel L, Kim YJ, et al. 2022. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat Genet* **54**: 560–572.

Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, Payne AJ, Steinthorsdottir V, Scott RA, Grarup N, et al. 2018. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet* **50**: 1505–1513.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.

Martinez-Ara M, Comoglio F, van Arensbergen J, van Steensel B. 2022. Systematic analysis of intrinsic enhancer-promoter compatibility in the mouse genome. *Mol Cell* **82**: 2519-2531.e6.

Mattioli K, Volders P-J, Gerhardinger C, Lee JC, Maass PG, Melé M, Rinn JL. 2019. High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity. *Genome Res* **29**: 344–355.

McAfee JC, Bell JL, Krupa O, Matoba N, Stein JL, Won H. 2022. Focus on your locus with a massively parallel reporter assay. *J Neurodev Disord* **14**: 50.

McCaw ZR, Lane JM, Saxena R, Redline S, Lin X. 2020. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics* **76**: 1262–1272.

Melloul D, Marshak S, Cerasi E. 2002. Regulation of insulin gene transcription. *Diabetologia* **45**: 309–326.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**: 271–277.

Mohan V, Radha V, Nguyen TT, Stawiski EW, Pahuja KB, Goldstein LD, Tom J, Anjana RM, Kong-Beltran M, Bhangale T, et al. 2018. Comprehensive genomic analysis identifies pathogenic variants in maturity-onset diabetes of the young (MODY) patients in South India. *BMC Med Genet* **19**: 22.

Nkonge KM, Nkonge DK, Nkonge TN. 2020. The epidemiology, molecular pathogenesis, diagnosis, and treatment of maturity-onset diabetes of the young (MODY). *Clin Diabetes Endocrinol* **6**: 20.

Okita K, Yang Q, Yamagata K, Hangenfeldt KA, Miyagawa J, Kajimoto Y, Nakajima H, Namba M, Wollheim CB, Hanafusa T, et al. 1999. Human insulin gene is a target gene of hepatocyte nuclear factor-1alpha (HNF-1alpha) and HNF-1beta. *Biochem Biophys Res Commun* **263**: 566–569.

Párrizas M, Maestro MA, Boj SF, Paniagua A, Casamitjana R, Gomis R, Rivera F, Ferrer J. 2001. Hepatic nuclear factor 1-alpha directs nucleosomal hyperacetylation to its tissue-

specific transcriptional targets. *Mol Cell Biol* **21**: 3234–3243.

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S-I, Cooper GM, et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**: 265–270.

Pugh BF, Tjian R. 1990. Mechanism of transcriptional activation by Sp1: evidence for coactivators. *Cell* **61**: 1187–1197.

Quang DX, Erdos MR, Parker SCJ, Collins FS. 2015. Motif signatures in stretch enhancers are enriched for disease-associated genetic variants. *Epigenetics Chromatin* **8**: 23.

Sekimata M, Homma Y. 2004. Sequence-specific transcriptional repression by an MBD2-interacting zinc finger protein MIZF. *Nucleic Acids Res* **32**: 590–597.

Sekimata M, Takahashi A, Murakami-Sekimata A, Homma Y. 2001. Involvement of a novel zinc finger protein, MIZF, in transcriptional repression by interacting with a methyl-CpG-binding protein, MBD2. *J Biol Chem* **276**: 42632–42638.

Smith T, Heger A, Sudbery I. 2017. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**: 491–499.

Spracklen CN, Horikoshi M, Kim YJ, Lin K, Bragg F, Moon S, Suzuki K, Tam CHT, Tabara Y, Kwak S-H, et al. 2020. Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature* **582**: 240–245.

Taylor BL, Liu F-F, Sander M. 2013. Nkx6.1 is essential for maintaining the functional state of pancreatic beta cells. *Cell Rep* **4**: 1262–1275.

Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS, Lander ES, Schaffner SF, et al. 2016. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**: 1519–1529.

Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, Melnikov A, McDonel P, Do R, Mikkelsen TS, et al. 2016. Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**: 1530–1545.

Vadnais C, Awan AA, Harada R, Clermont P-L, Leduy L, Bérubé G, Nepveu A. 2013. Long-range transcriptional regulation by the p110 CUX1 homeodomain protein on the ENCODE array. *BMC Genomics* **14**: 258.

Vahedi G, Kanno Y, Furumoto Y, Jiang K, Parker SCJ, Erdos MR, Davis SR, Roychoudhuri R, Restifo NP, Gadina M, et al. 2015. Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature* **520**: 558–562.

Varshney A, Kyono Y, Elangovan VR, Wang C, Erdos MR, Narisu N, Albanus RD, Orchard P, Stitzel ML, Collins FS, et al. 2021. A transcription start site map in human pancreatic islets reveals functional regulatory signatures. *Diabetes*. http://dx.doi.org/10.2337/db20-1087.

Varshney A, Scott LJ, Welch RP, Erdos MR, Chines PS, Narisu N, Albanus RD, Orchard P, Wolford BN, Kursawe R, et al. 2017. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc Natl Acad Sci U S A* **114**: 2301–2306.

Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, Claussnitzer M, Kellis M. 2018. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat Commun* **9**: 5380.

Watanabe K, Stringer S, Frei O, Umićević Mirkov M, de Leeuw C, Polderman TJC, van der Sluis S, Andreassen OA, Neale BM, Posthuma D. 2019. A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet* **51**: 1339–1348.

Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, Stark A. 2015. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**: 556–559.

Zahm AM, Owens WS, Himes SR, Rondem KE, Fallon BS, Gormick AN, Bloom JS, Kosuri S, Chan H, English JG. 2023. Discovery and Validation of Context-Dependent Synthetic Mammalian Promoters. *bioRxiv* 2023.05.11.539703. https://www.biorxiv.org/content/10.1101/2023.05.11.539703v1 (Accessed September 13, 2023).

Zhang H, Colclough K, Gloyn AL, Pollin TI. 2021. Monogenic diabetes: a gateway to precision medicine in diabetes. *J Clin Invest* **131**. http://dx.doi.org/10.1172/JCI142244.

Zorita E, Cuscó P, Filion GJ. 2015. Starcode: sequence clustering based on all-pairs search. *Bioinformatics* **31**: 1913–1919.

Figures

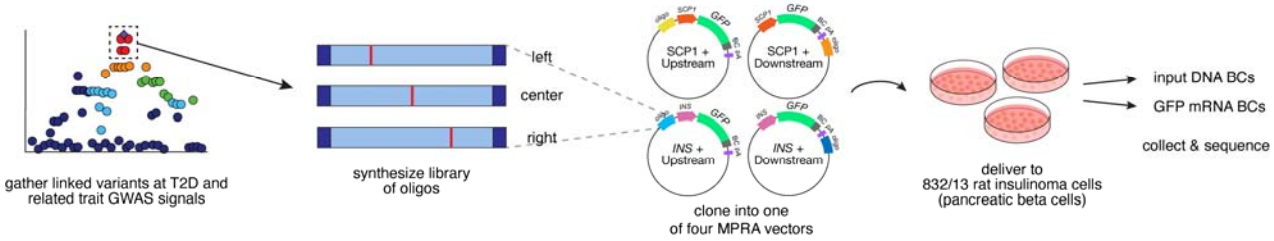

Figure 1. Study design. We synthesized a library of 79,511 198-bp fragments encompassing 13,226 genetic variants (6,177 associated with type 2 diabetes). We designed fragments such that the variant appeared in one of three positions within the fragment sequence at 50 bp, 99 bp or 149 bp. We cloned this library of fragments into each of four MPRA constructs: fragment upstream or downstream of the human insulin (*INS*) promoter or upstream or downstream of the super core promoter (SCP1). We simultaneously delivered all four libraries to 832/13 rat insulinoma cells (n = 3) and collected DNA and RNA for sequencing.

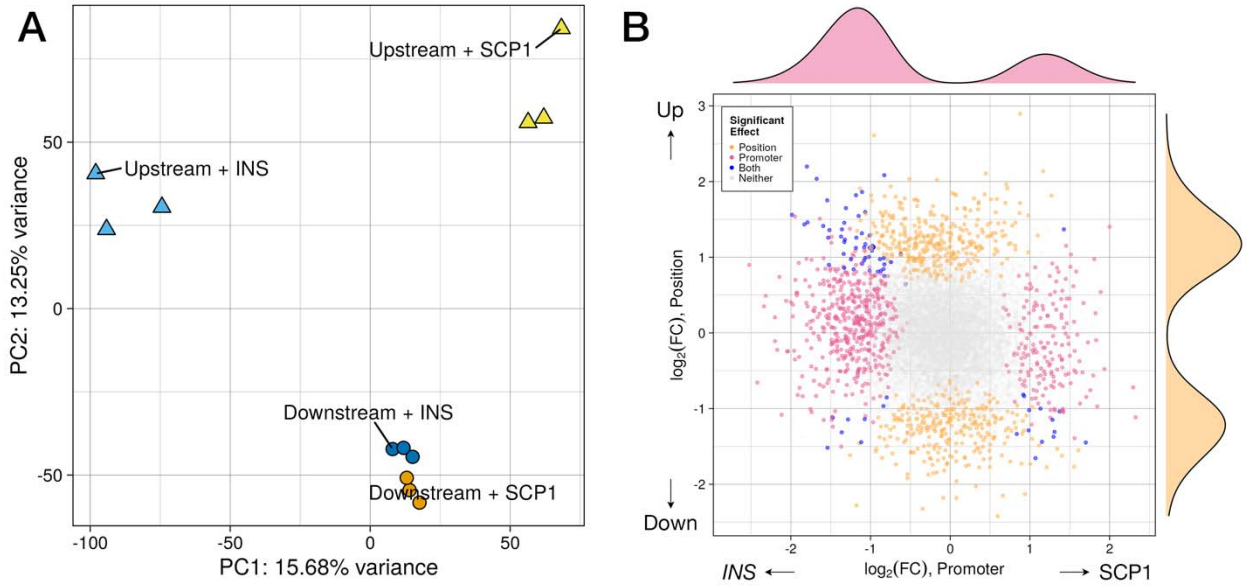

Figure 2. Choice of promoter and fragment position relative to the promoter and reporter gene influence measured MPRA activity. (A) After fragment activity quantification in R/MPRAnalyze, we performed principal components analysis to estimate broad influence of promoter (captured primarily by PC1, x-axis) and position (captured by PC2, y-axis) on MPRA activity. Each point represents an individual replicate, points are color-coded based on plasmid configuration, and point shape indicates position of the cloned fragment. (B) To jointly estimate effects of promoter and position across fragments in all four configurations, we included promoter and position as covariates in the MPRAnalyze model. This scatterplot displays the coefficients from Wald testing for the promoter (pink points along x-axis, *INS* promoter vs. SCP1 promoter with 5% local FDR) and position (orange points along y-axis, upstream versus downstream with 5% local FDR). PC: principal component, FC: fold-change

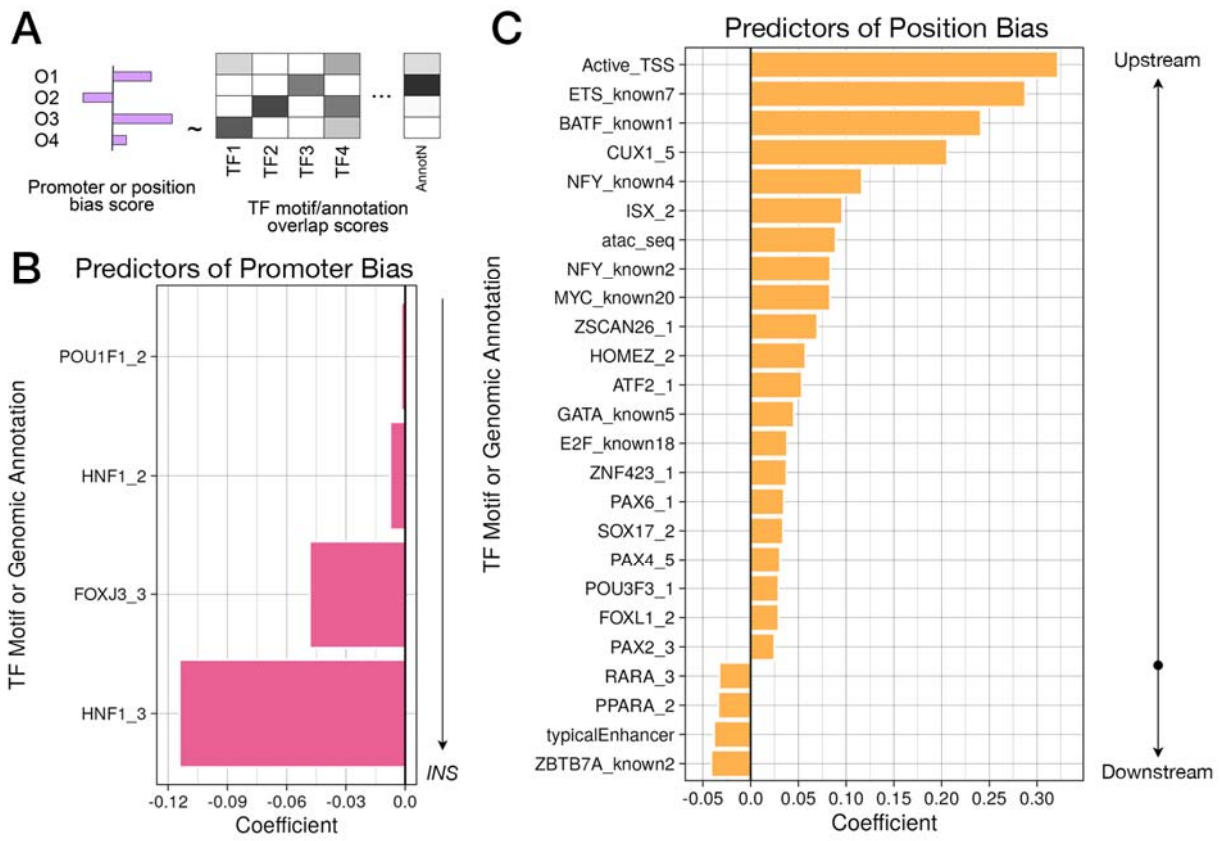

Figure 3. Position and promoter bias in enhancer activity is partially attributed to presence of relevant TF motifs and genomic annotations. (A) To select sequence features associated with position or promoter bias, we performed LASSO regression with 10-fold cross validation on position or promoter bias scores (signed Wald statistics) as a function of several thousand transcription factor motifs and genomic annotations. **(B)** Significant predictors of promoter bias are displayed. With the clustered set of annotations, we identified only predictors enriched in fragments that are more active with the *INS* promoter. **(C)** Top 25 predictors (i.e., nonzero coefficients) associated with position bias effects are shown. Positive coefficients indicate annotations that are enriched in fragments that are more active in the upstream position while negative coefficients are enriched in fragments that are more active in the downstream position.

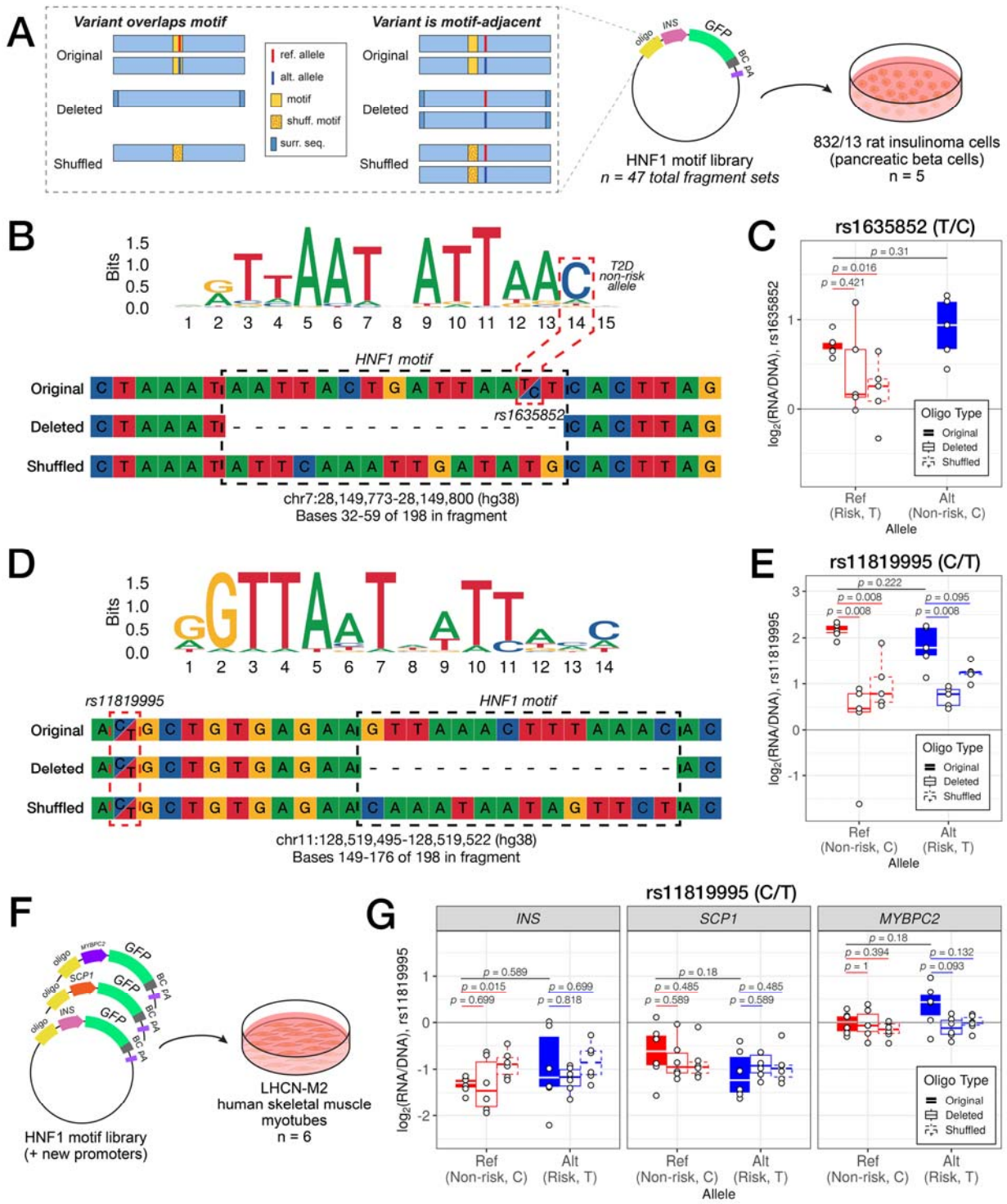
A
Variant overlaps motif
Variant is motif-adjacent
Original
Deleted
Shuffled
ref. allele
alt. allele
motif
shuff. motif
surr. seq.
oligo
INS
GFP
BC
pA
HNF1 motif library
n = 47 total fragment sets
832/13 rat insulinoma cells
(pancreatic beta cells)
n = 5
B
Bits
T2D non-risk allele
HNF1 motif
rs1635852
chr7:28,149,773-28,149,800 (hg38)
Bases 32-59 of 198 in fragment
C
rs1635852 (T/C)
log2(RNA/DNA), rs1635852
Oligo Type
Original
Deleted
Shuffled
Ref (Risk, T)
Alt (Non-risk, C)
Allele
D
rs11819995
HNF1 motif
chr11:128,519,495-128,519,522 (hg38)
Bases 149-176 of 198 in fragment
E
rs11819995 (C/T)
log2(RNA/DNA), rs11819995
Ref (Non-risk, C)
Alt (Risk, T)
F
MYBPC2
SCP1
HNF1 motif library
(+ new promoters)
LHCN-M2
human skeletal muscle
myotubes
n = 6
G
rs11819995 (C/T)
INS
SCP1
MYBPC2

Figure 4. HNF1 transcription factor motifs contribute to enhancer activity near selected T2D-associated variants. (A) For 47 fragments with significant *INS* promoter bias effects on activity that overlapped HNF1 motifs, we designed alternate fragment sequences: original, where the HNF1 motif is left intact; deleted, where the entire HNF1 motif was removed from the sequence and surrounding sequence was added to make up for the loss in total sequence length; and shuffled, where we performed a dinucleotide shuffle on the motif sequence. In cases where the tested variant was adjacent to the matched HNF1 motif, we synthesized two versions for each type to test both reference and alternate alleles. For fragments containing a variant that overlapped the matched HNF1 motif, we only synthesized one deletion fragment and one shuffled fragment. **(B)** We synthesized four fragments (two original, one motif-deleted, one motif-shuffled) corresponding to the variant rs1635852, which overlaps an HNF1 motif at a high information content position. The T2D risk allele (T) disrupts this motif whereas the non-risk allele (C) is the consensus base. **(C)** Deleting and shuffling the HNF1 motif overlapping the variant rs1635852 causes significantly decreased enhancer activity (Wilcoxon rank-sum test p = 0.016). **(D)** For the variant rs11819995, we synthesized six fragments (two original, two motif-deleted, two motif-shuffled). The HNF1 motif begins 11 bp downstream of this variant. **(E)** Deletion of the HNF1 motif near rs11819995 causes significant decreases in enhancer activity for both the reference (C, non-risk) and alternate (T, risk) allele (Wilcoxon rank-sum test p = 0.008). Shuffling the motif causes significant decline in enhancer activity for the reference allele (Wilcoxon rank-sum test p = 0.008). **(F)** To assess context-specific effects, we cloned the same set of fragments into MPRA vectors with the SCP1 or *MYBPC2* (skeletal muscle-specific gene) promoter and delivered all three libraries to LHCN-M2 human skeletal muscle myotubes (n = 6). **(G)** When paired with the *INS* promoter, the rs11819995-containing fragment where the HNF1 motif is shuffled displays a significant increase in activity compared to the original fragment (Wilcoxon rank-sum test p = 0.015); however, all fragments containing rs11819995 are not

acting as enhancers in the LHCN-M2 cell line regardless of which promoter they are paired with. Overall, their activity is highest when paired with the skeletal muscle-specific promoter.