nature genetics

Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants

Lorenzo Pasquali[1,2,14], Kyle J Gaulton[3–5,14], Santiago A Rodríguez-Seguí[1,13,14], Loris Mularoni[1,2], Irene Miguel-Escalada[6], İldem Akerman[1,2], Juan J Tena[7], Ignasi Morán[8], Carlos Gómez-Marín[7], Martijn van de Bunt[3–5], Joan Ponsa-Cobas[8], Natalia Castro[1,2], Takao Nammo[1,13], Inês Cebola[8], Javier García-Hurtado[1,2], Miguel Angel Maestro[1,2], François Pattou[9], Lorenzo Piemonti[10], Thierry Berney[11], Anna L Gloyn[4,5], Philippe Ravassard[12], José Luis Gómez-Skarmeta[7], Ferenc Müller[6], Mark I McCarthy[3–5] & Jorge Ferrer[1,2,8]

Type 2 diabetes affects over 300 million people, causing severe complications and premature death, yet the underlying molecular mechanisms are largely unknown. Pancreatic islet dysfunction is central in type 2 diabetes pathogenesis, and understanding islet genome regulation could therefore provide valuable mechanistic insights. We have now mapped and examined the function of human islet *cis*-regulatory networks. We identify genomic sequences that are targeted by islet transcription factors to drive islet-specific gene activity and show that most such sequences reside in clusters of enhancers that form physical three-dimensional chromatin domains. We find that sequence variants associated with type 2 diabetes and fasting glycemia are enriched in these clustered islet enhancers and identify trait-associated variants that disrupt DNA binding and islet enhancer activity. Our studies illustrate how islet transcription factors interact functionally with the epigenome and provide systematic evidence that the dysregulation of islet enhancers is relevant to the mechanisms underlying type 2 diabetes.

© 2014 Nature America, Inc. All rights reserved.

Despite recent progress in mapping transcriptional regulatory elements and chromatin states[1–7], there is still limited understanding of how transcription factor networks interact with chromatin to control genome function. Furthermore, little is known about how the dysregulation of such networks contributes to disease. As pancreatic islet cells are pivotal in diabetes pathogenesis[8–10] and genomic *cis*-regulatory maps in islets are still markedly underdeveloped compared to those for other cell types[5,7,11–16], we have now mapped chromatin states, binding sites of key transcription factors and transcripts in human pancreatic islets (**Fig. 1a**). Integrative analysis and genetic perturbations were combined to provide reference charts of the *cis*-regulatory elements functioning in human pancreatic islets. Our studies identified enhancer domains that are central determinants of islet-specific gene activity and linked sequence variation in these regions to susceptibility for type 2 diabetes (T2D) and to variation in fasting glycemia levels.

RESULTS

Topology of an essential islet transcription factor network

To characterize the transcription factor networks that control gene activity in human islet cells, we focused on five key β cell transcription factors—FOXA2, MAFB and NKX2.2, which are expressed in insulin-producing β cells and other major islet cell types; NKX6.1, which is specific to β cells; and PDX1, which is present in β cells and more scarce somatostatin-producing islet cells[17–21] (**Supplementary Fig. 1a–c** and **Supplementary Table 1**). Mouse genetic experiments have shown that these five transcription factors are essential for the differentiation or function of β cells[22–24], yet little is known about how these factors function at the level of regulatory networks or how they interact with human islet chromatin to create cell type–specific transcriptional activity. To determine the genomic binding sites of these islet transcription factors, we used chromatin immunoprecipitation and sequencing (ChIP-seq) in duplicate human islet samples

[1]Genomic Programming of Beta-Cells Laboratory, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [2]Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Barcelona, Spain. [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [4]Oxford Centre for Diabetes, Endocrinology and Metabolism, Churchill Hospital, Oxford, UK. [5]Oxford National Institute for Health Research (NIHR) Biomedical Research Centre, Churchill Hospital, Oxford, UK. [6]School of Clinical and Experimental Medicine, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK. [7]Centro Andaluz de Biología del Desarrollo (CABD), Consejo Superior de Investigaciones Científicas-Universidad Pablo de Olavide–Junta de Andalucía, Seville, Spain. [8]Department of Medicine, Imperial College London, London, UK. [9]University of Lille 2, INSERM U859 Biotherapies of Diabetes, Lille, France. [10]Diabetes Research Institute, San Raffaele Scientific Institute, Milan, Italy. [11]Cell Isolation and Transplantation Center, Department of Surgery, Geneva University Hospitals and University of Geneva, Geneva, Switzerland. [12]Centre de Recherche de l'Institut du Cerveau et de la Moelle, Biotechnology and Biotherapy Team, CNRS UMR 7225, INSERM U975, University Pierre et Marie Curie, Paris, France. [13]Present addresses: Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE-Consejo Nacional de Investigaciones Científicas y Técnicas), Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina (S.A.R.-S.) and Department of Metabolic Disorder, Diabetes Research Center, Research Institute, National Center for Global Health and Medicine, Shinjuku-ku, Tokyo, Japan (T.N.). [14]These authors contributed equally to this work. Correspondence should be addressed to J.F. (j.ferrer@imperial.ac.uk).

Received 27 March 2013; accepted 12 December 2013; published online 12 January 2014; doi:10.1038/ng.2870

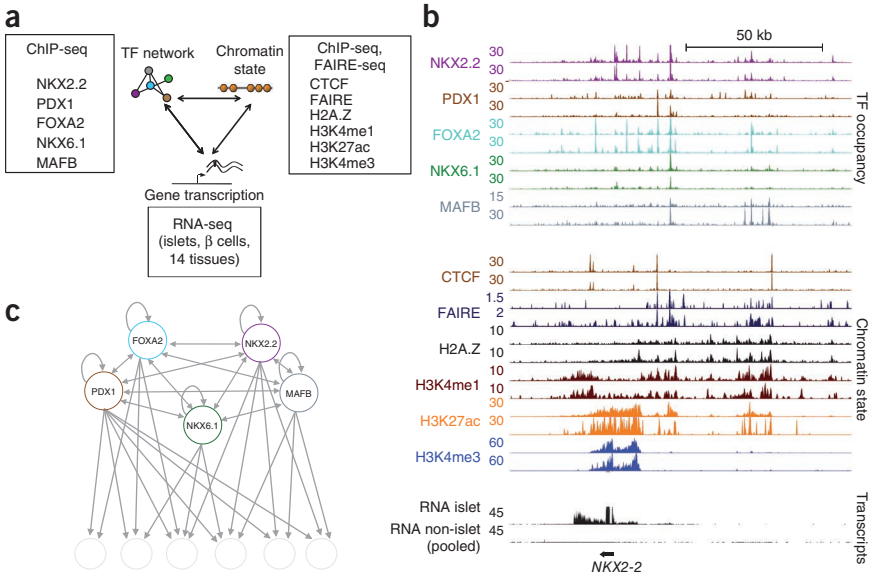

Figure 1 Integrative regulatory maps of human pancreatic islet cells. (**a**) Transcription factor (TF) binding, active chromatin and gene transcription maps in human islet cells. (**b**) Integrative map of the *NKX2-2* locus. Transcription factor binding and chromatin state profiles are shown for duplicate human islet samples. RNA-seq tracks correspond to human islets or pooled data from 14 non-pancreatic tissues are shown highlighting islet-specific transcripts. (**c**) Network topology diagram showing that all five islet-specific transcription factors have direct auto- and cross-regulatory interactions and frequently target adjacent genomic sites (**Supplementary Fig. 2**).

and identified 3,911–32,747 high-confidence sites per transcription factor (**Supplementary Figs. 1d**,**e** and **2a**,**b**). We found that all five transcription factors invariably bound to multiple sites in the vicinity of their own and each other's genes, a pattern indicating auto- and cross-regulatory interactions (**Fig. 1b**,**c** and **Supplementary Fig. 2c**–**f**). Furthermore, all five transcription factors frequently bound to overlapping genomic sites, and, consequently, their genomic binding signals were highly correlated (r = 0.36–0.68 between pairs of islet transcription factors in contrast with r = 0.03–0.11 between islet transcription factors and a control non-islet transcription factor) (**Fig. 1b**,**c** and **Supplementary Fig. 2c**–**h**). This systematic analysis shows that, in agreement with earlier descriptions of transcription factor networks in diverse species and tissues[25,26], human islet-specific transcription factors form a remarkably interconnected network (**Fig. 1c**). These genome-scale maps of binding sites for key islet transcription factors provide a framework for understanding how transcription factors control islet-specific gene transcription.

Islet transcription factors bind to distinct chromatin states

We next sought to understand the relationship between the binding sites of islet-specific transcription factors and underlying chromatin states in human islets. We first mapped all accessible, or open, chromatin sites in pancreatic islets. We combined two assays that capture overlapping, although not identical, sets of accessible chromatin sites: FAIRE-seq, which uses formaldehyde-assisted isolation of regulatory elements to identify open (nucleosome-depleted) sites[5], and ChIP-seq for H2A.Z, a histone variant that is enriched in accessible chromatin sites[27,28]. We defined ~95,000 sites that were enriched in either of these 2 accessible chromatin types in 2 or more human islet samples. Next, we mapped key histone modifications (monomethylation of histone H3 at lysine 4 (H3K4me1), trimethylation of histone H3 at lysine 4 (H3K4me3) and acetylation of histone H3 at lysine 27 (H3K27ac)) and CTCF-binding sites in human islets and used k-median clustering to divide all accessible chromatin sites into five classes (C1–C5) that displayed clearly recognizable enrichment patterns for histone modifications and CTCF binding[1–3], namely, promoters (C1), poised or inactive enhancers (C2), active enhancers (C3), CTCF-bound sites (C4) and additional accessible sites lacking active histone modifications and CTCF occupancy (C5) (**Fig. 2a** and **Supplementary Fig. 3a**–**c**). Consistent with our expectation that transcription factors would bind to accessible DNA, 92% of the genomic sites bound by any of the five islet-specific transcription factors mapped to sites that showed enrichment in FAIRE-seq or H2A.Z ChIP-seq in at least one sample, and islet transcription factors bound to all major classes of accessible chromatin (**Fig. 2b** and **Supplementary Fig. 4**). Integrating histone marks and transcription factor binding maps thus resolved discrete classes of transcription factor–bound accessible chromatin sites in human islets.

Cis-regulatory clusters drive islet-specific transcription

Next, we examined the functional relationship between these different classes of transcription factor–bound, accessible chromatin sites and islet-specific gene transcription. To study islet-specific gene transcription, we analyzed RNA sequencing (RNA-seq) profiles from 3 human islet samples[29], 2 fluorescence-activated cell sorting (FACS)-purified β cell samples[29] and 14 non-pancreatic tissues and used these profiles to define 3 classes of genes: (i) islet-specific genes, which included the 1,000 genes that showed the strongest transcript enrichment in human islet cells, (ii) ubiquitously transcribed genes, which included 1,000 genes that were similarly active in islet cells and other tissues, and (iii) inactive genes, which included the 1,000 genes that showed the lowest expression levels in islet cells. Contrary to our expectation that cell type–specific transcription factors should primarily bind to cell type–specific genes, islet-specific transcription factors bound to promoter (C1) accessible chromatin to a similar extent at both islet-specific and ubiquitously transcribed genes (**Fig. 2c**,**d** and **Supplementary Fig. 5a**,**b**). Similarly, binding events of islet-specific transcription factors at C2, C4 and C5 accessible chromatin sites were not enriched at loci harboring genes specifically expressed in islet cells (**Fig. 2c** and **Supplementary Fig. 5c**). In sharp contrast, gene loci with islet-specific expression had a higher density of islet transcription factor–bound active enhancer (C3) accessible chromatin sites (**Fig. 2c**,**e**). These genes with islet-specific expression had a median of 3 transcription factor–bound C3 sites (interquartile range (IQR) = 1–5) within 50 kb of their transcriptional start site (TSS), whereas ubiquitously transcribed genes had 1 (IQR = 0–3) and islet-inactive genes had a median of 0 (IQR = 0–1) transcription factor–bound C3 sites (Kruskall-Wallis $P < 1 \times 10^{-16}$). Consistent with studies showing that combinatorial transcription factor interactions are critical for enhancer activation[30,31], 72.8% of genomic sites bound by three or more islet-specific transcription factors were associated with C3 accessible chromatin (**Fig. 2f**). Although some promoter (C1) accessible chromatin sites were also bound by multiple islet transcription factors, these sites were not associated with islet-specific transcription, suggesting that the link between transcription

© 2014 Nature America, Inc. All rights reserved.

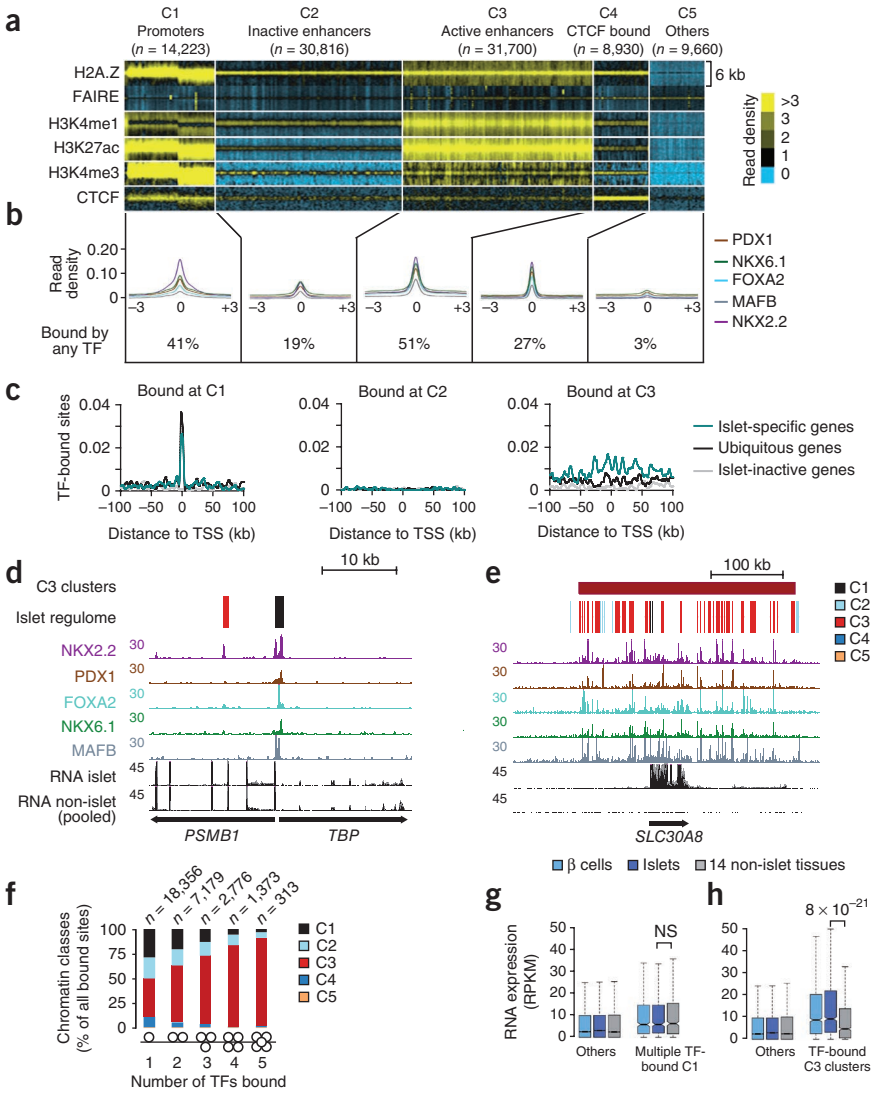

Figure 2 Transcription factor networks establish distinct types of interactions with the epigenome. (**a**) *k*-median clustering of 95,329 islet accessible chromatin sites (defined by FAIRE and/or H2A.Z enrichment in 2 samples) showing 5 classes of accessible chromatin (C1–C5) that we refer to as promoters, inactive enhancers, active enhancers, CTCF-bound sites and other, on the basis of previously defined patterns of histone modification and CTCF binding[1–3] (**Supplementary Fig. 3a**). Histone- and CTCF-bound reads were divided into 100-bp bins across 6-kb windows centered on merged FAIRE- or H2A.Z-enriched sites. (**b**) Average transcription factor binding signal distribution relative to the center of each accessible chromatin class and the percentage of sites in each class bound by at least one transcription factor. Non-quantile-normalized reads were processed as described for **a**. (**c**) Density of C1, C2 and C3 accessible chromatin sites bound by 2 or more transcription factors in the regions surrounding the TSSs of the 1,000 genes with the most islet-specific expression, 1,000 ubiquitously active genes and 1,000 islet-inactive genes. (**d**,**e**) Examples of islet-specific transcription factor binding to the 5′ ends of the ubiquitously active genes *TBP* and *PSMB1* (**d**) or to the islet-specific T2D susceptibility gene *SLC30A8* (**e**). The regulome track depicts colored chromatin states for classes C1–C5. (**f**) Binding by multiple transcription factors is more common at C3 (active enhancer) chromatin. The numbers of transcription factor–bound sites consistent in two samples for each category are shown above. (**g**) Binding by three or more transcription factors at C1 (promoter) chromatin is not associated with islet-specific activity of the adjacent gene. NS, not significant. (**h**) Genes located <25 kbfrom clusters of C3 sites that are highly bound by transcription factors show enriched expression in islets and β cells relative to 14 non-islet tissues. Boxes show IQR, notches indicate 95% confidence intervals of the median, and whiskers extend to either 1.5 times the IQR or to extreme values. Gene expression data from non-pancreatic tissues in **g**,**h** are presented by individual tissue in **Supplementary Figures 5d** and **7b**. *P* values were calculated with the Wilcoxon rank-sum test.

© 2014 Nature America, Inc. All rights reserved.

factor occupancy and cell type–specific transcription is largely confined to active enhancers (**Fig. 2g** and **Supplementary Fig. 5d**). Taken together, these findings showed remarkable differences between transcription factor binding at sites in the different accessible chromatin classes. These findings suggest that certain binding events of islet-specific transcription factors, such as those occurring at promoters, are not generally associated with cell type–specific gene transcription, which is instead tightly linked to clusters of active enhancer sites bound by multiple transcription factors.

We scanned the genome to identify all clusters of three or more enhancer sites active in islets (Online Methods and **Supplementary Fig. 6a**,**b**). We identified 3,677 such clusters, which mapped near genes with strong islet-enriched expression (median expression in islets of 8.3 reads per kilobase per million mapped reads (RPKM) (IQR = 2.9–19.6) and in non-islet tissues of 4.8 RPKM (IQR = 0.7–14.2); Wilcoxon $P < 1 \times 10^{-30}$), in contrast to more modest islet enrichment in genes near non-clustered (orphan) enhancers (median expression in islets of 6.0 RPKM (IQR = 1.5–14.3) and in non-islet tissues of 5.0 RPKM (IQR = 0.8–13.9); Wilcoxon $P < 1 \times 10^{-5}$) (**Supplementary Fig. 6c**). Islet-enriched transcription was most pronounced near the 1,813 enhancer clusters that showed higher than median occupancy by islet transcription factors (**Fig. 2h** and **Supplementary Fig. 7a–c**). Remarkably, most genes that are currently known to be important for islet cell identity, function or disease were associated with an islet enhancer cluster (92% of a manually annotated list of 65 such genes, 90% of which belonged to the subset with high transcription factor occupancy; **Supplementary Fig. 7d** and **Supplementary Table 2**). These findings reinforce the view that the genomic program that underlies islet cell identity is tightly associated with clusters of enhancers bound by islet transcription factors.

To investigate the function of transcription factor–bound enhancer clusters, we first verified that individual clustered C3 sites truly acted as cell type–specific enhancers. Consistent with their chromatin signature, 8 of 12 transcription factor–bound active enhancer C3 sites (but not transcription factor–bound C2 or C5 sites) displayed β cell–selective enhancer activity in episomal reporter assays in mouse cell lines (**Fig. 3a**). Five conserved transcription factor–bound C3 sites were tested in transgenic zebrafish assays, and all exhibited

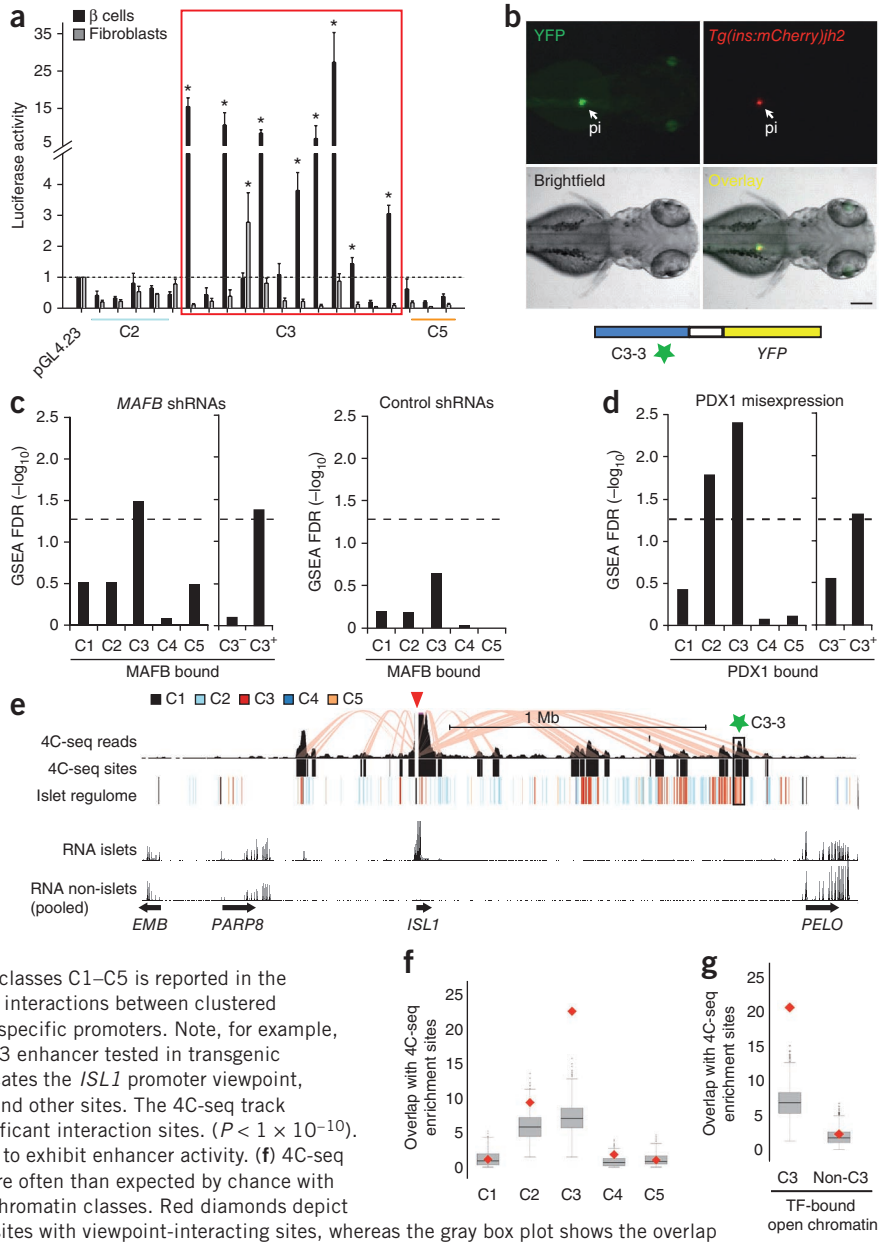

Figure 3 Enhancer clusters form functional three-dimensional chromatin domains. (**a**) Luciferase assays in mouse MIN6 β cells and NIH3T3 fibroblasts showing that 8 of 12 transcription factor–bound C3 sites (red box) but not transcription factor–bound C2 or C5 sites confer cell type–specific expression to a minimal promoter (*$P < 0.005$, *t* test for comparison with empty vector; $n = 3$ transfections per condition). Data are presented as mean ± s.d. (**b**) C3-3, a C3 element in a cluster >1 Mb from *ISL1*, shows selective enhancer activity in pancreatic islet from zebrafish embryos 70 h post-fertilization. The enhancer transgene (*YFP*) was injected into a transgenic line that exhibits fluorescence (mCherry) in insulin-positive cells. Scale bar, 0.1 mm. (**c**) *MAFB* knockdown in human EndoC-β H1 β cells causes the downregulation of genes bound by MAFB at clustered ($C3^+$) rather than orphan ($C3^-$) MAFB-bound enhancers. Bar plots show GSEA FDRs for different MAFB-bound gene sets for the genes that are downregulated after infecting cells with viruses expressing two short hairpin RNAs (shRNAs) for *MAFB* versus four non-targeting shRNAs. As a control, we repeated the analysis using the same number of arrays but compared sets of different control shRNAs. Horizontal dashed lines signify FDR = 0.05 as a reference. (See also **Supplementary Fig. 9a–e**.) (**d**) Misexpression of PDX1 with MAFA and NGN3 in HEK293T cells preferentially activates genes associated with PDX1-bound clustered ($C3^+$) enhancers but not genes bound by PDX1 at promoter accessible chromatin sites or those associated with orphan ($C3^-$) PDX1-bound enhancers. The number of transcription factor–bound genes for classes C1–C5 is reported in the Online Methods. (**e**) 4C-seq analysis shows selective interactions between clustered transcription factor–bound enhancers and cell type–specific promoters. Note, for example, interactions between the *ISL1* promoter and the C3-3 enhancer tested in transgenic zebrafish located >1 Mb away. The red triangle indicates the *ISL1* promoter viewpoint, and pink lines highlight interactions between *ISL1* and other sites. The 4C-seq track shows aligned sequences, and 4C-seq sites are significant interaction sites. ($P < 1 \times 10^{-10}$). A green star marks region C3-3, which is shown in **b** to exhibit enhancer activity. (**f**) 4C-seq analysis of nine loci shows that C3 sites interact more often than expected by chance with nearby promoters in contrast with other accessible chromatin classes. Red diamonds depict the mean overlap of different accessible chromatin sites with viewpoint-interacting sites, whereas the gray box plot shows the overlap of randomized accessible chromatin sites (classes C1–C5) after 1,000 iterations at the same loci. (**g**) Transcription factor–bound C3 but not other transcription factor–bound sites establish frequent interactions with nearby promoters. Boxes show IQR, and whiskers extend to 1.5 times the IQR.

enhancer activity (three specifically in the pancreatic islet), whereas transcription factor–bound C2 and C5 sites showed no activity (**Fig. 3b** and **Supplementary Fig. 8a–c**). These experiments show that clustered sites have islet enhancer activity.

We next designed experiments to directly test whether transcription factor binding sites that were present in clusters of enhancers, but not necessarily other transcription factor binding sites, were functionally important for the activation of islet-selective genes. First, we transduced human β cells (EndoC-β H1) with two independent interfering hairpin RNAs for *MAFB* or with four control hairpins (**Supplementary Fig. 9a**). Knockdown of *MAFB* caused downregulation of genes linked to MAFB-bound enhancer clusters (Gene Set Enrichment Analysis (GSEA) normalized enrichment score (NES) = 1.98, false discovery rate (FDR) $q < 1.5 \times 10^{-2}$) (**Fig. 3c**), such as *ROBO2* and *G6PC2* (**Supplementary Fig. 9b**,**c**). However, the expression of genes bound by MAFB only at their promoters or at other classes of accessible chromatin was not significantly affected (**Fig. 3c** and **Supplementary Fig. 9d**,**e**). Next, we transfected HEK293T embryonic kidney cells with expression vectors encoding PDX1 and assessed which PDX1-bound genes were activated. We coexpressed PDX1 with MAFA and NGN3 because this transcription factor combination, which has previously been employed to activate β cell genes[32], was more efficient than PDX1 alone in gene activation in heterologous cells. We observed transcriptional enhancement of genes near PDX1-bound enhancer clusters (GSEA NES = 2.2, FDR $q < 1 \times 10^{-3}$), whereas the expression of genes that were exclusively bound by PDX1 at promoter or other accessible chromatin classes was unaffected (GSEA NES = −1.3, FDR $q = 0.37$) (**Fig. 3d**). These studies confirm that chromatin profiling can identify subsets of transcription

© 2014 Nature America, Inc. All rights reserved.

factor binding sites that are important for cell type–specific gene activity.

Having shown that active enhancer clusters are pervasive at genes with islet-specific expression and, furthermore, that these clusters are critical for islet-specific gene activity, we examined the relationship between enhancer multiplicity and the formation of broad three-dimensional chromatin structures. We selected nine loci, eight of which contained active enhancer clusters near a gene with islet-enriched expression, and used human islet chromatin to perform circular chromosome conformation capture coupled with high-throughput sequencing (4C-seq) (**Fig. 3e**)[33]. Transcription factor–bound C3 sites from these clusters displayed frequent, strong interactions with the promoters of genes at the same loci with islet-enriched expression, including C3 sites located >1 Mb from their target promoter (**Fig. 3e–g**; additional loci are shown in **Supplementary Fig. 10a–e**). In contrast, transcription factor–bound non-C3 sites at the same loci did not show greater interactions than expected by chance ($P < 1 \times 10^{-3}$ for transcription factor–bound C3 sites versus randomized sites and $P = 0.41$ for transcription factor–bound non-C3 sites versus randomized sites) (**Fig. 3f,g**). These observations suggest that islet enhancer clusters are three-dimensional structural units in which multiple transcription factor–bound active enhancers (rather than other transcription factor–bound sites) interact with target promoters. Collectively, the results indicate that human islet-specific gene transcription is largely driven by combinatorial transcription factor binding at clusters of enhancers that form three-dimensional chromatin structures.

Islet *cis*-regulatory sequence code

The identification of the genomic sites that underlie islet gene activity allowed us to explore the sequence code that is recognized by islet transcription factor networks to drive cell type–specific transcription. We identified 46 highly enriched sequence motifs in clustered enhancers, including recognition sequences for key β cell transcription factors (HNF1A, RFX, FOXA, NEUROD1, NKX6.1, PDX1, MAFA and MAFB), as well as previously unrecognized motifs (**Fig. 4** and **Supplementary Table 3**). We further identified combinations of these 46 motifs that were most enriched in islet enhancers relative to enhancers from 9 non-pancreatic cell types[34] (**Supplementary Fig. 11a–c**). Motifs for RFX[35] or the pioneer FOXA factors[36] were present in all of the most islet-enriched combinations (**Supplementary Fig. 11b**), consistent with a major role for these factors in the activation of islet enhancers. To independently validate this analysis, we mapped all instances of the ten most enriched motif combinations to the mouse genome and found that they were also located near genes that show islet-enriched expression in mice (median expression of nearby genes in mouse islets versus non-islet tissues of 1.0 (IQR = 0.0–5.2) RPKM versus 0.6 (IQR = 0.0–3.7) RPKM; Wilcoxon $P = 9.6 \times 10^{-31}$) (**Supplementary Fig. 11d**). *In silico* analysis thus expanded the range of candidate combinatorial islet transcription factor interactions and provides a resource to discover new transcription factors and upstream signaling pathways that control β cell transcriptional programs.

Sequence variation in islet enhancers is associated with T2D

Genome-wide association studies (GWAS) have identified dozens of loci associated with T2D and glycemic traits[37,38], but the molecular mechanisms linking specific alleles to cellular functions remain poorly described. The catalog of islet *cis*-regulatory elements allowed us to explore to what extent genetic variants that contribute to T2D susceptibility and to variation in fasting glycemia in non-diabetic individuals act through islet regulatory mechanisms.

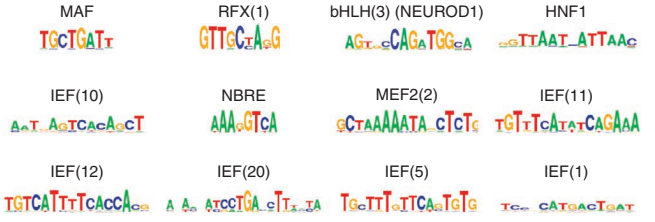

Figure 4 Known and new transcription factor motifs are enriched in clustered islet enhancers. Examples are shown of sequence motifs that are enriched in clustered enhancers at $P < 1 \times 10^{-60}$ (HOMER)[50]. Several motifs match known islet transcription factor recognition sequences, whereas others are candidate binding sites for new regulators. A complete list of motifs that showed enrichment at $P < 1 \times 10^{-60}$ (HOMER) is given in **Supplementary Table 3**. The name of each motif corresponds to the annotations shown in **Supplementary Table 3**. IEF, islet-enriched factor.

First, we examined loci with genome-wide significant association with T2D or fasting glycemia in European populations, considering all variants in high linkage disequilibrium (LD; 1000 Genomes Project CEU (Utah residents of Northern and Western European ancestry) $r^2 > 0.8$) with a lead SNP reported in the National Human Genome Research Institute (NHGRI) GWAS catalog[39]. These associated variants were enriched in C3 sites and occurred more prominently in clustered C3 rather than in orphan C3 sites (T2D: C3 $P = 0.004$; clustered C3 $P = 0.012$, orphan C3 $P = 0.20$; fasting glycemia: C3 $P < 1 \times 10^{-5}$; clustered C3 $P = 3 \times 10^{-5}$, orphan C3 $P = 0.12$) (**Fig. 5a**). There was no significant enrichment in variants associated with T2D or fasting glycemia for other accessible chromatin classes, and variants associated with other complex traits were not enriched at clustered islet C3 sites (all traits $P > 0.05$). Moreover, little enrichment was seen in the 200-bp sites immediately flanking C3 sites (**Fig. 5a**; T2D $P = 0.44$, fasting glycemia $P = 0.043$), suggesting that C3 enrichment is not driven solely by the genomic context of these sites. These findings indicate that a subset of loci with genome-wide association with T2D and fasting glycemia are likely to harbor functional *cis*-regulatory variants in islet active enhancers.

To further examine the contribution of islet *cis*-regulatory variation to T2D and fasting glycemia, we considered association results for the full set of ~2.5 million variants represented in the largest available genome-wide association meta-analyses for these traits rather than restricting analysis to signals reaching stringent genome-wide significance thresholds[37,38]. For each islet accessible chromatin class, we identified the overlapping variants that exceeded a series of association significance thresholds and calculated fold enrichment values using a null distribution of permuted variant counts (Online Methods). Variants in C3 sites were collectively enriched for T2D and fasting glycemia genome-wide association values of $P < 0.001$ (T2D, C3 $P = 2.1 \times 10^{-11}$; fasting glycemia, C3 $P < 1 \times 10^{-16}$). This enrichment was predominantly a feature of C3 clusters rather than of orphan C3 sites (T2D: clustered C3 $P = 5.1 \times 10^{-11}$, orphan C3 $P = 0.006$; fasting glycemia: clustered C3 $P < 1 \times 10^{-11}$, orphan C3 $P = 0.23$) (**Fig. 5b** and **Supplementary Fig. 12**). Fold enrichment patterns were unchanged when the genome-wide association data were aggressively pruned ($r^2 > 0.2$) to retain only the most associated signal in each LD cluster and remained significant for variants associated with T2D and fasting glycemia in clustered C3 sites (T2D: C3 $P = 0.2$, clustered C3 $P = 0.031$; fasting glycemia: C3 $P = 3.6 \times 10^{-6}$, clustered C3 $P = 5.0 \times 10^{-5}$). Enrichment was also retained when established (genome-wide significant)

© 2014 Nature America, Inc. All rights reserved.

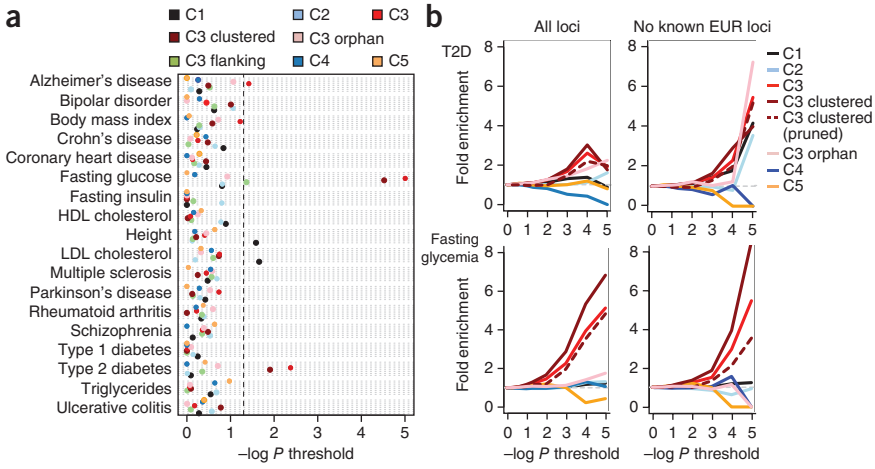

Figure 5 Islet enhancers are enriched in loci associated with T2D and fasting glycemia. (**a**) Loci associated with complex traits were tested for enrichment of overlap with classes of islet sites (C1–C5, clustered C3 and orphan C3) compared to matched background loci. Loci associated with T2D and fasting glycemia were enriched for islet active enhancer sites (C3) but not for directly flanking sites (C3 flanking) or for other islet classes. The enrichment of C3 sites for both T2D and fasting glycemia was stronger when considering clustered rather than orphan (non-clustered) enhancers. The most significant enrichments for other complex diseases were found for C1 sites with low-density lipoprotein (LDL) cholesterol ($P = 0.022$) and height ($P = 0.026$) and for C3 sites with Alzheimer's disease ($P = 0.038$). No significant enrichment of variants associated with other complex traits was seen in clustered C3 enhancers. HDL, high-density lipoprotein. The horizontal dashed line shows P value = 0.05 as a reference. (**b**) Enrichment of islet regulome sites in T2D[38] and fasting glycemia[37] genome-wide association data. We determined the number of variants overlapping sites from each islet accessible chromatin class that surpassed a series of association significance thresholds for T2D and fasting glycemia. We then calculated fold enrichment values at each threshold by comparison to the number of matched background variants significant at that threshold. C3 sites and clustered C3 sites were more enriched for T2D and fasting glycemia association at increasing P-value thresholds (left), even after removing known European (EUR) T2D and fasting glycemia loci (right). Patterns in C3 and clustered C3 sites were maintained when pruning variants to retain a single variant in each LD block ($r^2 > 0.2$). In all such analyses, clustered C3 sites displayed significant enrichment relative to the null distribution of permuted counts for T2D and fasting glycemia association P values of <0.001 (P values for these analyses are given in the text).

T2D and fasting glycemia loci were removed (**Fig. 5b**; T2D: C3 $P = 7.3 \times 10^{-5}$, clustered C3 $P = 1.1 \times 10^{-5}$; fasting glycemia: C3 $P = 1.1 \times 10^{-5}$, clustered C3 $P = 4.1 \times 10^{-9}$), suggesting that the association of variants at C3 sites (and, in the T2D analysis, also of variants at C1 sites) extends to new loci that have yet to be confirmed to show genome-wide significant association with these traits.

To further examine this point, we sought examples of variants overlapping islet accessible chromatin sites that mapped within such moderately associated loci. We found that rs72695654 disrupted a sequence motif and abolished binding of a β cell protein complex to a C3 site that is located in an enhancer cluster within the *ACSL1* gene, which encodes a long-chain fatty acyl-CoA synthase that is highly expressed in islets (**Supplementary Fig. 13a–c**). Long-chain fatty acyl-CoA synthesis has been implicated in β cell stimulus-secretion coupling, survival and lipotoxicity[40–42]. The rs72695654 variant was in high LD (1000 Genomes Project CEU $r^2 = 0.94$) with rs735949, which is strongly associated with T2D ($P = 3.7 \times 10^{-6}$) and, independently, with

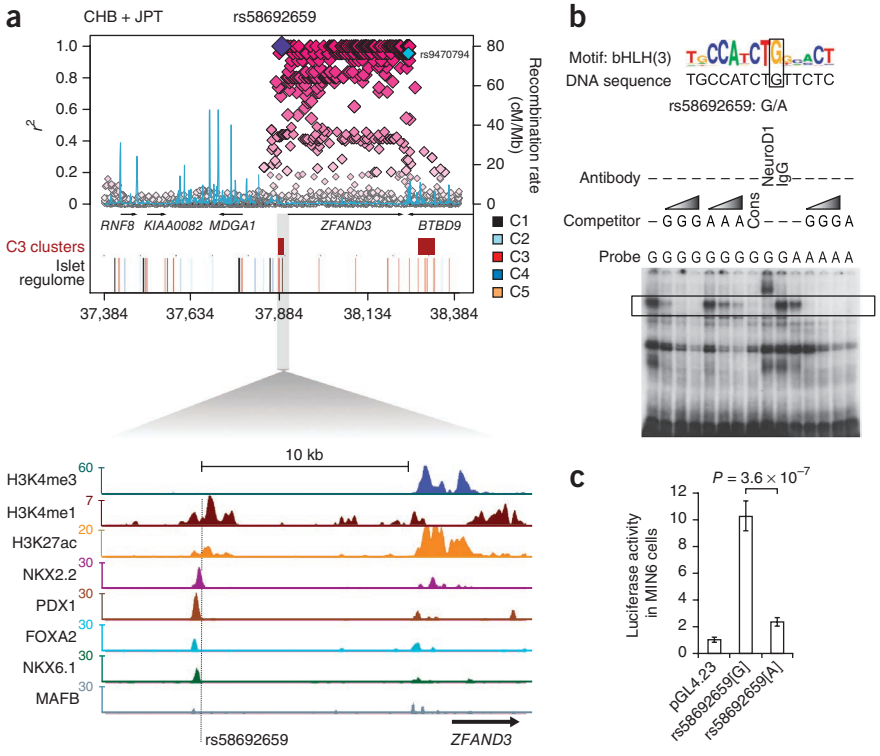

Figure 6 A T2D risk variant at *ZFAND3* disrupts islet enhancer activity. (**a**) Regional plot of 1000 Genomes Project variants and islet accessible chromatin elements at the *ZFAND3* locus associated with T2D in East Asian individuals (r^2 values based on 1000 Genomes Project CHB and JPT LD with rs58692659). rs9470794 (cyan) at this locus shows the strongest T2D association from Cho *et al.*[44]. The rs58692659 SNP maps to a C3 site >10 kb upstream of *ZFAND3* bound by PDX1, FOXA2, NKX2.2 and NKX6.1. (**b**) The rs58692659 SNP disrupts an islet-enriched bHLH-like motif that matches the recognition site for NEUROD1 (ref. 51), a known and important islet transcription factor[22]. The electrophoretic mobility shift assay (EMSA) shows that the minor A allele at this variant abolishes binding of a protein complex that is supershifted by antibody to NEUROD1. The black box highlights the sequence-specific retardation complex. Competition gradients, represented by gray triangles, correspond to 5-fold, 50-fold and 100-fold excess of cold probe. (**c**) Luciferase assays showing reduced enhancer activity of the A allele compared to the G allele of rs58692659 in MIN6 β cells. Data are presented as mean ± s.d. Three independent experiments were performed in triplicate. P values were calculated by two-sided t test.

© 2014 Nature America, Inc. All rights reserved.

fasting glycemia levels in non-diabetic individuals in large genome-wide association meta-analyses ($P = 1.6 \times 10^{-5}$)[37,38] (**Supplementary Fig. 13a**). This finding suggests that *cis*-regulatory maps can be used to prioritize loci on the basis of which genetic evidence is strong, although not genome-wide significant, to guide hypothesis-driven experiments that define susceptibility variants.

We next attempted to define functional enhancer variants in established genome-wide associated loci that might be causal in T2D susceptibility. We first catalogued all variants that resided within C3 sites and were in high LD with established genome-wide significant T2D and fasting glycemia association signals (**Supplementary Tables 4–6**)[37,38]. We confirmed that the *TCF7L2* intronic SNP[43] rs7903146, previously shown to be located in an islet FAIRE-enriched site and to affect chromatin state and enhancer function[5], mapped to a C3 site that is bound by NKX2.2, FOXA2 and MAFB, yet did not show active chromatin marks in an extensive panel of non-islet human cell types (**Supplementary Fig. 14a**,**b**). Similar observations were made at multiple established T2D and fasting glycemia regions, including at *SLC2A2*, *CDKN2A*, *C2CD4A*, *C2CD4B*, *SLC30A8*, *DGKB* and *PCSK1*, where we identified associated variants that mapped to transcription factor–bound C3 sites, many of which disrupted islet-enriched sequence motifs (see **Supplementary Fig. 15a**,**b**, which describes a public browser that facilitates the visualization of T2D and fasting glycemia genome-wide associated variants and the islet regulome, and **Supplementary Fig. 16**). For example, at the *ZFAND3* T2D locus discovered in East Asians[44], SNP rs58692659 formed part of an array of variants that was highly correlated with the reported lead SNP (rs9470794) (1000 Genomes Project CHB (Han Chinese in Beijing, China) and JPT (Japanese in Tokyo, Japan) $r^2 = 0.96$) and mapped to a C3 element bound by multiple islet transcription factors within an enhancer cluster (**Fig. 6a**). This variant altered sequence-specific DNA binding of a key islet-enriched transcription factor, NEUROD1, and abolished enhancer activity in β cells (**Fig. 6b**,**c**). We note that this SNP, along with many others in the region, is monomorphic in European samples (1000 Genomes Project CEU) and that there is no T2D association at *ZFAND3* in Europeans (rs9470794 DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) $P = 0.8$)[44], consistent with rs58692659 having a causal role at this locus. These results show how islet accessible chromatin maps can pinpoint functional *cis*-regulatory variants that are strong candidates for having a causal role in driving T2D association signals.

DISCUSSION

Our work provides reference *cis*-regulatory maps for ongoing efforts to dissect the transcriptional program of pancreatic β cells[45] and the molecular mechanisms of human T2D[10]. We show that islet-specific transcription factors establish widespread binding to accessible chromatin sites where they are apparently not engaged in regulation in *cis* of cell type–specific gene transcription. We further demonstrate that binding events that do drive the transcription of genes underlying islet cell identity reside in clusters of transcriptional enhancers. These clusters mirror previously reported clusters of accessible chromatin in human islets[5], and, although they constitute groups of discrete enhancers, they share many features and may thus represent the same phenomenon as recently reported for large enhancers (superenhancers)[46,47]. Our 4C-seq analysis allows us to speculate that enhancer multiplicity serves an architectural role by creating active chromatin structures in genomic domains that are transcriptionally silent in most non-islet cell types. These findings have therefore uncovered central *cis*-regulatory determinants of islet cell gene transcription.

Our systematic analysis implicates sequence variation at islet enhancer clusters in susceptibility to T2D and in variation in fasting glycemia levels. Recently, SNPs associated with common polygenic diseases were shown to be enriched in noncoding genomic elements defined by DNase I hypersensitivity, which marks accessible chromatin regions such as active promoters, insulators, repressors, and poised and active enhancers[48], or by H3K4me3, which is enriched at promoters and, more weakly, at enhancers[49]. We have now defined for the first time, to our knowledge, functionally distinct transcription factor–bound genomic sites in human islets and have thereby disclosed that T2D susceptibility is specifically associated with allelic variation in pancreatic islet distant enhancers. Our results therefore link islet *cis*-regulatory networks to the mechanisms underlying T2D susceptibility and variation in glycemic traits. The availability of integrated *cis*-regulatory maps in pancreatic islets will facilitate hypothesis-driven experiments to establish the exact manner in which common and lower frequency genetic variants affect islet cells in human diabetes.

The islet regulome, including information on transcription factor occupancy, chromatin states, motifs, enhancer clusters and genome-wide significant *P* values for association with T2D and fasting glycemia, can be accessed and visualized online (see URLs and **Supplementary Fig. 15** for a description of this browser).

URLs. Islet Regulome Browser, http://www.isletregulome.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Raw data are available at ArrayExpress under accession E-MTAB-1919. Illumina human BodyMap2 data are available from ArrayExpress under accession E-MTAB-513.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J. Rios (IDIBAPS) for expert statistical advice, M. Parsons (Johns Hopkins University) for *Tg*(*ins*:*mCherry*)*jh2* transgenics and R. Stein (Vanderbilt University), J. Habener (Harvard University) and G. Gradwohl (Institute of Genetics and Molecular and Cellular Biology) for MAFA, IDX1, and NGN3 DNA constructs. We thank the DIAGRAM and MAGIC consortia, the Singapore Prospective Study Program, the Singapore Consortium of Cohort Studies, the Singapore Indian Eye Study, the Singapore Malay Eye Study and Y.Y. Teo, E.S. Tai, T.Y. Wong, W.Y. Lim and X. Wang (National University of Singapore; funded by the National Medical Research Council of Singapore, Singapore Translational Researcher Award schemes, the Biomedical Research Council of Singapore and the National Research Foundation (NRF) Fellowship scheme). This work was carried out in part at the Centre Esther Koplowitz. This work was funded by grants from a European Foundation for the Study of Diabetes Lilly fellowship (L. Pasquali), the Ministerio de Economía y Competitividad (SAF2011-27086 to J.F., BFU2010-14839 and CSD2007-00008 to J.L.G.S.), the Innovative Medicines Initiative (DIRECT to M.I.M. and J.F.), the Andalusian Government (CVI-3488 to J.L.G.S.), the Biology of Liver and Pancreatic Development and Disease Marie Curie Initial Training Network (F.M. and J.F.), the Wellcome Trust (090532, 098381 and 090367 to M.I.M., 095101 to A.L.G., 101033 to J.F.), Juvenile Diabetes Research Foundation (31-2012-783 to T.B., F.P. and L. Piemonti) and Framework Programme 7 (HEALTH-F4-2007-201413 to M.I.M.).

AUTHOR CONTRIBUTIONS

L. Pasquali, K.J.G., J.L.G.S., I.M.-E., F.M., M.I.M. and J.F. analyzed integrated data and wrote and edited the manuscript, which all authors have approved. İ.A., S.A.R.-S., J.P.-C., J.G.-H., T.N., I.M.-E., C.G.-M., I.C., N.C., M.A.M. and J.J.T. performed and analyzed experiments. L. Pasquali, K.J.G., L.M., J.P.-C. and I.M. performed

© 2014 Nature America, Inc. All rights reserved.

computational analysis. A.L.G., M.v.d.B., F.P., L. Piemonti, T.B. and P.R. provided materials and reagents. L. Pasquali and J.F. conceived and coordinated the project.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
2. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
3. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA* **107**, 21931–21936 (2010).
4. Bonn, S. *et al.* Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.* **44**, 148–156 (2012).
5. Gaulton, K.J. *et al.* A map of open chromatin in human pancreatic islets. *Nat. Genet.* **42**, 255–259 (2010).
6. Shen, Y. *et al.* A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488**, 116–120 (2012).
7. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
8. Kahn, S.E. Clinical review 135: the importance of β-cell failure in the development and progression of type 2 diabetes. *J. Clin. Endocrinol. Metab.* **86**, 4047–4058 (2001).
9. Lyssenko, V. *et al.* Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N. Engl. J. Med.* **359**, 2220–2232 (2008).
10. Ashcroft, F.M. & Rorsman, P. Diabetes mellitus and the β cell: the last ten years. *Cell* **148**, 1160–1171 (2012).
11. Bhandare, R. *et al.* Genome-wide analysis of histone modifications in human pancreatic islets. *Genome Res.* **20**, 428–433 (2010).
12. 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).
13. Khoo, C. *et al.* Research resource: the Pdx1 cistrome of pancreatic islets. *Mol. Endocrinol.* **26**, 521–533 (2012).
14. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
15. Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
16. Tennant, B.R. *et al.* Identification and analysis of murine pancreatic islet enhancers. *Diabetologia* **56**, 542–552 (2013).
17. Artner, I. *et al.* MafB is required for islet β cell maturation. *Proc. Natl. Acad. Sci. USA* **104**, 3853–3858 (2007).
18. Ahlgren, U., Jonsson, J., Jonsson, L., Simu, K. & Edlund, H. β-cell-specific inactivation of the mouse *Ipf1/Pdx1* gene results in loss of the β-cell phenotype and maturity onset diabetes. *Genes Dev.* **12**, 1763–1768 (1998).
19. Sander, M. *et al.* Homeobox gene *Nkx6.1* lies downstream of *Nkx2.2* in the major pathway of β-cell formation in the pancreas. *Development* **127**, 5533–5540 (2000).
20. Sund, N.J. *et al.* Tissue-specific deletion of *Foxa2* in pancreatic β cells results in hyperinsulinemic hypoglycemia. *Genes Dev.* **15**, 1706–1715 (2001).
21. Sussel, L. *et al.* Mice lacking the homeodomain transcription factor Nkx2.2 have diabetes due to arrested differentiation of pancreatic β cells. *Development* **125**, 2213–2221 (1998).
22. Servitja, J.M. & Ferrer, J. Transcriptional networks controlling pancreatic development and β cell function. *Diabetologia* **47**, 597–613 (2004).
23. Wilson, M.E., Scheel, D. & German, M.S. Gene expression cascades in pancreatic development. *Mech. Dev.* **120**, 65–80 (2003).
24. Oliver-Krasinski, J.M. & Stoffers, D.A. On the origin of the β cell. *Genes Dev.* **22**, 1998–2021 (2008).
25. Gerstein, M.B. *et al.* Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330**, 1775–1787 (2010).
26. Kyrmizi, I. *et al.* Plasticity and expanding complexity of the hepatic transcription factor network during liver development. *Genes Dev.* **20**, 2293–2305 (2006).
27. Jin, C. *et al.* H3.3/H2A.Z double variant–containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nat. Genet.* **41**, 941–945 (2009).
28. Ong, C.T. & Corces, V.G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* **12**, 283–293 (2011).
29. Morán, I. *et al.* Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metab.* **16**, 435–448 (2012).
30. Stefflova, K. *et al.* Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell* **154**, 530–540 (2013).
31. Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E.E. Combinatorial binding predicts spatio-temporal *cis*-regulatory activity. *Nature* **462**, 65–70 (2009).
32. Zhou, Q., Brown, J., Kanarek, A., Rajagopal, J. & Melton, D.A. *In vivo* reprogramming of adult pancreatic exocrine cells to β-cells. *Nature* **455**, 627–632 (2008).
33. Noordermeer, D. *et al.* The dynamic architecture of *Hox* gene clusters. *Science* **334**, 222–225 (2011).
34. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
35. Smith, S.B. *et al.* Rfx6 directs islet formation and insulin production in mice and humans. *Nature* **463**, 775–780 (2010).
36. Zaret, K.S. *et al.* Pioneer factors, genetic competence, and inductive signaling: programming liver and pancreas progenitors from the endoderm. *Cold Spring Harb. Symp. Quant. Biol.* **73**, 119–126 (2008).
37. Scott, R.A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).
38. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
39. Hindorff, L.A. *et al. A Catalog of Published Genome-Wide Association Studies* http://www.genome.gov/26525384.
40. El-Assaad, W., *et al.* Saturated fatty acids synergize with elevated glucose to cause pancreatic β-cell death. *Endocrinology* **144**, 4154–4163 (2003).
41. Shimabukuro, M., Zhou, Y.T., Levi, M. & Unger, R.H. Fatty acid–induced β cell apoptosis: a link between obesity and diabetes. *Proc. Natl. Acad. Sci. USA* **95**, 2498–2502 (1998).
42. Hu, L. *et al.* Regulation of lipolytic activity by long-chain acyl-coenzyme A in islets and adipocytes. *Am. J. Physiol. Endocrinol. Metab.* **289**, E1085–E1092 (2005).
43. Helgason, A. *et al.* Refining the impact of *TCF7L2* gene variants on type 2 diabetes and adaptive evolution. *Nat. Genet.* **39**, 218–225 (2007).
44. Cho, Y.S. *et al.* Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat. Genet.* **44**, 67–72 (2012).
45. Zaret, K.S. & Grompe, M. Generation and regeneration of cells of the liver and pancreas. *Science* **322**, 1490–1494 (2008).
46. Whyte, W.A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
47. Lovén, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334 (2013).
48. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
49. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
50. 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).
51. Seo, S., Lim, J.W., Yellajoshyula, D., Chang, L.W. & Kroll, K.L. Neurogenin and NeuroD direct transcriptional targets and their regulatory enhancers. *EMBO J.* **26**, 5093–5108 (2007).

© 2014 Nature America, Inc. All rights reserved.

ONLINE METHODS

Human islets. Islets were isolated from donors without a history of glucose intolerance[52], shipped in culture medium and then cultured for 3 d before analysis. β cells were purified by FACS as described[53]. Islet samples were selected from a set of 120 islet samples with ascertainment of minimal exocrine contamination, using dithizone staining and quantitative RT-PCR of lineage-specific markers (**Supplementary Table 7**). High purity was also ascertained by strong H3K4me3 and H3K27ac enrichment at key β cell–specific genes, including *INS*, which shows no enrichment of histone marks in publically available human islet regulatory maps (**Supplementary Fig. 2i**)[11,12,14].

Ethics. Human pancreata were harvested from brain-dead organ donors after obtaining informed consent from family members. Islet isolation centers had permission to use islets for scientific research if they were insufficient for clinical islet transplantation, in accordance with national laws and institutional ethical requirements. Ethical approval for the project was given by the Clinical Research Ethics Committee of Hospital Clinic de Barcelona.

FAIRE and ChIP. Human islets processed for FAIRE and ChIP as described[5,54,55] using 1–1.5 μg of antibodies against H3K4me1 (ab-8895, Abcam), H3K4me3 (Upstate, 05-745), H3K27ac (ab4729, Abcam), CTCF (07-729, Millipore), H2A.Z (ab4174, Abcam; 09-862, Millipore), Pdx1 (AB2027, BCBC), FOXA2 (sc-6554, Santa Cruz Biotechnology), Nkx2.2 (HPA003468, Sigma), MafB (HPA005653, Sigma) and Nkx6.1 (HPA036774, Sigma). Expected cell-specific and subcellular patterns of formalin-fixed epitopes were ascertained by protein blotting on human islet extracts and dual immunofluorescence of human pancreas (**Supplementary Fig. 1b**,**c**).

RNA sequencing. PolyA$^+$ RNA-seq data from three highly purified human islet samples and two FACS-purified human β cells have been described[29]. Non-pancreatic RNA-seq reads were from the Illumina human BodyMap2 project. We processed ~2.5 billion 50-bp paired-end reads from 14 human tissues or cell types (adipose, adrenal gland, breast, colon, heart, kidney, liver, lung, lymphatic node, muscle, ovary, prostate, thyroid gland and white blood cell) identically as for the islet data.

We defined (i) 1,000 islet-specific, non-redundant RefSeq transcripts with the highest median expression ratio in islets versus non-islet samples that were expressed in islets at >10 RPKM, (ii) 1,000 ubiquitously expressed transcripts with median islet versus non-islet expression ratios closest to 1 that had expression in islets of >10 RPKM and (iii)1,000 inactive transcripts with the lowest median RPKM expression in islets.

ChIP-seq analysis. Sequencing was performed on an Illumina HiSeq or Genome Analyzer IIx platform. Quality-filtered 36- to 51-bp single-end reads were aligned to the human genome (NCBI36/hC18) using Bowtie v0.11.3 (ref. 56), allowing only one mismatch and unique mapping. Clonal reads were removed. To calculate genome coverage densities, reads were extended *in silico* to a final length equal to the fragment size estimation based on MACS[57] and were averaged by the number of aligned reads in each ChIP-seq experiment. Sites enriched for transcription factor binding and H2A.Z were detected with MACS[57] using default parameters. Transcription factor peaks were called at $P < 1 \times 10^{-10}$ and H2A.Z peaks were called at $P < 1 \times 10^{-5}$ over a background model derived by sequencing input DNA (**Supplementary Table 8**). We only retained enrichment sites found in replicate experiments. To calculate FDR for replicated peaks, we balanced the number of input and sample reads with MACS and ascertained that >94% of peaks called at these thresholds showed FDR of <0.01. For FAIRE, we tuned the local background noise to $\lambda_{local} = \max \lambda_{5000}, \lambda_{125000}$ and used a cutoff of $P < 1 \times 10^{-3}$, retaining only enrichment sites found in replicate experiments. We considered peaks to be overlapping if they shared a minimum of one base.

Accessible chromatin classes. Accessible chromatin was defined by merging all FAIRE- and H2A.Z-enriched sites that were found in at least two islet samples. These two established measures of accessible chromatin[5,28] proved to be complementary, as FAIRE enrichment was weak in promoters yet readily identified central regions of enhancers and CTCF-bound sites, whereas H2A.Z enrichment was strongest in promoters and also marked enhancers and CTCF-bound sites (**Fig. 2a** and **Supplementary Fig. 3a**). We generated 6-kb windows centered on each accessible chromatin site and computed the read coverage for each feature (H2A.Z, FAIRE, H3K4me3, H3K4me1, H3K27ac and CTCF binding) over 100-bp bins. Cluster3 (ref. 58) was used for *k*-median clustering with HI32, a single human islet sample for which data were available for all marks. For C1, two clusters that showed nearly identical histone modification enrichment patterns were artificially merged into a single class. Enrichment patterns at clusters (or accessible chromatin classes) were concordant in replicate samples (**Supplementary Fig. 3b**).

Genome scan of enhancer clusters. To define enhancer clusters, we first created 1,000 iterations of randomized C3 sites in the mappable genome of individual chromosomes. We then calculated for each chromosome the 25th percentile of the between-site distances of randomized C3 sites (**Supplementary Fig. 6a**). Next, we defined clusters of islet C3 sites as any group of ≥3 C3 sites in which all adjacent C3 sites were separated by less than the above-mentioned 25th-percentile distance for randomized sites in the same chromosome. The distribution of islet C3 clusters differed from those of clusters generated with randomized C3 sites (**Supplementary Fig. 6b**).

We next created a score to measure the average transcription factor occupancy of islet enhancer (C3) clusters. For each cluster, we summed the number of binding sites for each transcription factor at C3 sites and divided this value by the number of C3 sites in the cluster. Clusters with higher average transcription factor occupancy were associated with genes with enriched transcription in islets and β cells (**Supplementary Fig. 7a**). We defined high-occupancy clusters as those showing scores in the top two quartiles (**Fig. 2h** and **Supplementary Fig. 7a**,**b**).

For comparisons of RNA expression in islets and non-islet samples, we identified genes containing a transcription factor–bound enhancer cluster within 25 kb of their annotated TSS (**Fig. 2h** and **Supplementary Fig. 7b**).

For sequence motif analysis, we examined all 19,624 C3 sites that formed clusters, including C3 sites not bound by the transcription factors we had profiled, as we had only profiled a subset of all islet factors. This set of clustered enhancers was also associated with islet-enriched transcription (**Supplementary Fig. 6c**), and only 46% overlapped enhancer sites from nine non-pancreatic cells[34].

We used the Genomic Regions Enrichment of Annotations Tool (GREAT)[59] with default parameters to study enriched functional categories among genes linked to transcription factor–bound enhancer clusters (**Supplementary Fig. 7c**). We also used GREAT to determine which clusters were associated with a list of 65 genes with established importance for islet-specialized function and identity determined on the basis of a systematic literature search (**Supplementary Fig. 7d** and **Supplementary Table 2**).

Transcription factor occupancy, chromatin state and islet-specific transcription. To compare transcription factor occupancy at different chromatin states in islet-specific and ubiquitous versus islet-inactive genes, we identified sites bound by >1 transcription factor in each chromatin class and calculated the density of such sites ±100 kb from the TSSs of the three groups of genes. This level of transcription factor occupancy was chosen to highlight that even high transcription factor occupancy at promoters is not selective for islet-specific gene expression, although conclusions remained unaltered using different numbers of transcription factors.

To directly examine genes bound by islet-specific transcription factors at different chromatin states, we compared quantile-normalized RNA expression in human islets, β cells and 14 non-pancreatic tissues. Data for non-pancreatic tissues were pooled in **Figure 2g**,**h** and are given for individual tissues in **Supplementary Figures 5b**,**d** and **7b**. We compared the mRNA expression of genes bound by transcription factors at C1 sites but lacking transcription factor–bound C3 chromatin within 25 kb of the TSS versus all other genes or genes containing transcription factor–bound C3 clusters within 25 kb of the TSS versus all other genes.

Enhancer function. Human sequences were cloned upstream of the *hsp70* zebrafish core promoter[60] or the *gata2* promoter[61] linked to a Venus reporter and used to inject >200 eggs from wild-type (AB*) and *Tg*(*ins*:*mCherry*)*jh2* zebrafish in three independent experiments. Reporter expression patterns

© 2014 Nature America, Inc. All rights reserved.

were documented using NIS-Elements software (Nikon). Expression was quantified by counting the number of embryos with Venus-positive cells in different tissues (**Supplementary Table 9**). Islet-specific expression patterns from transient transgenic zebrafish were confirmed in stable transgenic lines (**Supplementary Fig. 8**). Transcription factor–bound C3 sites were also cloned in Gateway-adapted pGL4.23 and cotransfected in triplicate wells with pRL in MIN6 and NIH3T3 cells, and luciferase activity was measured after 48 h. Results were expressed as luciferase/*Renilla* ratios with vectors carrying putative enhancers relative to the ratio with empty pGL4.23 vector (**Fig. 3a**). See **Supplementary Table 10** for oligonucleotides.

Transcription factor functional studies. Human EndoC-β H1 β cells[62] were independently transduced with lentiviral vectors expressing two independent shRNAs that target *MAFB* mRNA or four negative-control, nontargeting shRNAs. Each shRNA was transduced in duplicate. See **Supplementary Table 10** for oligonucleotides. *MAFB* shRNAs led to 64% and 55% inhibition of *MAFB* mRNA. We note that *MAFB* was chosen as a transcription factor target on the basis of our ability to design two efficient inhibitory shRNAs.

Mycoplasma-negative HEK293T cells (ATCC CRL-11268) were transfected with pcDNA3-PDX1, pcDNA-NGN3 or pcDNA3.1-MAFA vector or with a control empty vector using Lipofectamine2000 (Invitrogen). After 72 h, RNA was extracted (TRIzol, Invitrogen) and hybridized to GeneChip Human Genome U133 Plus2.0 arrays. To assess the enrichment of predefined gene sets in transcription factor perturbations, robust multi-array average (RMA)-normalized data were analyzed with GSEA[63] using default parameters over 1,000 permutations to calculate FDR. The same analysis was carried out in parallel for an identical number of arrays that compared two non-targeting control and three non-targeting control shRNAs. Gene sets for GSEA were created with genes whose TSSs were closest to C1–C5 sites bound by either MAFB or PDX1. We thus created the following five non–mutually exclusive sets of genes for each transcription factor: MAFB-bound C1 ($n = 363$), C2 ($n = 511$), C3 ($n = 1{,}891$), C4 ($n = 155$) and C5 ($n = 16$); PDX1-bound C1 ($n = 830$), C2 ($n = 878$), C3 ($n = 2{,}874$), C4 ($n = 247$) and C5 ($n = 65$). Similar results were observed using mutually exclusive gene sets (i.e., only sites bound at C3 showed NES > 1 and FDR $q < 0.05$). We also compared the behavior of two mutually exclusive sets of genes, one associated with orphan C3 sites only and another that was only associated with clusters of C3. Gene set sizes were as follows: MAFB-bound orphan C3 ($n = 577$) and clustered C3 ($n = 657$); PDX1-bound orphan C3 ($n = 1{,}233$) and clustered C3 ($n = 1{,}331$).

4C-seq. Chromosome conformation capture assays were performed as described[64] and were adapted for 4C-seq as described[33,65]. Human islets (~5×10^6 cells) were treated with lysis buffer (10 mM Tris-HCl, pH 8, 10 mM NaCl, 0.3% IGEPAL-CA-630 (Sigma-Aldrich) and 1× protease inhibitor (Complete, Roche)). Nuclear DNA was digested with DpnII (New England Biolabs) and ligated with T4 DNA ligase (Promega). Csp6I endonuclease (Fermentas) was used in a second round of digestion, and DNA was religated. Locus-specific primers containing Illumina adaptors (**Supplementary Table 10**) were designed with Primer3 v. 0.4.0 (ref. 66) using the following gene promoters as viewpoints: *ISL1*, *GNAS*, *C2CD4A*, *C2CD4B*, *TM4SF4*, *TM4SF1*, *PDX1*, *MAFB* and *G6PC2*. Eight loci, with the exception of *TM4SF1*, contained enhancer clusters linked to genes with islet-enriched expression. PCR was performed with the Expand Long-Template PCR System (Roche) for each viewpoint, and reactions were pooled for sequencing. 4C-seq data were analyzed by adapting a previously described procedure[33]. Briefly, 4C-seq reads were sorted, aligned and translated to restriction fragments. A moving average of 30 fragments per window was used to smoothen reads. Next, we calculated for each fragment the Poisson probability of it containing a given number of smoothened reads. To this end, all aligned fragments were randomized 1,000 times in a 2-Mb window centered on the viewpoint and smoothened in the same way. We then defined significant interactions in the 4C-seq experiment as those with a Poisson probability of $< 1 \times 10^{-10}$. We tested two C3 promoter interactions by 3C capture, both of which were confirmed. To assess overlaps with chromatin sites, we computed for all nine loci the overlap of different classes of accessible chromatin with 4C-seq interaction sites located in a 2-Mb window centered on the viewpoint, excluding 40 kb on each side of the viewpoint. To contrast expected and observed overlap (**Fig. 3d**,**e**), we randomized different chromatin sites or transcription factor–bound sites 1,000 times in mappable genomic sequences of the same 2-Mb window and calculated the overlap of 4C-seq enrichment sites with these positions.

Electrophoretic mobility shift assays. EMSAs with mouse MIN6 β-cell nuclear extracts were performed as described[67]. We used 1 µl of antibody to NEUROD1 (sc-1084, Santa Cruz Biotchnology) for supershifts. Findings were confirmed with binding experiments carried out on a separate day. See **Supplementary Table 10** for oligonucleotides.

Motif analysis. *De novo* motif discovery was performed with HOMER[50], using a 500-bp window centered on the FAIRE or H2A.Z peaks of clustered C3 sites. We searched for motifs of 6–20 bp in length and retained 46 non-redundant matrices with $P < 1 \times 10^{-60}$. Motifs were annotated using HOMER[50], TOMTOM[68] and manual comparisons (**Supplementary Table 3**). All possible combinations of 3 motifs from the 46 enriched motifs were computed in clustered islet C3 versus analogous genomic sites (H3K4me1 and H3K27ac enriched not H3K4me3 enriched, referred to as strong enhancers by Ernst *et al.*[34]) in 9 non-pancreatic cell lines (GM12878, HepG2, HSMM, HUVEC, ESC, HMEC, K562, NHEK and NHLF). We limited the motif search window to ±250 bp from the center of the genomic site and computed the $\log_2$ value of ratios between motif combination frequencies in islet versus non-islet sites as a metric of islet enrichment. We focused on combinations enriched at χ^2 $P < 1 \times 10^{-3}$. To ensure that results were not affected by differences in the data types used to define enhancers, we also analyzed enhancers from four non-pancreatic cell lines (GM12878, HepG2, HUVEC and K562) for which data on the same chromatin marks were available, including for FAIRE and H2A.Z enrichment[69]. We identified non-islet strong enhancer sites that overlapped FAIRE or H2A.Z sites and calculated the density of the 100 most islet-enriched motif combinations in enhancers from islet and non-islet cells in a 2-kb window centered on merged FAIRE and/or H2A.Z enrichment sites (**Supplementary Fig. 11c**).

To compute the ability of motif combinations to predict islet gene activity in mice, we used HOMER to scan the mouse genome (mm9) for all instances of the ten most islet-enriched motif combinations. We then identified all possible three-motif combinations spanning <500 bp, extended sequences on both sides of motif combinations to create 500-bp segments and filtered for non-redundant combinations. We linked these motif combinations to nearby genes using GREAT with default parameters[59]. Then, RNA-seq RPKM values for mouse islets[29] and nine non-islet tissues (bone marrow, cerebellum, heart, kidney, liver, lung, embryonic fibroblast, embryonic stem cell and spleen; obtained from the Encyclopedia of DNA Elements (ENCODE)/Ludwig Institute for Cancer Research[70]) were aligned and processed as described for human RNA-seq, and expression values were quantile normalized across all tissues[71] (**Supplementary Fig. 11d**).

We also performed *de novo* motif discovery in all accessible chromatin classes (C1–C5) as described above for genome-wide association variant analyses.

Regulome browser. To facilitate the exploitation of integrated data sets, we created a browser (**Supplementary Fig. 15a**,**b**) that enables data downloads and visualization at desired levels of resolution for islet transcription factor binding, chromatin states, motifs and Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC)[37] or DIAGRAM *P* values[38] (see URLs).

Genome-wide association analysis. We identified SNPs with genome-wide association ($P < 5 \times 10^{-8}$) to any trait in European individuals present in the GWAS catalog. Index SNPs were pruned (CEU $r^2 > 0.2$) so that each independent locus was only represented by one index SNP. Each index SNP was then used to identify variants in 1000 Genomes Project pilot 1 data with CEU $r^2 > 0.8$. Thus, an associated locus consisted of an index SNP and the set of 1000 Genomes Project SNPs in high LD with it in the CEU population. We then created a background set containing all qualifying loci binned on the basis of the number of total variants (index SNP + high-LD SNPs) in the locus.

For each C3 site, 200-bp regions directly flanking the left and right ends were obtained. For each trait listed in **Figure 5a**, we calculated the number

© 2014 Nature America, Inc. All rights reserved.

doi:10.1038/ng.2870

of loci in the background set associated with that trait containing a variant overlapping a site for each class (C1–C5) as well as clustered C3, orphan C3 and C3 flanking sites. We then permuted the set of loci by drawing from matching bins in the background set and recalculated the number of loci with a variant overlapping a site for that class. The significance of each overlap was calculated as the number of permuted sets with greater overlap than the observed set divided by the number of permutations. We ran 10,000 permutations except for fasting glycemia, for which we ran 100,000 permutations.

We then identified a comprehensive set of T2D and fasting glycemia index SNPs identified in any population. For T2D, we used all genome-wide significant lead SNPs reported in Morris *et al.*[38] and further identified lead SNPs for loci not genome-wide significant in that study but reported as such by previous studies. For fasting glycemia, we used all genome-wide significant lead SNPs reported in Scott *et al.*[37] for both fasting glycemia levels and fasting glycemia levels adjusted for body mass index loci (the former was retained when different lead SNPs at the same locus were reported).

For all lead SNPs, we identified variants in high LD ($r^2 > 0.8$) in 1000 Genomes Project phase I data for CEU samples (European loci), CHB and JPT samples (Asian loci) or YRI samples (African loci; Yoruba from Ibadan, Nigeria).

HapMap enrichment analysis. We obtained *P* values for all HapMap variants from DIAGRAM[38] and MAGIC[37] studies for T2D and fasting glycemia, respectively. Variants were pruned using LD from HapMap CEU samples as follows: first, all variants with $r^2 > 0.2$ with a lead SNP for a trait in Europeans were removed; second, remaining variants were sorted by *P* value, and variants were included if not having $r^2 > 0.2$ with a more significant variant. Both sets of variants (all HapMap and LD-pruned HapMap variants) were binned on the basis of CEU minor allele frequency and distance to the closest GENCODEv12 TSS. For analyses excluding known associated loci, variants in the 500-kb region surrounding each European lead SNP were removed.

A series of trait association *P*-value thresholds were set, and, at each threshold, the number of variants overlapping an islet chromatin class (C1–C5, clustered C3 and orphan C3) was counted. For each variant, a matching variant was then selected at random from the same bin, and the number of matched variants with a *P* value below the same threshold was counted. Fold enrichment was calculated for each threshold by dividing the observed counts by the matched counts averaged over 1,000 permutations.

To evaluate the significance of these enrichments, we focused on variants overlapping each chromatin class that attained $P < 0.001$ in T2D or fasting glycemia meta-analyses. We obtained counts of matched variants significant at the same trait *P*-value threshold across all permutations. As the matched variant counts for these replicates were normally distributed, we calculated a *z* score for the observed islet class counts using the mean and s.d. from the distribution of permuted counts. Reported *P* values were then obtained from *z* scores using a one-sided test.

52. Bucher, P. *et al.* Assessment of a novel two-component enzyme preparation for human islet isolation and transplantation. *Transplantation* **79**, 91–97 (2005).
53. McCulloch, L.J. *et al. GLUT2* (*SLC2A2*) is not the principal glucose transporter in human pancreatic β cells: implications for understanding genetic association signals at this locus. *Mol. Genet. Metab.* **104**, 648–653 (2011).
54. Servitja, J.M. *et al.* Hnf1α (MODY3) controls tissue-specific transcriptional programs and exerts opposed effects on cell growth in pancreatic islets and liver. *Mol. Cell. Biol.* **29**, 2945–2959 (2009).
55. van Arensbergen, J. *et al.* Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing β-cells to adopt a neural gene activity program. *Genome Res.* **20**, 722–732 (2010).
56. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
57. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
58. de Hoon, M.J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453–1454 (2004).
59. McLean, C.Y. *et al.* GREAT improves functional interpretation of *cis*-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
60. Blechinger, S.R. *et al.* The heat-inducible zebrafish *hsp70* gene is expressed during normal lens development under non-stress conditions. *Mech. Dev.* **112**, 213–215 (2002).
61. Meng, A., Tang, H., Ong, B.A., Farrell, M.J. & Lin, S. Promoter analysis in living zebrafish embryos identifies a *cis*-acting motif required for neuronal expression of GATA-2. *Proc. Natl. Acad. Sci. USA* **94**, 6267–6272 (1997).
62. Ravassard, P. *et al.* A genetically engineered human pancreatic β cell line exhibiting glucose-inducible insulin secretion. *J. Clin. Invest.* **121**, 3589–3597 (2011).
63. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
64. Tena, J.J. *et al.* An evolutionarily conserved three-dimensional structure in the vertebrate Irx clusters facilitates enhancer sharing and coregulation. *Nat. Commun.* **2**, 310 (2011).
65. Splinter, E., de Wit, E., van de Werken, H.J., Klous, P. & de Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods* **58**, 221–230 (2012).
66. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–386 (2000).
67. Boj, S.F., Parrizas, M., Maestro, M.A. & Ferrer, J. A transcription factor regulatory circuit in differentiated pancreatic cells. *Proc. Natl. Acad. Sci. USA* **98**, 14481–14486 (2001).
68. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. & Noble, W.S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).
69. Rosenbloom, K.R. *et al.* ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.* **41**, D56–D63 (2013).
70. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
71. Hansen, K.D., Irizarry, R.A. & Wu, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204–216 (2012).

© 2014 Nature America, Inc. All rights reserved.