

# Functional characterization of T2D-associated SNP effects on baseline and ER stress-responsive $\beta$ cell transcriptional activation

Shubham Khetan<sup>1,2</sup>, Susan Kales <sup>3</sup>, Romy Kursawe<sup>1</sup>, Alexandria Jillette<sup>1</sup>, Jacob C. Ulirsch <sup>4,5</sup>, Steven K. Reilly <sup>5</sup>, Duygu Ucar <sup>1,2,6</sup>, Ryan Tewhey <sup>3,7,8</sup>  & Michael L. Stitzel <sup>1,2,6</sup> 

Genome-wide association studies (GWAS) have linked single nucleotide polymorphisms (SNPs) at >250 loci in the human genome to type 2 diabetes (T2D) risk. For each locus, identifying the functional variant(s) among multiple SNPs in high linkage disequilibrium is critical to understand molecular mechanisms underlying T2D genetic risk. Using massively parallel reporter assays (MPRA), we test the *cis*-regulatory effects of SNPs associated with T2D and altered *in vivo* islet chromatin accessibility in MIN6  $\beta$  cells under steady state and pathophysiologic endoplasmic reticulum (ER) stress conditions. We identify 1,982/6,621 (29.9%) SNP-containing elements that activate transcription in MIN6 and 879 SNP alleles that modulate MPRA activity. Multiple T2D-associated SNPs alter the activity of short interspersed nuclear element (SINE)-containing elements that are strongly induced by ER stress. We identify 220 functional variants at 104 T2D association signals, narrowing 54 signals to a single candidate SNP. Together, this study identifies elements driving  $\beta$  cell steady state and ER stress-responsive transcriptional activation, nominates causal T2D SNPs, and uncovers potential roles for repetitive elements in  $\beta$  cell transcriptional stress response and T2D genetics.

<sup>1</sup>The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. <sup>2</sup>Department of Genetics and Genome Sciences, University of Connecticut, Farmington, CT, USA. <sup>3</sup>The Jackson Laboratory for Mammalian Genetics, Bar Harbor, ME, USA. <sup>4</sup>Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA. <sup>5</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>6</sup>Institute of Systems Genomics, University of Connecticut, Farmington, CT, USA. <sup>7</sup>Graduate School of Biomedical Sciences and Engineering, University of Maine, Orono, ME, USA. <sup>8</sup>Tufts University School of Medicine, Boston, MA, USA. email: [ryan.tewhey@jax.org](mailto:ryan.tewhey@jax.org); [michael.stitzel@jax.org](mailto:michael.stitzel@jax.org)

Type 2 diabetes (T2D) is a complex disease with both genetic and environmental risk factors that ultimately manifests when pancreatic  $\beta$  cells are unable to secrete adequate amounts of insulin in response to elevated blood glucose levels<sup>1,2</sup>. Genome-wide association studies (GWAS) have identified single nucleotide polymorphisms (SNPs) representing 403 association signals in 243 regions of the human genome (loci) for genetic risk of developing T2D<sup>3,4</sup>. The overwhelming majority (~90%) of these T2D-associated GWAS SNPs are non-coding, suggesting that altered transcriptional regulation is a common molecular mechanism underlying disease risk in these loci<sup>3–6</sup>. Identifying the functional variant(s) among multiple SNPs that are in linkage disequilibrium (LD) at each T2D GWAS locus is an important step to convert these statistical association signals into molecular and biological insights.

Although studies over the past several years clearly implicate altered islet *cis*-regulatory element (CRE) activity in T2D genetic risk and progression, they have co-localized only ~1/4 of T2D-associated loci to altered chromatin accessibility and/or gene expression levels in islets<sup>5,7–10</sup>. This may be partially because previous studies measured the effect of genetic variants on chromatin accessibility (caQTLs) and gene expression levels (eQTLs) in islets under steady-state conditions<sup>7,11</sup>, consequently missing the role of genetic variants whose functions emerge only under certain cellular conditions. Uncovering such genotype-environment interactions is critical for a complex disorder like T2D.

Endoplasmic reticulum (ER) and the unfolded protein response (UPR) contribute to physiologic processes governing  $\beta$  cell protein quality control, insulin processing/secretion, and to pathophysiologic events contributing to islet failure in T2D<sup>12–15</sup>. Mild to moderate ER stress can elicit beneficial responses, such as  $\beta$  cell proliferation, to meet higher demand for insulin synthesis and secretion<sup>16</sup>. However, sustained insulin production demands of insulin resistance, associated with modern sedentary lifestyles<sup>17</sup> and overnutrition<sup>18–20</sup>, can intensify ER stress and activate terminal UPR, leading to  $\beta$  cell dysfunction and death<sup>13,14</sup>. Genetic modulation of  $\beta$  cell ER folding capacity can ameliorate<sup>21</sup> or exacerbate<sup>22</sup>  $\beta$  cell death. Non-coding T2D risk alleles may therefore modulate the transcription of genes and pathways that alter ER stress responses and the UPR.

Massively parallel reporter assays (MPRA) are functional genomic tools to interrogate the transcription activating potential of thousands of sequences simultaneously<sup>23</sup>. By introducing nucleotide changes in a given sequence of interest, the effect of naturally occurring variants in the human population on MPRA activity can also be elucidated<sup>24</sup>. Recent studies have employed MPRA to identify functional SNPs associated with different conditions including red blood cell traits<sup>25</sup>, adiposity<sup>26</sup>, osteoarthritis<sup>27</sup>, and eQTLs<sup>24</sup>.

Here, we use MPRA to comprehensively test 2512 index and genetically linked ( $r^2 \geq 0.8$ ) SNP/indel alleles representing 259 GWAS association signals for T2D and related quantitative trait from the NHGRI/EBI GWAS Catalogue, as well as 4124 SNPs residing in *in vivo* accessible chromatin sites in human islets, for their ability to modulate transcriptional activation in  $\beta$  cells under steady state and (patho)physiologic ER stress conditions. We identify 1982/6621 (29.9%) SNP-containing elements that activate transcription in MIN6 and 879 MPRA activity-modulating alleles. Multiple T2D-associated SNPs alter the activity of short interspersed nuclear element (SINE)-containing elements induced by ER stress. Importantly, MPRA uncovers 220 functional variants at 104 T2D association signals, narrowing 54 signals to a single candidate SNP. By identifying elements driving  $\beta$  cell steady state and ER stress-responsive transcriptional activation, we nominate putative causal T2D SNPs and uncover potential roles for repetitive elements in  $\beta$  cell transcriptional stress response and T2D genetics.

## Results

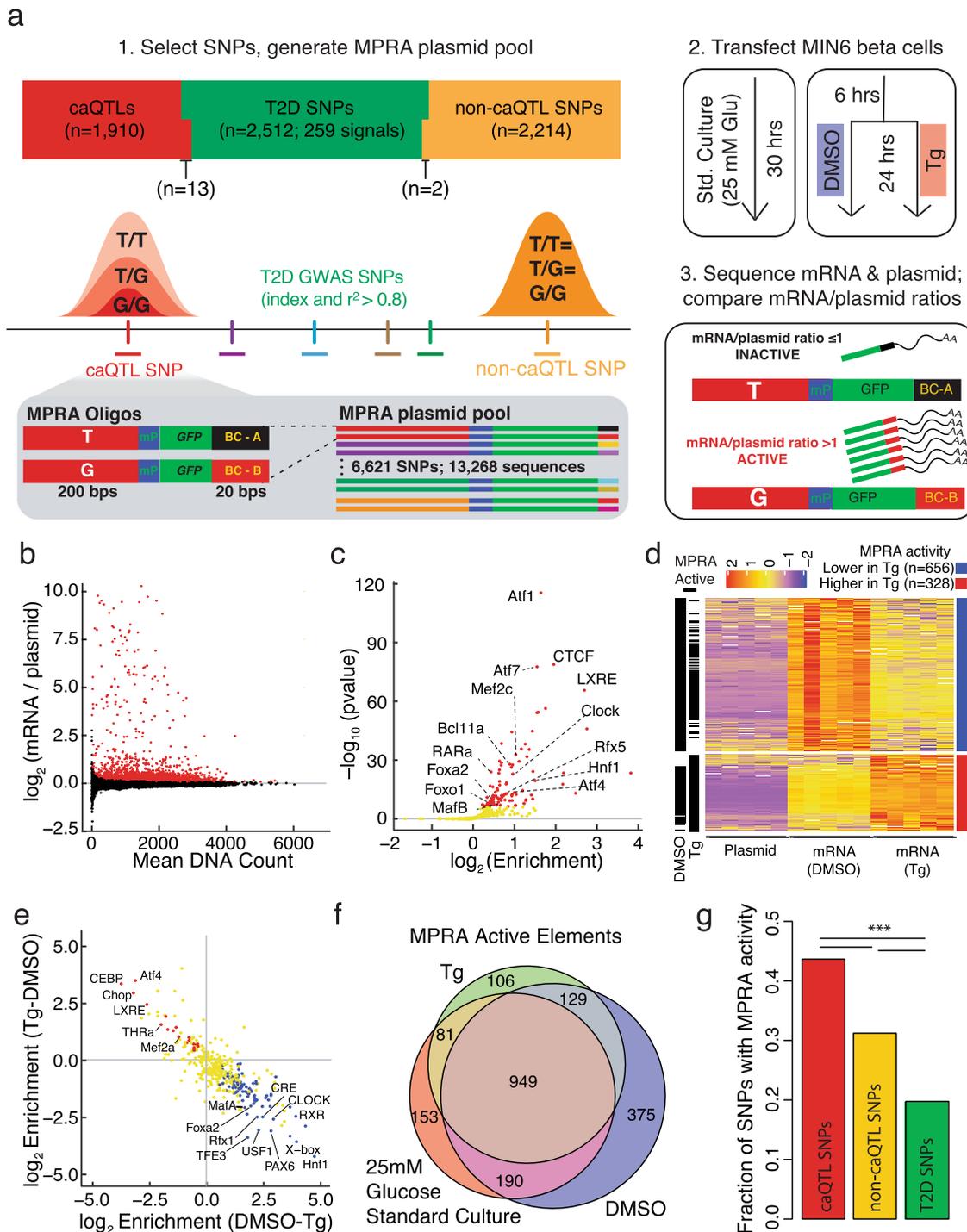
### Selection and testing of sequences for MPRA activity in $\beta$ cells.

To identify CRE sequences that activate  $\beta$  cell transcription and to determine how SNP alleles alter this activity, we employed MPRA in MIN6  $\beta$  cells. The MPRA library consisted of two-hundred base pair (bp) sequences from the human genome containing each allele for SNPs including: (i) 2512 index or linked (EUR  $r^2 \geq 0.8$ ) SNPs/indels from 259 T2D and related quantitative trait association signals in the NHGRI/EBI GWAS Catalogue<sup>28</sup> (“T2D SNPs”); (ii) 1910 SNPs significantly associated with changes in human islet chromatin accessibility (“caQTL SNPs”)<sup>7</sup>; and (iii) 2214 SNPs that overlapped human islet ATAC-seq peaks, but were not significantly associated with changes in human islet chromatin accessibility (“non-caQTL SNPs”)<sup>7</sup> (Methods; Fig. 1a).

Each sequence was cloned upstream of a minimal promoter controlling transcription of *GFP* mRNAs with distinct 20 bp barcodes in their 3' end (Fig. 1a). This MPRA plasmid library was transfected into biological replicates of MIN6 mouse  $\beta$  cells and tested for transcriptional activity in three conditions: standard culture, 24-h exposure to the ER stress-inducing agent thapsigargin (250 nM Tg), or DMSO solvent control conditions (Fig. 1a). For each experimental condition (standard culture, DMSO, Tg), cells were harvested a total of thirty hours after transfection for RNA isolation, *GFP* mRNA capture, and Illumina sequencing of the sequence-associated barcodes (Methods; Fig. 1a). RNA expression of the transfected MPRA libraries was highly correlated between all replicates for each condition and clustered distinctly from the MPRA plasmid library input (Supplementary Figs. 1a–c). Under standard culture conditions, 1373 SNPs (20.7%) had at least one allele exhibiting significantly higher RNA-seq counts compared to the plasmid library input at FDR < 1% (Fig. 1b). We refer to these sequences as ‘MPRA active’ throughout the remainder of the manuscript.

While prior work has routinely used MIN6 mouse  $\beta$  cells to provide meaningful insights into human islet *cis*-regulatory control and the transcriptional effects of T2D risk alleles<sup>5–7,9,11,29–31</sup>, we first confirmed that MIN6  $\beta$  cells appropriately modeled the *cis*-regulatory potential of human islets. Binding motifs of 99 transcription factors (TFs) were significantly enriched (FDR < 1%) in MPRA active elements in MIN6  $\beta$  cells, which notably included motifs for TFs with reported roles in modulating beta cell identity and function (Hnf1, MafA/B, Foxo1)<sup>32–36</sup>, glucose-stimulated insulin secretion (Bcl11a<sup>37</sup>, LXRE<sup>38</sup>, RARa<sup>39</sup>), ER quality control and insulin folding and processing (Atfs), circadian regulation of  $\beta$  cell functions (Clock), and regulation of T2D SNP-containing *cis*-REs (Foxa2<sup>29,40</sup>, Rfx5<sup>11</sup>) (Fig. 1c and Supplementary Data 1). Several of these same motifs were enriched in sorted human islet  $\beta$  cell ATAC-seq peaks (e.g., Fox, Mef2c)<sup>41</sup> and were among the top motifs identified as predictors of human islet regulatory features by recent computational analyses<sup>42</sup>. In addition, empiric binding sites (i.e., ChIP-seq peaks) for PDX1, NKX6.1, MAFB, and FOXA2 in human islets<sup>8</sup> were enriched in the MPRA active elements (Supplementary Figs. 2a). In a comparison with nine human tissues, the chromatin accessibility profile of MIN6  $\beta$  cells most resembled that of human islets (Supplementary Figs. 2b–c). Furthermore, human-mouse sequence similarity did not influence the probability that an element was MPRA active (Supplementary Fig. 3a), and elements overlapping human islet ATAC-seq peaks were more likely to be MPRA active in MIN6 mouse  $\beta$  cells than those that did not (Supplementary Fig. 3b).

Together, comprehensive testing, identification, and analysis of thousands of human sequences using MPRA in MIN6 revealed regulatory features and sequence motifs empirically linked to steady-state  $\beta$  cell transcriptional activation and reinforced MIN6



cells as a valid cellular model to test the  $\beta$  cell transcriptional potential of human sequences.

**Identification of ER stress-responsive  $\beta$  cell regulatory sequences.** The high burden of insulin production and secretion makes  $\beta$  cells particularly susceptible to ER stress<sup>43</sup>, and ER stress has been implicated in the genetic etiology and pathophysiology of both monogenic<sup>44,45</sup> and T2D<sup>22</sup>. Thapsigargin (Tg) blocks calcium transport into the ER lumen and has been widely used to induce the UPR in  $\beta$  cells<sup>46–53</sup>. To identify ER stress-responsive sequences, we compared MPRA activity of sequences in MIN6 cells treated with Tg versus DMSO solvent control (Fig. 1a).

Treatment with Tg increased the expression of ER stress response genes *Ddit3* (Chop), *Hspa5*, and *Edem1*, and reduced *Ins2* expression, confirming UPR induction in MIN6 cells (Supplementary Fig. 1d). ER stress increased the MPRA activity of 328 sequences (representing  $\geq 1$  allele(s) of 275 elements) and decreased MPRA activity of 656 sequences ( $\geq 1$  allele of 449 elements) (Fig. 1d). Elements with increased MPRA activity under ER stress were enriched for motifs of TFs that mediate transcriptional responses to uncompensated ER stress and whose activity and/or abundance is increased in T2D patient islets<sup>54</sup>, such as ATF4 and *DDIT3/CHOP*<sup>15,55–57</sup>, as well as factors linked to pathophysiologic epigenetic changes in the beta cells of diet-induced obese mice (Mef2a<sup>58</sup>) and beta-cell senescence (THRA<sup>59</sup>)

**Fig. 1 Massively parallel reporter assay (MPRA) identifies steady state and ER stress-responsive transcription activating sequences in  $\beta$  cells. a**

Schematic of MPRA study workflow to identify  $\beta$  cell transcription activating sequences. Sequences containing each allele of 2512 SNPs associated with T2D (T2D SNPs), 1910 associated with altered islet chromatin accessibility (caQTLs), and 2214 overlapping islet ATAC-seq peaks but not associated with altered in vivo islet chromatin accessibility (non-caQTL) were selected and tested for their ability to activate transcription in MIN6  $\beta$  cells under three conditions. Std standard, Glu glucose; DMSO dimethylsulfoxide, Tg thapsigargin, mP minimal promoter, BC barcode. **b** Log<sub>2</sub> ratios of barcode/sequence counts in *gfp* mRNA of transfected cells compared to those in the MPRA plasmid library in MIN6  $\beta$  cells grown under standard culture conditions. Each point represents a 200 bp sequence that was tested. Red points denote MPRA active sequences at FDR < 1% ( $n = 2224/13,628$ ). **c** Transcription factor (TF) binding motifs enriched in MPRA active elements ( $\geq 1$  allele) in MIN6 cells grown under standard culture conditions (25 mM glucose). Red dots denote TF binding motifs enriched at FDR < 1%. **d** Heatmap of z-scores for sequences with significantly (FDR < 1%) higher (far right column, red bar;  $n = 328$ , mapping to 275 elements) or lower MPRA activity (blue bar;  $n = 656$ , mapping to 449 elements) in Tg-treated cells compared to DMSO solvent control. Black annotation bars to the left of the heatmap indicate sequences identified as MPRA active in DMSO and/or Tg based on their sequence counts in RNA vs. plasmid DNA input. **e** Transcription factor (TF) binding motifs significantly enriched (FDR < 1%) in elements identified in d with higher (red dots) or lower (blue dots) MPRA activity under ER stress. Yellow dots denote TF motifs with no significant enrichment in either comparison. **f** Venn diagram showing the number of elements ( $\geq 1$  allele) with MPRA activity (FDR < 1%) in each experimental condition. **g** Fraction of 200 nucleotide sequence elements containing caQTL ( $n = 824/1910$ ), non-caQTL (707/2218), or T2D-associated ( $n = 492/2215$ ) SNPs with  $\geq 1$  allele identified as MPRA active in  $\geq 1$  experimental condition. Two-sided Fisher's Exact  $p = 3.12e-16$  (caQTL vs. non-caQTL SNPs),  $1.50e-63$  (caQTL vs. T2D SNPs), and  $1.23e-19$  (non-caQTL vs. T2D SNPs). \*\*\* $p < 0.001$ .

(Fig. 1e; Supplementary Data 2). Elements with decreased MPRA activity were enriched for motifs of  $\beta$  cell TFs that control insulin transcription and secretion<sup>60,61</sup>, such as MAFA, FOXA2, and PAX6 (Fig. 1e; Supplementary Data 2). Consistently, *Ins2* expression was significantly decreased by Tg (Supplementary Fig. 1d), suggesting ER stress leads to the inactivation of the  $\beta$  cell-specific TFs. MPRA thus revealed  $\beta$  cell regulatory elements that respond to ER stress and provided a functional readout of TF dynamics in this (patho)physiologic  $\beta$  cell stress response.

MPRA identified 1938 elements in total for which one or both alleles activated transcription in MIN6  $\beta$  cells for at least one of the experimental conditions tested (Fig. 1f). T2D SNP-containing elements had the lowest proportion of active elements among the three categories of SNPs tested by MPRA (Fig. 1g; Fisher's exact  $p = 1.50e-63$ , caQTL vs. T2D;  $p = 1.23e-19$ , non-caQTL vs. T2D), presumably because the vast majority of those tested ( $n = 2299/2512$ ) did not overlap islet ATAC-seq peaks. Although both caQTL and non-caQTL SNPs overlapped islet ATAC-seq peaks, a significantly higher fraction of elements containing caQTL SNPs were MPRA active (Fig. 1g; Fisher's exact  $p = 3.12e-16$ , caQTL vs. non-caQTL). We hypothesize this is due to the closer proximity of caQTL SNPs to islet ATAC-seq peak summits, as SNP-to-ATAC-seq peak summit proximity was associated with increased MPRA activity (Supplementary Fig. 4).

**MPRA identifies SNPs altering  $\beta$  cell transcriptional activity.**

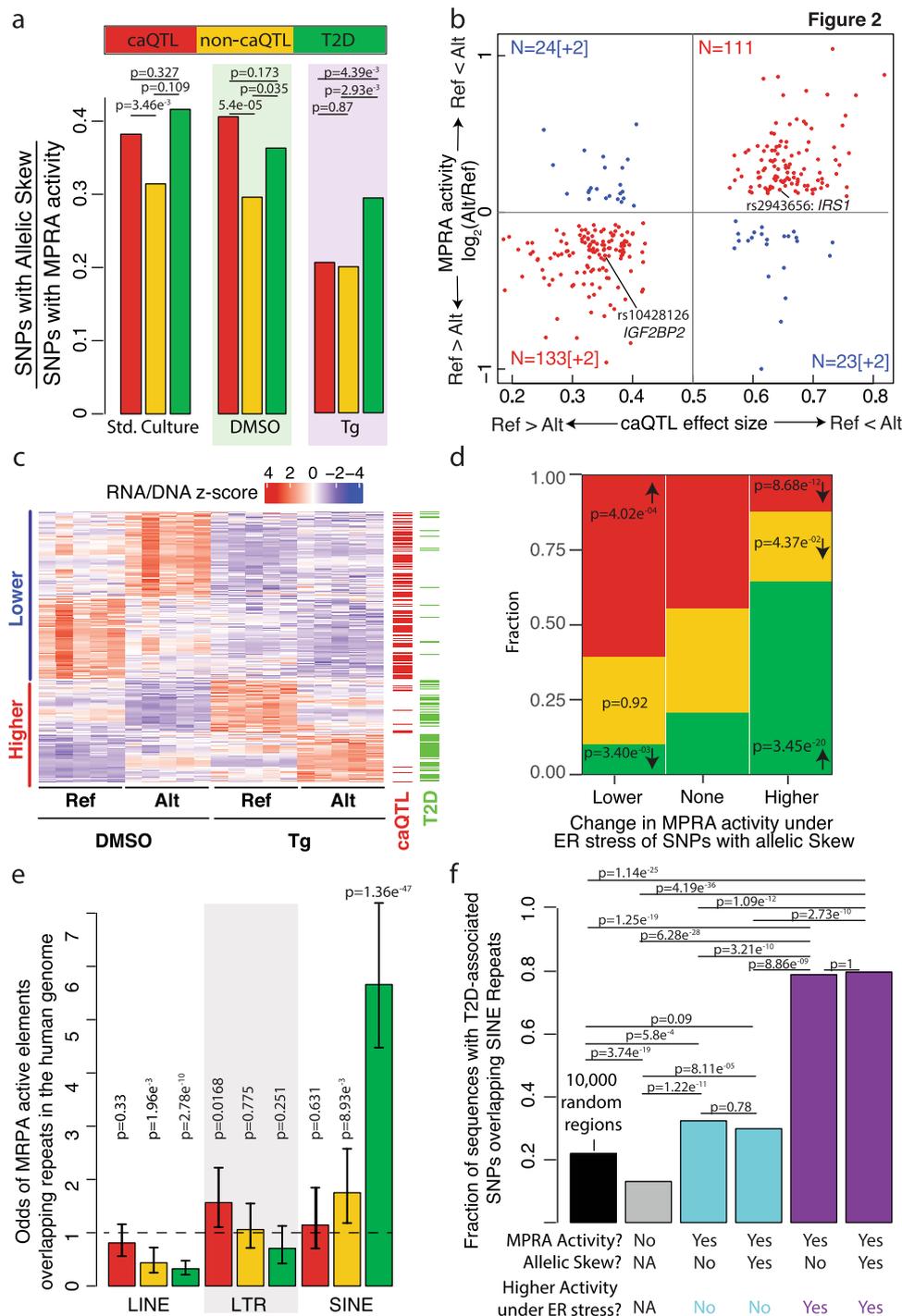
To identify SNPs that altered transcriptional activation in  $\beta$  cells, we compared MPRA activity of each allele under each experimental condition (standard culture, DMSO, Tg). Combining all three categories (caQTL, T2D, and non-caQTL), 879 SNPs exhibited allelic effects on MPRA activity (FDR < 10%) in one or more experimental condition (Fig. 2a, Supplementary Fig. 5a–c, Supplementary Data 3 and 4). For 98.2% ( $n = 332/338$ ; Binomial test  $p = 7.2 \times 10^{-90}$ ) of SNPs that altered MPRA activity in multiple conditions, the direction of allelic effects was concordant (Supplementary Fig. 5a–c). To assess if MPRA properly captured in vivo allelic effects, we assessed their concordance with caQTL effects. Strikingly, 82.8% ( $n = 246/297$ ;  $p = 8.6 \times 10^{-32}$ , binomial test) of caQTL alleles that increased MPRA activity were associated with increased chromatin accessibility in islets (Fig. 2b), underscoring MPRA's ability to report allelic effects relevant to their endogenous, in vivo consequences.

We next studied the impact of ER stress on SNP allelic effects and  $\beta$  cell transcriptional activation. As anticipated based on motif enrichment analyses (Fig. 1e), fewer caQTL SNPs exhibited significant allelic effects on MPRA activity in Tg-treated cells

(Fig. 2a). This was driven by an overall reduction of activity of caQTL elements under ER stress (Fig. 2c) and associated with the enrichment of islet-specific TF binding motifs (e.g., FOXA2, MAFA) in islet caQTL ATAC-seq peaks (Fig. 1e)<sup>7</sup>. In contrast, a larger proportion of elements containing T2D SNPs with allelic effects by MPRA exhibited higher activity under ER stress conditions (Fig. 2c, d). Most of these T2D SNPs ( $n = 190/220$ ) were not islet caQTL and did not overlap islet ATAC-seq peaks in our extensive set of >150,000 islet open chromatin sites from 19 donors. This may be due in part to the steady-state nature of these ATAC-seq profiles, which do not capture regulatory elements made accessible by ER stress-responsive TFs. Alternatively, the stress-responsive increases in these T2D SNP-containing elements may be mediated by repetitive elements, which have been shown to become activated by and enhance transcriptional stress responses<sup>62–64</sup>. Consistent with the latter, 63% (139/220) of T2D SNPs with allelic effects on MPRA activity overlapped repetitive elements. Among three repetitive element categories, only SINES were enriched in MPRA active elements. SINES containing T2D-associated SNPs were five times more likely to be active than any other SNP-containing repetitive elements tested (Fig. 2e). Therefore, we next asked whether T2D SNP-containing elements that overlapped SINES showed higher activity under ER stress. Indeed, 79.4% of T2D SNP-containing elements with higher MPRA activity under ER stress overlapped SINES, a significantly higher proportion compared to T2D SNP-containing elements with lower or no change in MPRA activity under ER stress (Fig. 2f; Supplementary Data 3). *Alu* elements, the most common SINES in the human genome, are primate-specific<sup>65</sup>. When we investigated conservation in 20 mammalian genomes, T2D SNP-containing elements with higher MPRA activity under ER stress were indeed less likely to be conserved beyond the primate lineage (Supplementary Fig. 6).

In summary, MPRA revealed 879 SNPs that alter  $\beta$  cell transcriptional activation, including 220 T2D-associated SNPs. These SNP-containing elements exhibited distinct patterns of transcriptional activity in Tg- vs. DMSO-treated  $\beta$  cells, with caQTL and T2D SNP-containing elements losing or gaining activity under ER stress, respectively.

**MPRA nominates functional SNPs at T2D GWAS signals.** We next sought to nominate T2D causal variants within each locus from the set of 2512 SNPs tested by MPRA. We identified allelic effects for 220 T2D-associated SNPs in 104 distinct association signals (Fig. 3a; Supplementary Data 5–7 and Supplementary Fig. 5d–f), only 17 of which were reported index SNPs. The



number of candidate functional variants identified within the association signals ranged from one to as many as ten, suggesting that some T2D association signals might have more complex effects on multiple genes and/or regulatory elements (Fig. 3b). Importantly, 6/6 T2D risk alleles that significantly altered MPRA activity and in vivo islet chromatin accessibility (islet caQTL)<sup>7</sup> did so in the same direction, and T2D SNP effects on MPRA activity were consistent with their previously reported effects on in vitro luciferase reporter activity, including rs7903146 (*TCF7L2*)<sup>6,66</sup>, rs1635852 (*JAZF1*)<sup>30</sup>, rs12189774 (*VEGFA*)<sup>67</sup>, rs2943656 (*IRS1*)<sup>7</sup>, and rs10428126 (*IGF2BP2*)<sup>7,68</sup>.

10.4% ( $n = 23/220$ ) of T2D SNPs exhibited specific allelic effects on MPRA activity when treated with Tg compared to baseline conditions (Fig. 3a; Supplementary Data 5 and Supplementary

Fig. 5d–f). This included rs12189774, which is predicted to alter the binding of ATF5 (Supplementary Data 7), a TF that forms complexes with PDX1 and ATF4 in stressed  $\beta$  cells to modulate the expression of stress response and apoptosis genes<sup>69</sup>. Although rs12189774 has not been empirically connected to any target gene(s) to date, it may modulate ER stress-responsive VEGFA expression based on reports that ER stress elicits XBP-1(s) and ATF4 binding to the VEGFA promoter and induces VEGFA expression in pancreatic beta cells and other cell types<sup>70–72</sup>. 6.8% ( $n = 15/220$ ) of T2D SNPs altered MPRA activity exclusively in ER stressed cells compared to baseline and DMSO-treated cells (Supplementary Fig. 5d–f; Supplementary Data 5).

For 54 T2D GWAS signals, a single SNP among those tested exhibited allelic effects on MPRA activity (Fig. 3b, black dots;

**Fig. 2 Identification of  $\beta$  cell transcription-modulating allele.** **a** Fraction of caQTL, non-caQTL, and T2D SNP alleles that significantly alter MPRA activity in standard culture, DMSO, and Tg conditions. **b** Correlation between allelic effects on in vivo islet chromatin accessibility (x-axis) and MPRA activity (y-axis) for MPRA active elements containing islet caQTL SNPs. Quadrants 1 and 3 (red) contain SNPs where the allelic effects on chromatin accessibility and MPRA activity were concordant. The number of SNPs in each quadrant is indicated; the number of SNPs for which the relative MPRA activity lies outside the y-axis boundaries are indicated in brackets. Ref = hg19 reference allele; Alt = alternate allele. **c** Heatmap of elements containing SNPs that significantly alter MPRA activity for which overall MPRA activity is higher (red) or lower (blue) in Tg-treated vs. DMSO control-treated cells. Scale bar on top indicates z-scores obtained by centering and scaling the normalized RNA/DNA ratios for the reference and alternate alleles (5 replicates each per condition). Annotations on the right indicate whether the element (row) contains a caQTL and/or T2D-associated SNP. Note that the majority of elements with higher MPRA activity under ER stress contain T2D-associated SNPs. **d** Proportion of elements containing caQTL (red), non-caQTL (yellow), or T2D-associated (green) SNPs with allelic effects on MPRA activity (under any condition) that show lower, no change, or higher MPRA activity under ER stress conditions compared to DMSO control. A significant proportion of caQTL SNPs with allelic skew showed lower MPRA activity under ER stress, while a significant proportion of T2D-associated SNPs showed higher activity in this condition. \*\*\* $p < 0.001$ ; \*\* $p < 0.01$ ; and \* $p < 0.05$ , Fisher's exact test comparing elements with higher or lower MPRA activity under ER stress to those showing no change. **e** Odds of MPRA active elements ( $\geq 1$  allele; any experimental condition) containing caQTL, non-caQTL, or T2D-associated SNPs for overlapping long interspersed nuclear element (LINE), long terminal repeat (LTR), or short interspersed nuclear element (SINE) repetitive elements. Error bars indicate the 95% confidence interval of the odds ratio estimates. **f** Fraction of T2D SNP-containing elements or 10,000 randomly selected 200 bp genomic regions (black bar) overlapping SINE repeats in the human genome. T2D SNP-containing elements are grouped based on whether they show (i) MPRA activity, (ii) allelic skew in MPRA activity, and/or (iii) higher MPRA activity under ER stress. A significantly higher fraction of T2D SNP-containing elements with higher activity under ER stress overlap SINEs compared to those without higher activity under ER stress. \*\*\* $p < 0.001$ ; \*\* $p < 0.01$ ; and \* $p < 0.05$  using two-sided Fisher's exact test (corrected for multiple testing) in panels (a), (d), (e), and (f).

Supplementary Data 6). For instance, rs17748864 was the only SNP among the eight tested in the *PEX5L* locus that altered MPRA activity (Fig. 3c). In contrast to the non-risk C allele, which was inactive in both DMSO and Tg conditions, the T2D risk T allele exhibited an 11-fold greater activity in MPRA and had significantly higher MPRA activity under ER stress (Fig. 3d). These allelic differences in MPRA activity were supported biochemically in electrophoretic mobility shift assays (EMSAs), which identified allele-specific binding of the non-risk C allele by a protein or complex in MIN6 nuclear extracts (Fig. 3e, arrow). Together, these data suggest that the rs17748864 risk T allele conveys robust transcriptional activation activity to this sequence by abrogating the binding of a transcriptional repressor.

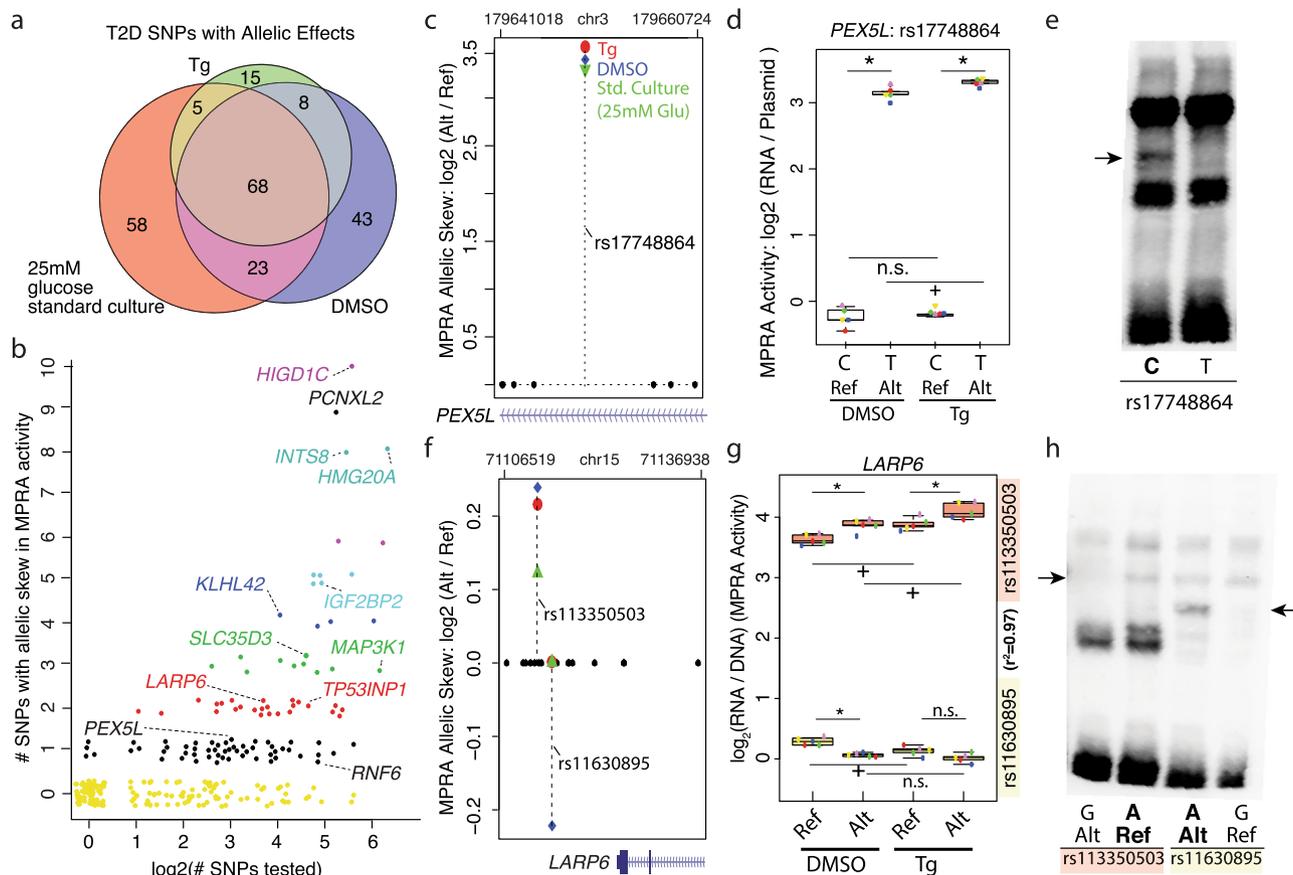
For 25 T2D-associated GWAS signals, we nominated two functional SNPs (Fig. 3b, red points; Supplementary Data 6). For example, in the *LARP6* locus, rs11630895 and rs113350503 were the only functional SNPs among the 16 tested which all resided on the same tightly linked haplotype ( $r^2 = 0.97$ ; Fig. 3f, g). The major allele for each of these SNPs displayed opposing directions-of-effect on MPRA activity in DMSO control conditions and were affected differently by ER stress (Fig. 3f, g). As with the *PEX5L* locus, EMSAs indicate that both transcription-lowering T2D SNP alleles are bound by distinct MIN6 nuclear factors (Fig. 3h, arrows). Thus, while our results provide conclusive support for allelic effects on in vitro binding and transcriptional output, further investigation to understand their relative contributions and directions-of-effect in their endogenous context is needed. Because oligo design and synthesis for this study occurred prior to the report of credible SNP sets for T2D-associated loci<sup>4</sup>, our LD-based approach to select SNPs for testing did not exhaustively and comprehensively test the functionality of all T2D credible set SNPs. However, for this and other GWAS signals, MPRA has nominated multiple high-priority candidate causal SNPs for targeted and mechanistic investigation in the future (Supplementary Data 5–7).

**Integrating MPRA data with QTL maps refines T2D causal alleles/mechanisms.** Finally, we sought to understand the in vivo consequences of the putative functional T2D SNPs nominated by MPRA. Integration with islet caQTL revealed striking correlations between SNP effects on in vitro MPRA activity and in vivo islet chromatin accessibility (Fig. 2b), which confirmed that the short sequence element tested by MPRA has function when residing in

its broader, endogenous context. To define target genes, we determined islet transcripts whose abundance<sup>73</sup> was linked to the functional T2D SNP alleles nominated by MPRA. We found that T2D functional SNPs nominated by MPRA were more enriched for islet eQTLs than T2D SNPs without MPRA activity (Fig. 4a). Integration of MPRA with caQTLs and eQTLs confirmed previously reported effects of the rs10428126 T2D risk allele, which increased reporter activity (Fig. 2b), in vivo chromatin accessibility (Fig. 2b), and *IGF2BP2* expression in islets (Fig. 4a).

Importantly, this integrated approach refined QTL maps to improve insights at additional T2D loci. For example, at the *RNF6* locus, rs4630391 was the only SNP to exert allelic effects on MPRA activity among 28 T2D SNPs in high LD that were tested (Fig. 4b). The C risk allele conveyed 30% higher MPRA activity than the T allele (Fig. 4c) and was associated with increased *RNF6*, *CDK8*, and *WASF3* expression in human islets<sup>73</sup> (Fig. 4d). rs4630391 is one of eight SNPs in or immediately adjacent to an ATAC-seq peak containing an islet caQTL (Fig. 4e). Islet donors with AA genotype at the T2D-associated, lead caQTL SNP rs34584161 ( $n = 10$ ) exhibited higher chromatin accessibility than those with AG ( $n = 8$ ) or GG ( $n = 1$ ) genotypes. Although rs34584161 has been recently reported as the SNP with the highest genetic posterior probability of being the causal allele for T2D association in this locus (PPAg = 0.67)<sup>4</sup>, only rs4630391 (PPAg = 0.037) exhibited transcription-modulating effects in MPRA (Fig. 4b). Interestingly, this SNP overlaps an Alu SINE element at the edge of an islet ATAC-seq peak and exhibits allelic effects under ER stress conditions (Fig. 4c, e). Importantly, allelic effects on MPRA activity and chromatin accessibility were concordant, i.e., the rs4630391-C allele was associated with both higher in vitro MPRA activity in MIN6 and increased in vivo chromatin accessibility in human islets (Fig. 4b, e). EMSA revealed that decreased transcriptional activity of the rs4630391-T allele was accompanied by increased T allele-specific binding in MIN6 nuclear extracts (Fig. 4f, arrow). While this does not definitively rule out that the lead caQTL SNP (rs34584161) or other SNPs in high LD are also functional, these data provide compelling functional support for rs4630391 as a putative causal variant despite its lower reported genetic posterior probability.

MPRA also refined the relative functional contributions of two T2D-associated SNPs in high LD. At the *SLC35D3* locus, we previously identified rs6937795 and rs6917676, located 15 bp apart, as islet caQTLs<sup>7</sup> for which the risk alleles were associated



**Fig. 3** MPRA identifies functional SNPs at 104 T2D-associated GWAS signals. **a** Venn diagram showing the number of T2D-associated SNP alleles altering MPRA activity (FDR < 10%) in each experimental condition. Tg thapsigargin, DMSO dimethyl sulfoxide. **b** Scatter plot of 220/2515 T2D-associated index and high LD ( $r^2 \geq 0.8$ ) SNP alleles, representing 104/259 association signals from the NHGRI/EBI GWAS catalog (x-axis; log scale), that significantly altered MPRA activity at FDR < 10%. Jitter in the plot is used to visually separate individual points. **c** Among eight SNPs in high LD at the *PEX5L* locus, only rs17748864 exhibited significant allelic effects on MPRA activity. Higher MPRA activity was detected for the rs17748864-T allele in all three experimental conditions compared to rs17748864-C. **d** MPRA activity for rs17748864 reference ‘C’ and alternate ‘T’ alleles in DMSO and Tg conditions. Each of the five paired biological replicates for each MPRA experiment are indicated by distinct shapes and colors. \*, FDR < 10% for allelic effects on MPRA activity; +, FDR < 1% for increased MPRA activity; n.s., not significant. **e** Allele-specific binding of sequences containing rs17748864 alleles by MIN6  $\beta$  cell nuclear factors. EMSA was completed with biotin-labeled probes containing the rs17748864 ‘C’ or ‘T’ allele incubated with MIN6 nuclear extract. Arrow indicates the allele-specific difference in nuclear factor binding. **f** Among 19 SNPs in high LD in the *LARP6* locus, two exhibited significant allelic effects on MPRA activity (rs113350503, rs11630895). rs113350503 alternate ‘G’ allele increased MPRA activity compared to the reference ‘A’ allele in all experimental conditions. rs11630895 reference ‘G’ allele conferred higher MPRA than the alternate ‘A’ allele in DMSO control. **g** MPRA activity for the rs113350503 reference ‘A’ and alternate ‘G’ alleles (top) and for the reference ‘G’ and alternate ‘A’ alleles of rs11630895 (bottom). Distinct data point shapes and colors of indicate each of five paired biological replicates for each MPRA experiment. \*, +, and n.s. assignments are as in panel (d). **h** Allele-specific binding of sequences containing different rs113350503 or rs11630895 alleles by  $\beta$  cell nuclear factors. EMSA was completed with biotin-labeled probes containing reference or alternate alleles for each SNP and incubated with MIN6 nuclear extract. Arrows indicate allele-specific differences in nuclear factor binding. Box plots in panels (d) and (g) display the minimum, maximum, median, first quartile, and third quartile of each data set.

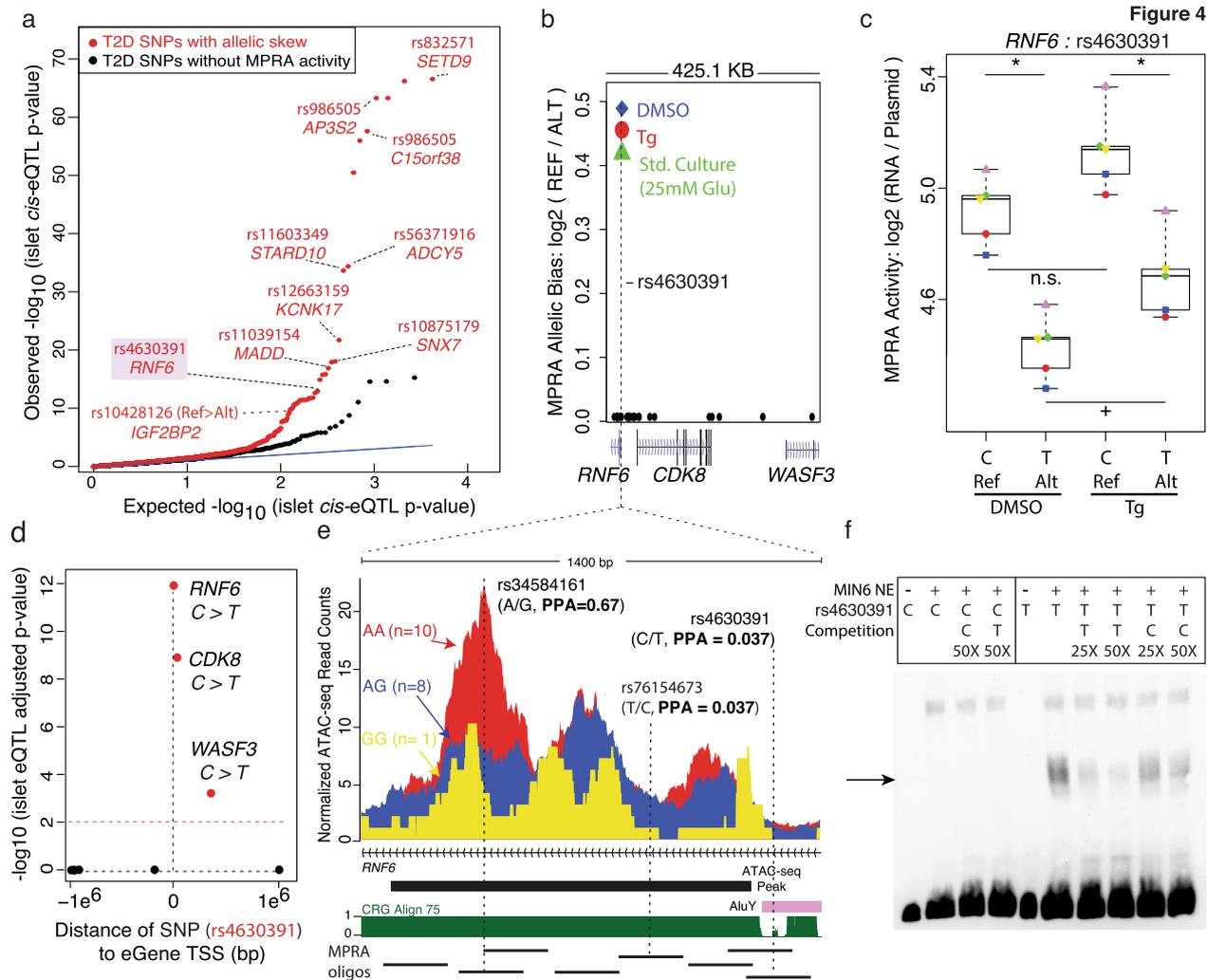
with significantly higher chromatin accessibility in islets (Fig. 5a). Their high LD ( $r^2 = 0.99$ ) make it impractical to separate the individual contributions of each SNP allele to altered regulatory function using population genetic approaches. MPRA, however, allowed each pairwise, synthetic combination to be tested independently to identify which of the SNPs has functional effects on  $\beta$  cell transcription. This revealed rs6917676, not rs6937795, as the SNP that altered transcriptional activity (Fig. 5b). Interrogation of recent steady-state islet eQTL data<sup>73</sup> indicated that the rs6917676 T allele, which increased MPRA activity, was associated with increased *SLC35D3* expression (Fig. 5c). Allelic effects of rs6917676 were lost in ER stressed cells, while the activity and allelic effects of rs947734 and rs947735, two SNPs in high LD ( $r^2 = 1$ ) ~4.6 kb away in a SINE, increased (Fig. 5d). While additional mechanistic investigation is clearly needed to understand the causal relationship to T2D, this

observation illustrates the value of evaluating variant haplotypes and their effects in multiple conditions.

**Discussion**

In this study, we tested 13,252 sequences containing alleles of 6,621 SNPs for their ability to activate and modulate transcription in MIN6  $\beta$  cells under standard culture conditions and after ER stress or paired solvent control exposures. In total, 29.9% of elements ( $n = 1982/6621$ ) exhibited increased  $\beta$  cell transcriptional activity from a minimal promoter. SNP alleles in 44.3% of these elements altered MPRA activity ( $n = 879/1982$ ), including 220 SNPs associated with T2D risk by GWAS.

Multiple lines of evidence indicate that MPRA in MIN6 mouse  $\beta$  cells is capable of identifying features of  $\beta$  cell transcription activation of human sequences and allelic effects of SNPs on this activity, despite potential limitations of cross-species testing of

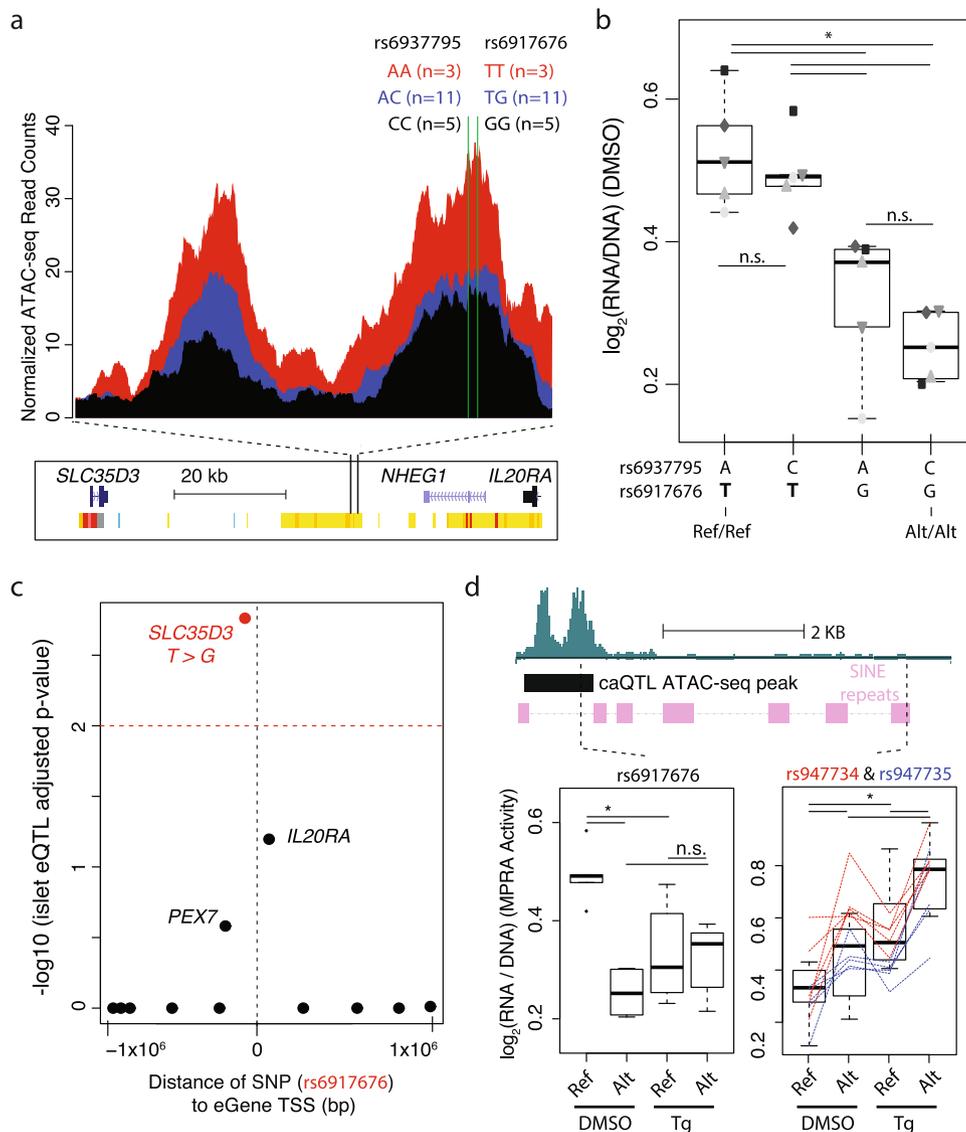


**Fig. 4** MPRA identifies putative T2D causal SNPs altering islet expression (eQTL) and chromatin accessibility (caQTL). **a** Quantile-quantile plot of observed (y-axis) vs. expected (x-axis) islet eQTL p-values from the InsPIRE Consortium for T2D SNPs, categorized based on whether  $\geq 1$  SNPs in LD had allelic skew (red dots), or 0 SNPs in LD had significant MPRA activity (black dots). **b** Plot of genome location of 27 SNPs in high LD ( $r^2 \geq 0.8$ ) with the T2D-associated index SNP rs10507349 tested with MPRA (*RNF6* locus). Allelic effects on MPRA activity were detected for rs4630391 in the same direction across all three experimental conditions. **c** MPRA activity for the reference ('C') and alternate ('T') alleles of rs4630391 in DMSO and Tg experimental conditions. Each of the five paired biological replicates for each MPRA experiment are indicated by distinct shapes and colors. \* indicates significant (FDR < 10%) allelic effects on MPRA activity; + or n.s. indicates significant (FDR < 1%) or not significant changes in MPRA activity in each condition, respectively. Box plots display the minimum, maximum, median, first quartile and third quartile of each data set. **d** Plot of InsPIRE Consortium islet eQTL association p-values between rs4630391 genotypes and expression of genes  $\pm 1$  megabase. **e** Magnified view of genomic region containing the MPRA-nominated functional T2D SNP (rs4630391) and lead islet caQTL SNP (rs34584161). Credible set SNP genetic posterior probabilities of association (PPA) are included for SNPs for which they were reported<sup>4</sup>. Normalized ATAC-seq reads were higher for islets from donors with rs34584161 AA homozygous genotypes (red;  $n = 10$  donors) than those from AG heterozygous (blue;  $n = 8$ ) or GG homozygous (yellow;  $n = 1$ ) genotypes. 'CRG Align 75' indicates mappability of 75-mer sequences to the hg19 reference genome. Thick black bar indicates ATAC-seq peak in human islets. Thin black lines below indicate the MPRA oligos tested in this region. **f** Allele-specific binding of sequences containing reference and alternate alleles for rs4630391. EMSA with biotin-labeled probes containing rs4630391 C or T alleles were incubated with MIN6 nuclear extract. The arrow indicates allele-specific nuclear factor binding to the T allele, which was more efficiently competed with unlabeled T probe than unlabeled C probe.

human sequences using episomal assays. First, MPRA active elements were enriched for both motifs and empiric binding of several islet TFs governing human islet cell identity and function under steady-state conditions. Second, under ER stress, changes in MPRA activity reflect reported in vivo changes in TF levels and their activity in  $\beta$  cells<sup>54–56,58–61</sup>. Finally, allelic effects on MPRA activity in MIN6 exhibited a significant, positive correlation with their effects on in vivo human islet chromatin accessibility. Moreover, the likelihood of MPRA activity was higher for elements in the vicinity of ATAC-seq peak summits, and SNPs closer to ATAC-seq peak summits were more likely to alter

MPRA activity and chromatin accessibility. These results confirm cross-species transferability of this assay and help prioritize sequences within open chromatin regions for their importance in regulating human  $\beta$  cell transcriptional activity.

In total, 10.4% ( $n = 23/220$ ) of T2D SNPs exhibiting allelic effects on MPRA activity did so only when treated with Tg compared to baseline conditions. This included a SNP that has a comparable PPA<sub>g</sub> to the top SNP in this association signal (PPA<sub>g</sub> = 0.093 vs. 0.12, respectively) and resides near *VEGFA*, a putative T2D effector gene whose expression is induced by ER stress via direct binding of the ER stress-responsive TFs XBP-1(s) and ATF4 to its promoter in



**Fig. 5** MPRA refines T2D-associated haplotype in *SLC35D3* locus. **a** Islet caQTL in the *SLC35D3* locus. Two T2D SNPs, rs6937795 and rs6917676 (15 bps apart;  $r^2 = 0.99$ ; green lines), are both significantly associated with altered chromatin accessibility changes (caQTLs) in the *SLC35D3* locus in islets from 19 different individuals. **b** MPRA activity ( $\log_2$ RNA/plasmid ratios) for each of the four possible rs6937795 and rs6917676 allelic combinations tested in the MPRA library. Only rs6917676 alleles altered MIN6 MPRA activity. The rs6917676 T allele is associated with both increased chromatin accessibility and MPRA activity. Shape/shading of points indicate the five paired biological replicates for each MPRA experiment. **c** Plot of InsPIRE Consortium islet eQTL  $-\log_{10}$  association p-values between rs6917676 genotypes and expression of genes  $\pm 1$  megabase from the SNP. **d** MPRA activity ( $\log_2$ RNA/plasmid ratio) of caQTL SNP rs6917676 (left), and two additional SNPs [rs947734 (red) and rs947735 (blue); 43 bp apart] that overlap the same SINE, in ER stress (Tg-treated) compared to DMSO control conditions. Red and blue dashed lines indicate relationships between alleles for each biological replicate. \* denotes FDR < 10%. Box plots in panels (b) and (d) display the minimum, maximum, median, first quartile and third quartile of each data set. Points outside the minimum and maximum values are outliers.

pancreatic beta cells and other cell types<sup>70–72</sup>. 6.8% ( $n = 15/220$ ) of T2D SNPs that altered MPRA activity were detected exclusively in ER stressed cells. This proportion of T2D SNPs with ER stress-specific effects is similar to that of the broader set of SNPs altering ER stress-specific MPRA activity in this study (7.2%). Moreover, it is comparable to the reported percentage of context-specific regulatory element use or activation<sup>74–78</sup>, including cytokine-responsive elements in islets (4.5%;  $n = 3798/84,162$ )<sup>76</sup> and latent enhancers in monocytes (8.1–15%)<sup>79</sup>, and within the range of genetic variants reported to alter stimulus-responsive gene expression in monocytes exposed to three immunogenic stimuli (3, 9, and 17% response expression QTL upon exposures to LPS, MDP, and dsRNA, respectively)<sup>80</sup>. In the future, it will be interesting and important to determine if the variants altering ER stress-responsive MPRA

activity alter stress-responsive islet chromatin accessibility, active histone modifications, or target gene expression in vivo and, more broadly, to test SNP effects on cis-regulatory element use or activity in response to a range of (patho)physiologic stimuli and stressors.

Surprisingly, we found that multiple T2D SNP-containing elements in SINEs were active in MPRA and responded to ER stress with increased activity. These data suggest that SINEs, and SNPs within them, may play underappreciated roles in modulating  $\beta$  cell transcriptional programs in response to stress or other stimuli. Recent studies have contributed to an emerging appreciation of the importance of these elements in epigenetic and transcriptional regulation, demonstrating repetitive element-mediated oncogene activation and modulation of chromatin structure<sup>81–89</sup>. Alu/SINEs have been shown to be

transcriptionally induced by cellular stressors<sup>63,90</sup> and are emerging as pervasive transcriptional modulators of cellular functions and stress<sup>62,64,81,85,91</sup>. Three of four SNPs tested by EMSA overlapped SINEs (rs113350503 and rs11630895 in the *LARP6* locus, rs4630391 in the *RNF6* locus) and showed specific binding for the allele with lower MPRA activity, suggesting that the activating alleles disrupt binding of a transcriptional repressor. Future studies elucidating target genes of these and other MPRA active, SINE-containing regulatory elements will be necessary to fully understand the functional consequences of sequence variation in these transcriptionally active repetitive element sequences and the potential role of SINE/Alu exaptation<sup>83,92,93</sup> in the genetics of islet (dys)function and T2D.

Finally, a key challenge in T2D genetics is to identify the functional SNPs from multiple variants in high LD per association signal. This is the first study to test thousands of T2D-associated SNPs for their empiric effects on transcriptional activation in  $\beta$  cells. MPRA identified 220 SNP alleles associated with T2D, representing 104 distinct association signals. Candidate causal T2D SNPs nominated by MPRA include those previously studied using targeted, low throughput luciferase assays, such as rs7903146 (*TCF7L2*)<sup>6,66</sup>, rs1635852 (*JAZF1*)<sup>30</sup>, rs12189774 (*VEGFA*)<sup>67</sup>, rs2943656 (*IRS1*)<sup>7</sup>, and rs10428126 (*IGF2BP2*)<sup>7,68</sup>. Importantly, directions-of-effect detected for the T2D risk alleles by MPRA were consistent with those observed in each previous study. At 54 T2D association signals, only one SNP among all SNPs tested exhibited significant allelic effects on MPRA, nominating it as a putative causal SNP for that respective signal. Two or more candidate causal SNPs were identified for 50 T2D association signals, including *LARP6*, wherein EMSA and MPRA demonstrated allelic effects on both nuclear factor binding and transcriptional activity for two SNPs in high LD. The components of T2D risk at this and the 49 other GWAS signals may therefore result from a combined effect of multiple functional SNPs. Although this study was underway before T2D credible set SNPs were reported<sup>4</sup>, select loci illustrate how MPRA may help to evaluate and prioritize them by providing functional evidence in support of SNPs with high genetic posterior probabilities (e.g., rs7903146 (*TCF7L2*, PPAg = 0.59), rs3802177 (*SLC30A8*, PPAg = 0.57), rs10811661 and rs10811660 (*CDKN2A/B*, PPAgs = 0.47, 0.41), rs11603349 (*CENTD2/ARAP1*, PPAg = 0.22), rs4846567 (*LYPLAL1*, PPAg = 0.17), rs2879813 (*TP53INP1*, PPAg = 0.14)) or by identifying SNPs with lower genetic posterior probability for T2D association signals, such as those in the *RNF6*, *SLC35D3*, *ANK1/NKX6-3*, *JAZF1*, *SPRY2*, *THADA*, and *VEGFA* loci, as candidate causal SNPs. However, it is possible that SNPs with low T2D association posterior probabilities identified as functional by MPRA may not be causal, and that technical limitations such as MPRA design, fragment size, cell line used, and conditions tested, may preclude the identification of some true causal SNPs. Therefore, comprehensive testing of T2D credible set SNPs by MPRA across multiple metabolic cell types and relevant (patho)physiologic states in the near future will be critical to modify poster probabilities and nominate functional T2D variants.

## Methods

**MPRA library design.** 200 base pair sequences, with 100 bps flanking each side of 6621 SNPs were included in our MPRA library. The SNPs belong to three categories:

1. T2D-associated SNPs/indels: SNPs ( $n = 2299$ ), small insertions ( $n = 72$ ), and small deletions ( $n = 129$ ) in linkage disequilibrium ( $r^2 \geq 0.8$ ) with T2D-associated index SNPs ( $n = 259$ ) were selected as previously described<sup>94</sup> for synthesis and testing. Briefly, T2D-associated SNPs were retrieved from the NHGRI/EBI GWAS Catalog (accessed 19 January 2017) and LD-pruned using PLINK version 1.9<sup>95</sup> with parameters “-maf 0.05-clump-clump-p1 0.0001-clump-p2 0.01-clump-r2 0.8-clump-kb1000” to remove index SNPs representing redundant association signals. Additional SNPs from the 1000 Genomes Phase 3 reference panel were identified and included based on their high LD ( $r^2 \geq 0.8$ ) in EUR with each retained index SNP.

2. Islet chromatin accessibility quantitative trait loci (caQTLs): 1910 SNPs previously identified as having a significant association with altered in vivo chromatin accessibility in islet samples were also included<sup>7</sup>. Only SNPs within a given ATAC-seq peak were considered and tested for their association with altered accessibility of that peak. For 1816 caQTLs, one SNP was found to show a significant correlation with chromatin accessibility changes in human islets, all of which were included in the MPRA library (Bonferroni adjusted  $p$ -values  $< 0.023$ ). For 94 caQTLs, two SNPs  $< 25$  bp apart showed significant correlations with islet chromatin accessibility changes, so all four allelic combinations were synthesized and tested.
3. Non-caQTL SNPs: 2214 SNPs that overlapped islet ATAC-seq peaks but did not significantly alter the accessibility of those peaks<sup>7</sup> were also synthesized and tested. Since islet caQTLs were identified in a relatively small cohort of individuals ( $n = 19$ ), the following criteria were used to include SNPs for which the caQTL study was more appropriately powered to detect associations with chromatin accessibility: (i) unadjusted  $p$  value  $> 0.2$ ; and (ii) minor allele frequency  $> 0.125$ . SNPs overlapping individual-specific peaks or sharing peaks with other SNPs were removed. The 2,214 non-caQTL SNPs were randomly selected for inclusion in the MPRA library from the 15,178 SNPs that passed these inclusion criteria.

The vast majority of SNPs and elements tested belonged to only one of three categories. However, T2D-associated SNPs overlapping 13 ATAC-seq peaks were significantly associated with chromatin accessibility in islets (caQTLs). Therefore, for analysis purposes, whenever SNPs were required to belong to only one of the three categories above (such as Fig. 2a), they were not categorized as caQTLs, but as being T2D-associated only.

**MPRA library construction.** The MPRA library was constructed as previously described<sup>24</sup>. Briefly, oligos were synthesized (Agilent Technologies) as 230 bp sequences containing 200 bp of genomic sequences and 15 bp of adaptor sequence on either end. Unique 20 bp barcodes were added by PCR along with additional constant sequence for subsequent incorporation into a backbone vector by Gibson assembly. The oligo library was expanded by electroporation into *E. coli*, and the resulting plasmid library was sequenced by Illumina 2 × 150 bp chemistry to acquire oligo-barcode pairings. The library underwent restriction digestion, and GFP with a minimal TATA promoter was inserted by Gibson assembly resulting in the 200 bp oligo sequence positioned directly upstream of the promoter and the 20 bp barcode falling in the 3' UTR of GFP. After expansion within *E. coli* the final MPRA plasmid library was sequenced by Illumina 1 × 31 bp chemistry to acquire a baseline representation of each oligo-barcode pair within the library. Barcodes mapping to more than 1 sequence were discarded from all downstream analyses. Note: Two separate batches of the MPRA library were prepared. The first batch was used to perform MPRA under standard culture conditions. This MPRA library was then electroporated into *E. coli* to obtain a second batch of the MPRA library, which was used for the paired DMSO-Tg experiments.

**MPRA library transfection into MIN6 cells.** 10 million MIN6 cells were seeded in each of seven 15 cm<sup>2</sup> dishes. The cells were 60–70% confluent the next day. Each 15 cm<sup>2</sup> dish was replaced with 20 ml of fresh media and transfected with 7  $\mu$ g of the MPRA plasmid library using 55  $\mu$ l Lipofectamine 2000 (38% transfection efficiency). Six hours after transfection, media was either (i) not changed (MPRA under standard culture conditions), (ii) replaced with media containing 250 nM thapsigargin (Tg) dissolved in 0.025% DMSO, or (iii) replaced with media containing 0.025% DMSO. Thirty hours after transfection, cells were trypsinized and collected by centrifugation. Cell pellets were frozen at  $-80^\circ\text{C}$ . For each condition (standard culture, DMSO, or Tg), MIN6 cells were transfected on five separate days to generate biological replicates.

**RNA isolation and MPRA RNA-seq library generation.** RNA was extracted from frozen cell pellets using the Qiagen RNeasy Midi kit. Following DNase treatment, a mixture of 3 GFP-specific biotinylated primers (Supplementary Data 8; #120, #123, and #126) were used to immunoprecipitated GFP transcripts using Streptavidin C1 Dynabeads (Life Technologies). Following another round of DNase treatment, cDNA was synthesized from GFP mRNA using SuperScript IV and purified with AMPure XP beads. Quantitative PCR using primers specific for GFP (Supplementary Data 8; #34 and #52) was used to determine the cycle at which linear amplification begins for each replicate. Replicates were diluted to approximately the same concentration based on the qPCR results, and PCR with primers #34 and #52 was used to amplify barcodes associated with the  $\sim 13.5$ k sequences included in the MPRA library for each replicate (9 cycles for standard culture, and 13 cycles for DMSO/250 nM Tg). A second round of PCR (6 cycles) was used to add Illumina sequencing adaptors to the DNA/RNA replicates. The resulting MPRA barcode libraries were spiked with 5% PhiX and sequenced using Illumina single-end 31 bp chemistry (with 8 bp index read), clustered at 80–90% maximum density.

**MPRA data analysis.** Data from the MPRA was analyzed as previously described<sup>24</sup>. Briefly, the sum of the barcode counts for each oligo within replicates was median normalized, and oligos showing differential expression relative to the

plasmid input were identified by modeling a negative binomial distribution with DESeq<sup>26</sup> and applying a false discovery rate (FDR) threshold of 1%. For sequences that displayed significant MPRA activity, a paired *t*-test was applied on the log-transformed mRNA/plasmid ratios for each experimental replicate to test whether the reference and alternate allele had similar activity. An FDR threshold of 10% was used to identify SNPs with significant effects on MPRA activity between alleles. Because the MPRA testing standard culture conditions was performed with a separate MPRA library preparation, the DMSO-Tg MPRA results were not directly compared to MPRA performed under standard culture conditions.

**Annotating repetitive elements tested with MPRA.** The ‘RepeatMasker’ track for hg19 was downloaded from the UCSC genome browser. Among the ten different classes of repeats, only three classes (long interspersed nuclear element (LINE), long terminal repeat (LTR), and SINE) overlapped more than 100 elements tested with MPRA. Therefore, only these three classes of repeats were assessed for associations with MPRA activity.

**TF motif enrichment.** Homer<sup>97</sup> findMotifsGenome.pl script was used to investigate TF motifs enriched in a given set of elements. Elements with lower MPRA activity under ER stress were used as background to identify TF motifs enriched in elements with higher MPRA activity under ER stress, and vice-versa (parameters: hg19, -size given). 2008 T2D-associated elements with no MPRA activity were used as background to identify TF motifs enriched in the 492 T2D SNP-containing elements with significant MPRA activity (parameters: hg19, -size given). For the cross-species ATAC-seq peak analysis, ATAC-seq peaks shared with other human cell types were used as background to identify TF motifs enriched in unique ATAC-seq peaks (parameters: mm9, -size given).

**Analysis of islet ChIP-seq data.** Chromatin immunoprecipitation sequencing (ChIP-seq) data from Pasquali et al.<sup>8</sup> were aligned to the hg19 reference human genome as previously described<sup>6</sup>. Elements tested with MPRA were then overlapped with ChIP-seq peaks to conduct Fisher’s exact tests using R.

**Electrophoretic mobility shift assay (EMSA).** 21-bp biotin end-labeled complementary oligonucleotides were designed with each SNP allele of interest in the 11th position of the oligo (Integrated DNA Technologies; Supplementary Data 8). For each SNP tested, complementary oligos were annealed to create double-stranded probes for each allele tested. Nuclear extract was prepared from MIN6 β cells using the NE-PER Extraction Kit (Thermo Scientific), and EMSA were completed using the LightShift Chemiluminescent EMSA kit (Thermo Scientific) according to the manufacturer’s instructions. Binding reactions consisted of 1× binding buffer, 1 μg poly dI-dC, 4 μg MIN6 nuclear extract, and 200 fmol labeled probe. Reactions were incubated at 25 °C for 25 m. For competition reactions, 25- and 50-fold excess of non-biotinylated double-stranded probes for either allele was included and pre-incubated in the reaction mixture for 15 m. DNA-protein complexes were detected by chemiluminescence. EMSAs were completed on two or more separate occasions to ensure that results were consistent.

**MPRA-based interrogation of islet eQTLs.** InsPIRE Consortium<sup>73</sup> islet eQTL *p*-values were retrieved for genes within 1 megabase (Mb) of each T2D-associated SNP included in the MPRA library. In Fig. 4a, nominal islet eQTL *p*-values of T2D-associated SNPs for which ≥1 SNP in high LD ( $r^2 > 0.8$ ) exhibited significant allelic effects on MPRA activity were plotted and compared to those for which no SNPs in high LD exhibited MPRA activity. For locus-specific plots in Figs. 4d and 5c, nominal *p*-values were adjusted for multiple testing of genes within 1 Mb on either side of the SNP (Bonferroni corrected *p*-value cutoff = 0.01).

**Mapping human regulatory sequences tested with MPRA to mammalian genomes.** The UCSC genome browser Liftover tool was used to map human sequences (hg19) tested with MPRA to 20 mammalian genomes (with a minimum ratio of 0.20 bases that must remap; allowing for multiple output regions). The 20 mammalian genomes are: papAnu2 (Baboon), felCat5 (Cat), PanTro6 (Chimpanzee), BosTau7 (Cow), canFAM3 (Dog), loxAfr3 (Elephant), nomLeu3 (Gibbon), gorGor3 (Gorilla), equCab2 (Horse), mm9 (mouse), ponAbe2 (Orangutan), aiMel1 (Panda), susScr11 (Pig), ochPri3 (Pika), oryCun2 (Rabbit), rn5 (Rat), rheMac8 (Rhesus), oviAri3 (Sheep), sorAra2 (Shrew), speTri2 (Squirrel). Human sequences that did not lift over to the genome assembly of a given species were subsequently classified as not conserved (with a minimum ratio of 0.20 bases that must remap; allowing for multiple output regions).

To obtain human-mouse sequence similarity measures, Liftover was performed 99 times with the minimum ratio of bases that must remap ranging from 0.01 to 1.00 in increments of 0.01 (allowing for multiple output regions). The R package ‘sm’ was used plot density of human-mouse sequence similarity and perform non-parametric bootstrap hypothesis tests of equality. Human sequences that did not lift over to the mm9 mouse genome with even 1% sequence similarity were classified as having 0% sequence similarity.

### Cross-species mapping of human ATAC-seq peaks to MIN6 ATAC-seq peaks.

Human and mouse ATAC-seq data were processed as previously described<sup>7</sup>. Briefly, low-quality portions of reads were trimmed using Trimmomatic<sup>98</sup> and aligned to the hg19 or mm9 genome assembly using Burrows Wheeler Aligner-MEM. For each replicate, duplicate reads were removed after shifting. Technical replicates were merged using SAMtools and peaks were called using MACS2<sup>99</sup> (with parameters -callpeak-nomodel -f BAMPE) at FDR < 1%. ATAC-seq peak summit positions were obtained from MACS2 output files. The liftover tool in the UCSC genome browser was used to map human ATAC-seq peaks to the mouse (mm9) genome using a minimum ratio of 0.10 bases that must remap (not allowing for multiple output regions). Using bedtools, human ATAC-seq peaks mapping to the mouse genome were then overlapped with MIN6 ATAC-seq peaks.

**Identification of TF binding motifs disrupted by T2D-associated SNPs.** Genes expressed in MIN6 beta cells were identified by fitting a Gaussian mixture model (two components) to RNA-seq FPKM values. After identifying genes expressed in MIN6 beta cells (FPKM values ≥ 1.63), mouse gene names were converted to corresponding human homologs using the R Package, ‘biomaRt’<sup>100</sup>. After filtering for expression in MIN6 beta cells, the R Package, ‘MotifBreakR’<sup>101</sup>, was used to identify TF binding motifs disrupted by SNPs with allelic skew in MPRA activity (Supplementary Data 4 and 7). Parameters used were: pwmList = homococo, threshold = 1e-2, filterp = TRUE, method = “ic”. In addition to MotifBreakR, predictions for TF binding motifs disrupted by T2D-associated and caQTL SNPs with allelic skew in MPRA activity were also obtained from SNP2TFBS<sup>102</sup> (default options) (Supplementary Data 4 and 7).

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

### Data availability

All datasets generated and analyzed during the current study are publicly available in GEO under Accession GSE145643 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145643>), which includes fastq files containing barcode sequences in the 3’UTR of gfp in the plasmid library and MIN6 RNA samples and processed files containing the sum of all barcode counts for each test construct in the plasmid DNA and MIN6 RNA samples. Human islet ATAC-seq data were obtained from NCI Sequence Read Archive Accession SRP117935. Summary InsPIRE Consortium<sup>73</sup> islet eQTL statistics were obtained from <https://zenodo.org/record/3408356> and SNP2TFBS predictions for transcription factor motifs altered by MPRA-modulating SNP alleles were obtained from <https://ccg.epfl.ch/snp2tfbs/>.

### Code availability

Code used for analysis in this paper is available at [https://github.com/UcarLab/MPRA\\_Khetan](https://github.com/UcarLab/MPRA_Khetan) and linked to Zenodo at <https://doi.org/10.5281/zenodo.4974390>

Received: 13 February 2020; Accepted: 10 August 2021;

Published online: 02 September 2021

### References

- Prentki, M. & Nolan, C. J. Islet beta cell failure in type 2 diabetes. *J. Clin. Invest.* **116**, 1802–1812 (2006).
- Lawlor, N., Khetan, S., Ucar, D. & Stitzel, M. L. Genomics of islet (Dys) function and type 2 diabetes. *Trends Genet.* **33**, 244–255 (2017).
- Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
- Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
- Parker, S. C. J. et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. USA* **110**, 17921–17926 (2013).
- Stitzel, M. L. et al. Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metab.* **12**, 443–455 (2010).
- Khetan, S. et al. Type 2 diabetes-associated genetic variants regulate chromatin accessibility in human islets. *Diabetes* **67**, 2466–2477 (2018).
- Pasquali, L. et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143 (2014).
- Roman, T. S. et al. A type 2 diabetes-associated functional regulatory variant in a pancreatic islet enhancer at the Adcy5 locus. *Diabetes* <https://doi.org/10.2337/db17-0464> (2017).
- Kycia, I. et al. A common type 2 diabetes risk variant potentiates activity of an evolutionarily conserved islet stretch enhancer and increases C2CD4A and C2CD4B expression. *Am. J. Hum. Genet.* **102**, 620–635 (2018).

11. Varshney, A. et al. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc. Natl Acad. Sci. USA* <https://doi.org/10.1073/pnas.1621192114> (2017).
12. Lytrivi, M., Castell, A.-L., Poitout, V. & Cnop, M. Recent insights into mechanisms of  $\beta$ -cell lipo- and glucolipotoxicity in type 2 diabetes. *J. Mol. Biol.* <https://doi.org/10.1016/j.jmb.2019.09.016> (2019).
13. Cnop, M., Toivonen, S., Igoillo-Esteve, M. & Salpea, P. Endoplasmic reticulum stress and eIF2 $\alpha$  phosphorylation: the Achilles heel of pancreatic  $\beta$  cells. *Mol. Metab.* **6**, 1024–1039 (2017).
14. Back, S. H. & Kaufman, R. J. Endoplasmic reticulum stress and type 2 diabetes. *Annu. Rev. Biochem.* **81**, 767–793 (2012).
15. Shrestha, N., Reinert, R. B. & Qi, L. Endoplasmic reticulum protein quality control in  $\beta$  cells. *Semin. Cell Dev. Biol.* **103**, 59–67 (2020).
16. Sharma, R. B. et al. Insulin demand regulates  $\beta$  cell number via the unfolded protein response. *J. Clin. Invest.* **125**, 3831–3846 (2015).
17. Hu, F. B. Sedentary lifestyle and risk of obesity and type 2 diabetes. *Lipids* **38**, 103–108 (2003).
18. Perng, W., Oken, E. & Dabelea, D. Developmental overnutrition and obesity and type 2 diabetes in offspring. *Diabetologia* **62**, 1779–1788 (2019).
19. Gupta, D., Krueger, C. B. & Lastra, G. Over-nutrition, obesity and insulin resistance in the development of  $\beta$ -cell dysfunction. *Curr. Diabetes Rev.* **8**, 76–83 (2012).
20. Jiang, X., Ma, H., Wang, Y. & Liu, Y. Early life factors and type 2 diabetes mellitus. *J. Diabetes Res.* **2013**, 485082 (2013).
21. Teodoro-Morrison, T., Schuiki, I., Zhang, L., Belsham, D. D. & Volchuk, A. GRP78 overproduction in pancreatic beta cells protects against high-fat-diet-induced diabetes in mice. *Diabetologia* **56**, 1057–1067 (2013).
22. Dooley, J. et al. Genetic predisposition for beta cell fragility underlies type 1 and type 2 diabetes. *Nat. Genet.* **48**, 519–527 (2016).
23. Melnikov, A. et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
24. Tewhey, R. et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).
25. Ulirsch, J. C. et al. Systematic functional dissection of common genetic variation affecting red blood. *Cell Traits Cell* **165**, 1530–1545 (2016).
26. Vockley, C. M. et al. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res.* **25**, 1206–1214 (2015).
27. Klein, J. C. et al. Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. *Nat. Commun.* **10**, 2434 (2019).
28. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
29. Fogarty, M. P., Cannon, M. E., Vadlamudi, S., Gaulton, K. J. & Mohlke, K. L. Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMK1D type 2 diabetes GWAS locus. *PLoS Genet.* **10**, e1004633 (2014).
30. Fogarty, M. P., Panhuis, T. M., Vadlamudi, S., Buchkovich, M. L. & Mohlke, K. L. 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 (2013).
31. Kulzer, J. R. et al. A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. *Am. J. Hum. Genet.* <https://doi.org/10.1016/j.ajhg.2013.12.011> (2014).
32. Kitamura, Y. I. et al. FoxO1 protects against pancreatic beta cell failure through NeuroD and MafA induction. *Cell Metab.* **2**, 153–163 (2005).
33. Cinti, F. et al. Evidence of  $\beta$ -cell dedifferentiation in human type 2 diabetes. *J. Clin. Endocrinol. Metab.* **101**, 1044–1054 (2016).
34. Kim-Muller, J. Y. et al. Aldehyde dehydrogenase 1a3 defines a subset of failing pancreatic  $\beta$  cells in diabetic mice. *Nat. Commun.* **7**, 12631 (2016).
35. Buteau, J. & Accili, D. Regulation of pancreatic beta-cell function by the forkhead protein FoxO1. *Diabetes Obes. Metab.* **9**(Suppl 2), 140–146 (2007).
36. Nakae, J. et al. Regulation of insulin action and pancreatic beta-cell function by mutated alleles of the gene encoding forkhead transcription factor Foxo1. *Nat. Genet.* **32**, 245–253 (2002).
37. Peiris, H. et al. Discovering human diabetes-risk gene function with genetics and physiological assays. *Nat. Commun.* **9**, 3855 (2018).
38. Ogihara, T. et al. Liver X receptor agonists augment human islet function through activation of anaplerotic pathways and glycerolipid/free fatty acid cycling. *J. Biol. Chem.* **285**, 5392–5404 (2010).
39. Brun, P.-J. et al. Retinoic acid receptor signaling is required to maintain glucose-stimulated insulin secretion and  $\beta$ -cell mass. *FASEB J.* **29**, 671–683 (2015).
40. Gaulton, K. J. et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).
41. Ackermann, A. M., Wang, Z., Schug, J., Naji, A. & Kaestner, K. H. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Mol. Metab.* **5**, 233–244 (2016).
42. Wesolowska-Andersen, A. et al. Deep learning models predict regulatory variants in pancreatic islets and refine type 2 diabetes association signals. *eLife* **9** (2020).
43. Papa, F. R. Endoplasmic reticulum stress, pancreatic  $\beta$ -cell degeneration, and diabetes. *Cold Spring Harb. Perspect. Med.* **2**, a007666 (2012).
44. Bonnycastle, L. L. et al. Autosomal dominant diabetes arising from a Wolfram syndrome 1 mutation. *Diabetes* **62**, 3943–3950 (2013).
45. Delépine, M. et al. EIF2AK3, encoding translation initiation factor 2-alpha kinase 3, is mutated in patients with Wolcott-Rallison syndrome. *Nat. Genet.* **25**, 406–409 (2000).
46. Booth, C. & Koch, G. L. Perturbation of cellular calcium induces secretion of luminal ER proteins. *Cell* **59**, 729–737 (1989).
47. Sehgal, P. et al. Inhibition of the sarco/endoplasmic reticulum (ER) Ca<sup>2+</sup>-ATPase by thapsigargin analogs induces cell death via ER Ca<sup>2+</sup> depletion and the unfolded protein response. *J. Biol. Chem.* **292**, 19656–19673 (2017).
48. Stone, S. et al. Pancreatic stone protein/regenerating protein is a potential biomarker for endoplasmic reticulum stress in beta cells. *Sci. Rep.* **9**, 5199 (2019).
49. Robbins, R. D. et al. Inhibition of deoxyhypusine synthase enhances islet {beta} cell function and survival in the setting of endoplasmic reticulum stress and type 2 diabetes. *J. Biol. Chem.* **285**, 39943–39952 (2010).
50. Cunha, D. A. et al. Pancreatic  $\beta$ -cell protection from inflammatory stress by the endoplasmic reticulum proteins thrombospondin 1 and mesencephalic astrocyte-derived neurotrophic factor (MANF). *J. Biol. Chem.* **292**, 14977–14988 (2017).
51. Syed, I. et al. PAHSAs attenuate immune responses and promote  $\beta$  cell survival in autoimmune diabetic mice. *J. Clin. Invest.* **129**, 3717–3731 (2019).
52. Chen, Y.-C. et al. PAM haploinsufficiency does not accelerate the development of diet- and human IAPP-induced diabetes in mice. *Diabetologia* **63**, 561–576 (2020).
53. Suetomi, R. et al. Adrenomedullin has a cytoprotective role against ER stress for pancreatic  $\beta$ -cells in autocrine and paracrine manners. *J. Diabetes Investig.* <https://doi.org/10.1111/jdi.13218> (2020).
54. Eizirik, D. L., Pasquali, L. & Cnop, M. Pancreatic  $\beta$ -cells in type 1 and type 2 diabetes mellitus: different pathways to failure. *Nat. Rev. Endocrinol.* **16**, 349–362 (2020).
55. Cullinan, S. B. & Diehl, J. A. Coordination of ER and oxidative stress signaling: the PERK/Nrf2 signaling pathway. *Int. J. Biochem. Cell Biol.* **38**, 317–332 (2006).
56. Oyadomari, S. & Mori, M. Roles of CHOP/GADD153 in endoplasmic reticulum stress. *Cell Death Differ.* **11**, 381–389 (2004).
57. Ghosh, R., Colon-Negron, K. & Papa, F. R. Endoplasmic reticulum stress, degeneration of pancreatic islet  $\beta$ -cells, and therapeutic modulation of the unfolded protein response in diabetes. *Mol. Metab.* **27S**, S60–S68 (2019).
58. Nammo, T. et al. Genome-wide profiling of histone H3K27 acetylation featured fatty acid signalling in pancreatic beta cells in diet-induced obesity in mice. *Diabetologia* **61**, 2608–2620 (2018).
59. Aguayo-Mazzucato, C. et al. T3 induces both markers of maturation and aging in pancreatic  $\beta$ -cells. *Diabetes* **67**, 1322–1331 (2018).
60. Gao, N. et al. Foxa1 and Foxa2 maintain the metabolic and secretory features of the mature beta-cell. *Mol. Endocrinol. Baltim. Md* **24**, 1594–1604 (2010).
61. Fu, Z., Gilbert, E. R. & Liu, D. Regulation of insulin synthesis and secretion and pancreatic beta-cell dysfunction in diabetes. *Curr. Diabetes Rev.* **9**, 25–53 (2013).
62. Hernandez, A. J. et al. B2 and ALU retrotransposons are self-cleaving ribozymes whose activity is enhanced by EZH2. *Proc. Natl Acad. Sci. USA* **117**, 415–425 (2020).
63. Liu, W. M., Chu, W. M., Choudary, P. V. & Schmid, C. W. Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res.* **23**, 1758–1765 (1995).
64. Sundaram, V. & Wang, T. Transposable element mediated innovation in gene regulatory landscapes of cells: re-visiting the ‘gene-battery’ model. *BioEssays News Rev. Mol. Cell. Dev. Biol.* **40** (2018).
65. Deininger, P. Alu elements: know the SINEs. *Genome Biol.* **12**, 236 (2011).
66. Gaulton, K. J. et al. A map of open chromatin in human pancreatic islets. *Nat. Genet.* **42**, 255–259 (2010).
67. Miguel-Escalada, I. et al. Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nat. Genet.* **51**, 1137–1148 (2019).
68. Greenwald, W. W. et al. Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. *Nat. Commun.* **10**, 2078 (2019).
69. Juliana, C. A. et al. A PDX1-ATF transcriptional complex governs  $\beta$  cell survival during stress. *Mol. Metab.* **17**, 39–48 (2018).

70. Ghosh, R. et al. Transcriptional regulation of VEGF-A by the unfolded protein response pathway. *PLoS ONE* **5**, e9575 (2010).
71. Pereira, E. R., Frudd, K., Awad, W. & Hendershot, L. M. Endoplasmic reticulum (ER) stress and hypoxia response pathways interact to potentiate hypoxia-inducible factor 1 (HIF-1) transcriptional activity on targets like vascular endothelial growth factor (VEGF). *J. Biol. Chem.* **289**, 3352–3364 (2014).
72. Pereira, E. R., Liao, N., Neale, G. A. & Hendershot, L. M. Transcriptional and post-transcriptional regulation of proangiogenic factors by the unfolded protein response. *PLoS ONE* **5** (2010).
73. Viñuela, A. et al. Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. *Nat. Commun.* **11**, 4912 (2020).
74. Reschen, M. E. et al. Lipid-induced epigenomic changes in human macrophages identify a coronary artery disease-associated variant that regulates PPAP2B expression through altered C/EBP-beta binding. *PLoS Genet.* **11**, e1005061 (2015).
75. Calderon, D. et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet.* **51**, 1494–1505 (2019).
76. Ramos-Rodríguez, M. et al. The impact of proinflammatory cytokines on the  $\beta$ -cell regulatory landscape provides insights into the genetics of type 1 diabetes. *Nat. Genet.* **51**, 1588–1595 (2019).
77. Vockley, C. M. et al. Direct GR binding sites potentiate clusters of TF binding across the human genome. *Cell* **166**, 1269–1281.e19 (2016).
78. Umans, B. D., Battle, A. & Gilad, Y. Where are the disease-associated eQTLs? *Trends Genet.* <https://doi.org/10.1016/j.tig.2020.08.009> (2020).
79. Ostuni, R. et al. Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**, 157–171 (2013).
80. Kim-Hellmuth, S. et al. Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nat. Commun.* **8**, 266 (2017).
81. Pehrsson, E. C., Choudhary, M. N. K., Sundaram, V. & Wang, T. The epigenomic landscape of transposable elements across normal human development and anatomy. *Nat. Commun.* **10**, 5640 (2019).
82. Su, M., Han, D., Boyd-Kirkup, J., Yu, X. & Han, J.-D. J. Evolution of Alu elements toward enhancers. *Cell Rep.* **7**, 376–385 (2014).
83. Jang, H. S. et al. Transposable elements drive widespread expression of oncogenes in human cancers. *Nat. Genet.* **51**, 611–617 (2019).
84. Franchini, L. F. et al. Convergent evolution of two mammalian neuronal enhancers by sequential exaptation of unrelated retroposons. *Proc. Natl Acad. Sci. USA* **108**, 15270–15275 (2011).
85. Choudhary, M. N. et al. Co-opted transposons help perpetuate conserved higher-order chromosomal structures. *Genome Biol.* **21**, 16 (2020).
86. Sundaram, V. et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* **24**, 1963–1976 (2014).
87. Xie, M. et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat. Genet.* **45**, 836–841 (2013).
88. Sundaram, V. et al. Functional cis-regulatory modules encoded by mouse-specific endogenous retrovirus. *Nat. Commun.* **8**, 14550 (2017).
89. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res* **23**, 169–180 (2013).
90. Li, T., Spearow, J., Rubin, C. M. & Schmid, C. W. Physiological stresses increase mouse short interspersed element (SINE) RNA expression in vivo. *Gene* **239**, 367–372 (1999).
91. Zhang, X.-O., Gingeras, T. R. & Weng, Z. Genome-wide analysis of polymerase III-transcribed Alu elements suggests cell-type-specific enhancer function. *Genome Res.* **29**, 1402–1414 (2019).
92. Bejerano, G. et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**, 87–90 (2006).
93. Mita, P. & Boeke, J. D. How retrotransposons shape genome regulation. *Curr. Opin. Genet. Dev.* **37**, 90–100 (2016).
94. Lawlor, N. et al. Multiomic profiling identifies cis-regulatory networks underlying human pancreatic  $\beta$  cell identity and function. *Cell Rep.* **26**, 788–801.e6 (2019).
95. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
96. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
97. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
98. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* **30**, 2114–2120 (2014).
99. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
100. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
101. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinforma. Oxf. Engl.* **31**, 3847–3849 (2015).
102. Kumar, S., Ambrosini, G. & Bucher, P. SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.* **45**, D139–D144 (2017).

## Acknowledgements

We thank members of the Stitzel and Ucar labs for helpful discussion and critiques during study design and execution and Francis S. Collins, D. Leland Taylor, Cassandra Spracklen, Christine Beck, and Taneli Helenius for helpful comments on the manuscript. This work was supported by The Jackson Laboratory Director's Innovation Fund (to M.L.S. and R.T.), W81XWH-18-0401 (to M.L.S. and D.U.), R01DK118011 (to M.L.S.), and R00HG008179 (to R.T.).

## Author contributions

M.L.S., R.T., S.K., and D.U. designed the study. S.K. and S. Kales designed, generated, and tested the MPRA library. S.K., R.K., and A.J. optimized and completed MPRA experiments in MIN6. S.K. and M.L.S. designed and completed EMSA experiments. S.K., D.U., RT, and M.L.S. analyzed and discussed MPRA results and interpretations. S.K. and S.K.R. completed evolutionary conservation analyses. J.C.U. and R.T. completed GWAS and credible set SNP enrichment analyses. S.K., S.K.R., D.U., R.T., and M.L.S. wrote the manuscript. All authors reviewed and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-25514-6>.

**Correspondence** and requests for materials should be addressed to R.T. or M.L.S.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021