CRISPR Screening Uncovers a Long-Range Enhancer for *ONECUT1* in Pancreatic Differentiation and Links a Diabetes Risk Variant

Samuel Joseph Kaplan[1,2], Wilfred Wong[1,3], Jielin Yan[2,4], Julian Pulecio[2], Hyein S. Cho[2], Qianzi Li[1,3], Jiahui Zhao[1], Jayanti Leslie-Iyer[5], Jonathan Kazakov[5], Dylan Murphy[1], Renhe Luo[2,4], Kushal K. Dey[3], Effie Apostolou[6], Christina S. Leslie[3], Danwei Huangfu[2,*]

1. Weill Cornell Graduate School of Medical Sciences, Weill Cornell Medical College, New York, NY, 10065, USA.

2. Developmental Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA.

3. Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA.

4. Louis V. Gerstner Jr. Graduate School of Biomedical Sciences, Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA.

5. Intern, Developmental Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA.

6. Meyer Cancer Center, Division of Neuro-Oncology, Department of Neurology, Sandra and Edward Meyer Cancer Center, New York-Presbyterian Hospital/Weill Cornell Medicine, New York, NY, 10065, USA.

* Correspondence: huangfud@mskcc.org

Abstract

Functional enhancer annotation is a valuable first step for understanding tissue-specific transcriptional regulation and prioritizing disease-associated non-coding variants for investigation. However, unbiased enhancer discovery in physiologically relevant contexts remains a major challenge. To discover regulatory elements pertinent to diabetes, we conducted a CRISPR interference screen in the human pluripotent stem cell (hPSC) pancreatic differentiation system. Among the enhancers uncovered, we focused on a long-range enhancer ~664 kb from the *ONECUT1* promoter, since coding mutations in *ONECUT1* cause pancreatic hypoplasia and neonatal diabetes. Homozygous enhancer deletion in hPSCs was associated with a near-complete loss of *ONECUT1* gene expression and compromised pancreatic differentiation. This enhancer contains a confidently fine-mapped type 2 diabetes associated variant (rs528350911) which disrupts a GATA motif. Introduction of the risk variant into hPSCs revealed substantially reduced binding of key pancreatic transcription factors (GATA4, GATA6 and FOXA2) on the edited allele, accompanied by a slight reduction of *ONECUT1* transcription, supporting a causal role for this risk variant in metabolic disease. This work expands our knowledge about transcriptional regulation in pancreatic development through the characterization of a long-range enhancer and highlights the utility of enhancer discovery in disease-relevant settings for understanding monogenic and complex disease.

38 **Main**

39 A spectrum of genetic causality often influences the onset and varying severity of symptoms for
40 a disease[1]. This concept is exemplified in metabolic conditions like diabetes, where a single
41 nucleotide alteration in a protein coding region can cause severe neonatal diabetes mellitus
42 (NDM) or maturity-onset diabetes of the young (MODY), while the aggregation of polygenic
43 effects is thought to confer susceptibility to type 2 diabetes (T2D)[2,3]. However, most genetic
44 variants have moderate effects and are in unannotated non-coding genomic regions, with the
45 affected genes typically unknown[3,4]. Therefore, there is a pressing need to understand the
46 functional impact of disease-associated variants uncovered by genome-wide association studies
47 (GWAS) and manage the burgeoning set of variants of unknown significance identified through
48 the widespread adoption of high-throughput DNA sequencing.

49

50 It is hypothesized that many genetic variants affect enhancers[5], and alterations in their sequences
51 can affect gene expression, contributing to both Mendelian and complex disease traits. A clear
52 illustration of this impact is shown by the limb malformation caused by a mutation in a *SHH*
53 enhancer ~1 Mb from the gene promoter[6]. This represents one of a handful of clinical genetics
54 examples where enhancer mutations have a pronounced pathological impact[7]. On the opposite
55 end of the disease severity spectrum, functional consequences of individual sequence alterations
56 in common complex diseases are generally unknown, but a significant proportion of disease-
57 associated genomic variations identified through GWAS are located within putative enhancers[5].
58 Together, these findings have contributed to the inception of the enhanceropathy concept[7].
59 However, diagnosis of suspected enhanceropathies remains difficult due to the lack of functional
60 annotation of most enhancers. While putative enhancers can be identified based on genomic

61 features like chromatin accessibility, complementing descriptive epigenetics through scalable
62 functional characterization in tissue-relevant settings poses a considerable challenge. One
63 obstacle is the variable distance enhancers can have from their target genes. In our recent
64 interrogation of developmental enhancers, we found that 41% of the enhancer identified were
65 located >100 kb away from the target gene promoter[8], yet practical considerations often lead
66 many functional enhancer screens to confine their search to regions within 100 kb of the target
67 gene promoter[9]. Consequently, the overall prevalence of long-distance enhancer-promoter
68 regulations remains unclear despite a growing set of long-range enhancers characterized through
69 genetic studies[10].
70
71 Human pluripotent stem cell (hPSC) differentiation represents a powerful system for disease
72 modeling. The integration of genetic perturbation and genomic approaches has proven
73 instrumental for investigating the cascade of gene and enhancer activation during development
74 and holds promise for understanding long-term health implications. We and others have shown
75 NDM gene knockout in hPSCs recapitulates key aspects of human genetic conditions[11]. In
76 addition, the hPSC platform has been used to study the impact of clinically observed recessive
77 mutations in a *PTF1A* enhancer, shedding light on its role in pancreatic agenesis and NDM[12,13].
78 However, many cases of NDM lack a known genetic cause even with whole exome sequencing,
79 underscoring the importance of identifying developmental enhancers[14]. Significant progress in
80 biochemical characterization has narrowed the search space for enhancers, but many putative
81 enhancers have not been functionally interrogated. Furthermore, the extensive epigenetic
82 rewiring during development further swells the number of potential enhancers, adding
83 complexity to the characterization task[15,16]. This challenge is becoming more tractable with the

84 use of CRISPR-Cas9 and dCas9-based CRISPR interference (CRISPRi) tools, enabling the
85 simultaneous perturbation of genes or enhancers en masse[9]. We and others have applied
86 CRISPR-Cas9 in hPSCs to uncover genes important for gastrulation and pancreatic
87 development[17-22]. Recent CRISPRi screens have further identified enhancers involved in hPSC
88 endoderm and cardiac differentiation[8,23,24]. These advances motivated us to leverage our
89 epigenetic characterization of the hPSC pancreatic differentiation system[25] to uncover enhancers
90 with potential roles in diabetes.
91
92 The hPSC pancreatic differentiation system directs cells progressively through the definitive
93 endoderm (DE), gut tube (GT), and posterior foregut (PFG, also known as PP1) stages that
94 model embryonic development in humans[26]. Building on our recent success in using a cell
95 identity gene as a readout for enhancer discovery in cell state transitions[8], we utilized the
96 developmental expression initiation of diabetes-associated gene *PDX1* at the PFG stage as a
97 readout for the onset of pancreatic differentiation[11,17]. We then selected candidate transcription
98 factors (TFs) likely to influence PDX1+ identity commitment and hypothesized that repressing
99 enhancers of these TFs would affect PDX1 expression. Our subsequent screen indeed uncovered
100 enhancers of multiple TFs that affected PDX1 activation during pancreatic specification. We
101 focused on a distal enhancer of *ONECUT1*, located ~664 kb from the gene promoter. This choice
102 was motivated by the presence of a high-confidence fine-mapped T2D-associated single
103 nucleotide polymorphism (SNP) within this enhancer[27] and clinical genetics findings indicating
104 that coding mutations in *ONECUT1* cause NDM[28,29]. We show that *ONECUT1* enhancer deletion
105 in hPSCs resulted in an almost complete absence of *ONECUT1* transcript and protein, along with
106 a reduced proportion of cells expressing PDX1 during pancreatic differentiation. Enhancer

107 deletion had locus-wide effects within the topologically associating domain (TAD), with reduced
108 antisense transcription and decreased H3K27ac levels. Finally, we generated hPSC lines
109 harboring the T2D variant (rs528350911) and found the risk variant significantly impaired the
110 binding of key pancreatic transcription factors GATA4, GATA6, and FOXA2, but had a smaller
111 effect on *ONECUT1* transcription compared to enhancer knockout. Our study not only identified
112 an essential enhancer of *ONECUT1* in pancreatic development but also developed and reduced to
113 practice a framework for prioritizing disease-associated variants for investigation.
114

115 ## Results

116 ### CRISPRi enhancer screen design

117 To discover enhancers involved in pancreatic development and diabetes, we conducted a
118 CRISPRi screen in a hPSC pancreatic differentiation system, using PDX1 expression as a
119 readout. PDX1 is a key marker for the initiation of pancreatic differentiation and is indispensable
120 for mammalian pancreatic development[30-33]. Homozygous and heterozygous mutations in *PDX1*
121 cause NDM and maturity onset diabetes of the young (MODY), respectively[34,35]. Additionally,
122 genomic variants in the *PDX1* locus are associated with elevated T2D risk[36]. To identify putative
123 enhancer regions for study, we first selected 23 candidate TFs (including PDX1) based on their
124 likely impact on PDX1+ induction and diabetes by integrating findings from murine
125 experiments[30,31,37-55], clinical genetics, and our previous CRISPR-Cas9 screens in hPSC
126 pancreatic differentiation (Fig. 1B). 11 of these TFs have known roles in human metabolic
127 disorders or NDM[29,34,56-61]. 15 TFs, including GATA6, RFX6, and ONECUT1, have also been
128 shown to affect pancreatic differentiation and PDX1 expression in hPSC disease models or
129 CRISPR-Cas9 coding screens for regulators of DE and PFG specification[17,18,28,62,63].

130

131 TADs have been proposed to play a key role in enhancer function[64], so we conducted Hi-C
132 assays on PFG stage cells and used TAD information to delineate a ~2-6 Mb linear window
133 around each TF for interrogation (Sup. Table 1). Within each window, we collated all accessible
134 regions identified based on ATAC-seq peaks across the DE, GT, and PFG stages for
135 interrogation (Fig. 1C, Sup. Table 1)[25]. These accessible regions also included 758 annotated
136 promoters, facilitating benchmarking of the CRISPRi screen results against our previous
137 genome-scale CRISPR-Cas9 coding screen conducted in the same differentiation context[17].
138 Thus, we designed a library with 37,184 gRNAs targeting 6,443 regions (avg. length of 300 bp),
139 1100 safe-targeting[65], and 463 non-targeting gRNAs[66], and cloned the sequences into a lentiviral
140 backbone (Sup. Table 1).

141

142 **CRISPRi screen identifies enhancers and genes necessary for human pancreatic**
143 **differentiation**

144 We conducted a pooled CRISPRi screen in our hPSC pancreatic differentiation system[17], with
145 two biological replicates (independent infections and differentiations) referred to as Screen 1 and
146 2 (Fig. 1A). For each screen, we infected inducible dCas9-KRAB H1 hPSCs[67] with the gRNA
147 library and induced dCas9-KRAB expression. Following differentiation, we enriched for PDX1+
148 and PDX1- cells at the PFG stage through fluorescence-activated cell sorting (FACS) (Fig. 1D,
149 Sup. Fig. 1A). Subsequently, next-generation sequencing determined gRNA abundance within
150 the sorted populations, and MAGeCK analysis[68] led us to prioritize 36 top hits that caused a
151 PDX1 decrease (Fig. 1E, Sup. Fig 1B, Sup. Table 1). Both promoter (19) and non-promoter (17)
152 hits have greater sequence conservation compared to all regions investigated, as shown by the

153 average conservation scores calculated from PhyloP100[69] (Sup. Fig. 1C). Furthermore, we
154 examined all interrogated promoter regions and found the gene hits to be largely concordant with
155 our previous Cas9 coding screen conducted in a similar pancreatic differentiation context[17] (Fig.
156 1F). These findings support the utility of CRISPRi repression for the discovery of both coding
157 and non-coding regulators of hPSC differentiation.
158
159 We next focused on non-promoter hits and examined H3K27ac, a histone mark associated with
160 active enhancers[70]. Non-promoter hit regions did not have H3K27ac peaks in undifferentiated
161 hPSCs, but gained H3K27ac at various stages during pancreatic differentiation (Fig. 1G,H),
162 indicating stage-specific enhancer activity. We further reasoned that each enhancer hit is likely to
163 correspond to a promoter hit for the regulated gene. Indeed, almost all discovered enhancers
164 were linked to a single promoter hit within the linear neighborhood, resulting in a total of 15
165 enhancer-gene pairs (Fig. 1I, Sup. Fig. 1D). This approach demonstrates the specificity of the
166 enhancer-gene regulatory mechanism. In addition, we observed that the enhancer-promoter
167 distances ranged from 2 to 664 kb, with three enhancers found beyond 100 kb from the
168 transcription start site (TSS) (Fig. 1J). Thus, unbiased screening in hPSC pancreatic
169 differentiation uncovered many previously undiscovered enhancers and expanded our
170 understanding of gene regulation during human pancreas development.
171
172 ***ONECUT1e-664kb* deletion causes allele-specific loss of *ONECUT1* expression**
173 The CRISPRi screen identified the *ONECUT1* promoter as well as a site ~664 kb from the
174 *ONECUT1* TSS, denoted as *ONECUT1e-664kb* henceforth. We observed that at the PFG stage,
175 the *ONECUT1* promoter showed significantly increased Hi-C contact with a region

176 encompassing *ONECUT1e-664kb* compared to the DE stage (Fig. 2A, 2B). This observation
177 suggests a physical enhancer-promoter interaction and substantial enhancer rewiring during
178 differentiation. We chose *ONECUT1e-664kb* for further investigation because *ONECUT1*
179 mutations cause pancreatic hypoplasia in humans[28,29] and *ONECUT1* is necessary for efficient
180 hPSC differentiation to pancreatic progenitors[28]. Consistent with the clinical findings, *ONECUT1*
181 null mice had a reduced number of endodermal cells expressing PDX1 at pancreatic budding[55]
182 accompanied by compromised pancreatic organogenesis[71-73]. To validate the regulatory impact
183 of *ONECUT1e-664kb* on *ONECUT1*, we used CRISPR-Cas9 deletion as an orthogonal method
184 to complement our dCas9-KRAB based repression approach. Using paired gRNAs flanking a
185 ~750 bp region, we generated two heterozygous and one homozygous clonal enhancer deletion
186 lines (Fig. 2C). *ONECUT1* transcription is activated during the GT to PFG transition (Sup. Fig.
187 2A), so we differentiated the homozygous and heterozygous enhancer knockout (eKO) lines to
188 the PFG stage and assessed PDX1 and ONECUT1 protein levels via flow cytometry (Fig. 2D).
189 Homozygous eKO led to a near-complete depletion of ONECUT1+ cells compared to unedited
190 cells (Fig. 2E). In addition, the percentage of cells activating PDX1 expression was significantly
191 lower in homozygous eKO cells, showing that the loss of ONECUT1 impaired pancreatic
192 differentiation (Fig. 2F). These findings were further confirmed through immunofluorescence
193 staining for PDX1 and ONECUT1 (Sup. Fig. 2B). To assess the impact of enhancer knockout at
194 a later pancreatic differentiation stage, we differentiated eKO cells to pancreatic progenitors
195 (PP2) (Sup. Fig. 2C). Homozygous enhancer knockout cells remained ONECUT1 negative at the
196 PP2 stage (Sup. Fig. 2F,G), and had a significant decrease in the percentage of PDX1+NKX6.1+
197 cells (Sup. Fig. 2D, E). This recapitulates the previously observed phenotype of *ONECUT1* gene
198 knockout in hPSC differentiation[28]. To explore the mechanisms by which *ONECUT1* may be

199 affecting *PDX1* expression, we examined the *PDX1* enhancers uncovered in our screen for
200 ONECUT1 binding[17]. While all *PDX1* enhancers gained ATAC-seq and H3K27ac ChIP-seq
201 signals during differentiation, we observed ONECUT1 ChIP-seq signal primarily at *PDX1e-2kb*,
202 suggesting ONECUT1 may activate *PDX1* through this enhancer (Sup. Fig. 2I). Additionally,
203 this enhancer region has binding of multiple other transcription factors such as HHEX, FOXA2,
204 GATA4, and GATA6. This is a potential mechanism by which the cells are able to overcome
205 loss of ONECUT1 and activate PDX1 at the PP2 stage.
206
207 While heterozygous lines did not exhibit a significantly lower percentage of PDX1+ or
208 ONECUT1+ cells, there was a significant decrease in the mean fluorescence intensity signal of
209 ONECUT1 among the PDX1+ cells in heterozygous cells compared to wild-type (WT) cells at
210 both the PFG and PP2 stages (Fig. 2G,H , Sup. Fig. 2H). This suggested that the heterozygous
211 cells had lower *ONECUT1* expression compared to WT cells. To directly assess the *cis* effect of
212 heterozygous eKO on *ONECUT1* transcription, we pursued a more quantitative approach and
213 designed an allele-specific droplet digital PCR (ddPCR) assay for the *ONECUT1* transcript.
214 Given the substantial linear genomic distance between *ONECUT1e-664kb* and the *ONECUT1*
215 gene, we first called and phased all SNPs in the *ONECUT1* locus after conducting targeted long-
216 read Nanopore sequencing with adaptive sampling (Sup. Fig. 2J). This approach allowed us to
217 identify a SNP in the 3'UTR of *ONECUT1* and design an allele-specific ddPCR assay (Fig. 2I,
218 K). We also identified a genotyping SNP close to *ONECUT1e-664kb* for phasing the deleted
219 enhancer in the heterozygous lines via PCR and found both heterozygous lines had enhancer
220 deletions on the same allele (designated as "Allele 2") (Fig. 2J). ddPCR assays performed on
221 RNA extracted from cells differentiated to the PFG stage showed that homozygous eKO lines

222 lost >95% of transcription from both alleles (Fig. 2L). Heterozygous lines exhibited a specific
223 loss of transcription from Allele 2, resulting in a significantly skewed ratio of transcripts from
224 Allele 2 over Allele 1. This skewed ratio was not observed in control ddPCR assays conducted
225 on genomic DNA (gDNA) (Sup. Fig. 2K). These findings demonstrate that the transcriptional
226 impact of *ONECUT1e-664kb* deletion on *ONECUT1* expression is due to direct *in cis* regulation,
227 rather than indirect effect through the downregulation of the ONECUT1 TF or other
228 mechanisms. Notably, one of the heterozygous eKO lines (Het. 1) had the enhancer sequence
229 inverted on Allele 1, and *ONECUT1* transcription from Allele 1 was not significantly affected by
230 this inversion. Together, these results establish *ONECUT1e-664kb* as a critical *cis*-regulatory
231 element for *ONECUT1* activation in pancreatic development.
232

233 ***ONECUT1* enhancer deletion causes transcriptional and epigenomic changes in the locus**

234 The large impact of *ONECUT1e-664kb* deletion on *ONECUT1* transcription led us to investigate
235 its effects on transcription in the broader genomic locus since some enhancers may serve as
236 "hubs" that simultaneously regulate the expression of multiple genes in a locus[74]. We performed
237 TruSeq Stranded Total (Ribodepletion) RNA-seq on WT, heterozygous, and homozygous eKO
238 cells differentiated to the GT and PFG stages. Consistent with our ddPCR results, homozygous
239 eKO cells had an approximately 30-fold decrease in *ONECUT1* reads at the PFG stage compared
240 to reads in WT cells. While we did not detect any significant transcription changes of genes in
241 the adjacent TADs, we did observe a statistically significant decrease in *WDR72* expression in
242 homozygous *ONECUT1e-664kb* cells (Fig. 3A). Notably, *WDR72* is the only other protein-
243 coding gene in the same TAD as *ONECUT1*. However, ddPCR assays did not show a
244 statistically significant difference in *WDR72* transcripts, despite a noticeable trend towards a

245 decrease (Sup. Fig. 3A), which may be attributed to the low expression of *WDR72* in the PFG
246 stage. In summary, we conclude *ONECUT1e-664kb* primarily regulates *ONECUT1* transcription
247 during pancreatic differentiation.
248
249 We further investigated the impact of *ONECUT1e-664kb* deletion on *ONECUT1* promoter
250 antisense (PAS) RNA transcription. Transcription-associated PAS RNAs have been extensively
251 documented[75,76], and antisense RNAs can be quantified in strand-specific RNA-seq datasets[77].
252 Indeed, we observed an adjacent non-coding region antisense to *ONECUT1* that had lower
253 antisense transcription in homozygous eKO cells compared to WT cells (Fig. 3B). Quantifying
254 antisense transcripts in this region across three differentiation replicates showed that during the
255 GT-PFG transition, *ONECUT1* transcriptional activation in WT cells was associated with a
256 significant increase in *ONECUT1* antisense transcripts. These transcripts were significantly
257 reduced in both heterozygous and homozygous eKO lines at the PFG stage (Fig. 3C). In contrast,
258 sense transcripts across the same non-coding region showed no change during the GT-PFG
259 transition and were unperturbed upon enhancer deletion (Sup. Fig. 3B). Thus, enhancer deletion
260 has transcriptional effects beyond the decrease of the *ONECUT1* transcript, illustrating an
261 additional transcriptional phenotype of enhancers loss.
262
263 The transcriptional changes we observed upon enhancer deletion motivated us to investigate the
264 effect of enhancer deletion on H3K27ac: an epigenetic mark associated with active enhancers
265 and transcription. Additionally, a previous study documented that *PTF1A* enhancer deletion
266 caused H3K27ac loss at the *PTF1A* promoter as well as adjacent regions in the genomic locus[12].
267 However, given that *PTF1A* encodes a TF, it is challenging to distinguish *trans* and *cis* effects of

268 enhancer deletion. We reasoned that analyzing allele-specific H3K27ac ChIP-seq reads could aid
269 in examining *cis*-consequences of enhancer deletion, with a *cis*-effect manifesting as an
270 imbalance in allele-specific reads within heterozygous eKO cells. We performed H3K27ac ChIP-
271 seq on PFG cells differentiated from WT, heterozygous, and homozygous eKO lines. First, we
272 compared WT and homozygous eKO cells. The decrease in H3K27ac in the *ONECUT1* TAD
273 was the most pronounced of all regions on chromosome 15 (Sup. Fig. 3C), strongly suggesting
274 *cis* effects. Among the 16 regions with the most significant H3K27ac reduction (adjusted p-value
275 < 1E-15) (Fig. 3D), 10 contained heterozygous SNPs. Since these SNPs were phased with long-
276 read sequencing (Sup. Fig. 2J), we could use the SNPs to quantify changes occurring *in cis* with
277 enhancer deletion by calculating allele-specific H3K27ac ChIP-seq read ratios in both WT and
278 heterozygous eKO cells (Sup. Fig. 3D). As controls, we used regions with heterozygous SNPs
279 that did not meet the significance threshold of the differential analysis. For all top differential
280 peaks, H3K27ac was decreased *in cis* to the allele with enhancer deletion in heterozygous eKO
281 cells, with some positions losing nearly all H3K27ac (Fig. 3E), whereas control regions showed
282 a near 1:1 allele read ratio average in both WT and heterozygous eKO cells (Fig. 3F). This
283 demonstrates *ONECUT1e-664kb* affects H3K27ac at multiple genomic positions within the
284 locus, supporting a model of direct crosstalk among the putative regulatory elements. Thus, the
285 reduction of H3K27ac signal and PAS transcripts *in cis* with enhancer deletion illustrates the
286 impact of enhancer dysregulation on genome function beyond gene transcription.
287

288 **T2D-associated SNP decreases pancreatic TF recruitment to *ONECUT1e-664kb***

289 To explore the significance of *ONECUT1e-664kb* in a clinical genetics context, we examined
290 SNP metabolic disease-association statistics in the *ONECUT1* locus using data from a recent

291 T2D GWAS meta-analysis[27] and other studies[78,79]. We observed one distinct T2D disease
292 association signal around the *ONECUT1* gene with a largely unresolved 99% credible set
293 spanning ~100kb genomic region (hg38, chr15:52777944-52873484). We identified a second
294 independent T2D disease-causal signal ~664 kb away from *ONECUT1* and close to *WDR72*, that
295 is driven by a single variant, rs528350911 (G/C, reverse strand, minor allele frequency =
296 0.0068)[27] (Fig 4A). This variant is directly within *ONECUT1e-664kb* and is highly confidently
297 fine-mapped with the posterior probability of association (PPA) close to 1 (0.994) after linkage
298 disequilibrium (LD) correction that is part of SuSiE fine-mapping[27,80]. To predict if rs528350911
299 would affect enhancer activation in pancreatic development, we trained a ChromBPNet[81,82]
300 model on PFG-stage ATAC-seq data[25]. Then, we computed per nucleotide DeepLift[83] attribution
301 scores across the region for both the reference and alternate sequences of rs528350911. The
302 decreased attribution for an entire GATA TF motif suggests that the G in the GATA motif is
303 crucial for the model's genomic accessibility prediction, and the alternate sequence associated
304 with T2D risk leads to an attribution loss at the site (Fig. 4B).
305
306 GATA factors such as GATA4 and GATA6 are known for their important role in pancreatic
307 differentiation[62] and are linked to neonatal diabetes[84,85]. At the PFG stage, the key pancreatic TFs
308 GATA4, GATA6, and FOXA2, bind to the *ONECUT1e-664kb* enhancer region (Sup. Fig.
309 4A)[25,62]. This motivated us to investigate the molecular impact of rs528350911. We engineered
310 three heterozygous hPSC lines with the risk variant: two carried the variant on Allele 2, and one
311 on Allele 1 (Fig. 4C). While all lines differentiated to the PFG stage had a similar percentage of
312 ONECUT1 and PDX1 expressing cells (Fig. 4D, Sup. Fig. 4B) and did not affect transcription of
313 WDR72 (Sup. Fig. 4C), we observed a significant reduction in ONECUT1 MFI for two of the

314 three lines (Fig. 4E). Allele-specific ddPCR quantification of *ONECUT1* transcription (Fig. 2K)
315 confirmed that the same two lines had a small but significant allele-specific reduction in
316 *ONECUT1* transcription (Fig. 4F).
317
318 We further investigated the impact of rs528350911 on TF binding at *ONECUT1-e664kb*. To
319 accomplish this, we developed an allele-specific ddPCR assay targeting the reference and
320 alternate (T2D associated) nucleotides (Fig. 4G) to quantify TF binding following chromatin
321 immuno-precipitation (ChIP) in an allele-specific manner. By utilizing the heterozygous lines,
322 this allele-specific assay supports an internally controlled experimental system that mitigates
323 technical variabilities arising from *in vitro* differentiation, ChIP, and PCR. We selected two
324 hPSC lines that contained the risk variant on different alleles; one line had a significant impact
325 on *ONECUT1* expression, while the other did not show a significant effect as mentioned above.
326 After differentiating the lines to the PFG stage, we performed ChIP for GATA4, GATA6, and
327 FOXA2. The precipitated DNA was then used as the template in the ddPCR assay, revealing a
328 ~4-fold reduction in GATA4 binding on the allele containing the disease-associated SNP in both
329 lines (Fig. 4H). We also observed significantly decreased GATA6 and FOXA2 binding. Thus,
330 despite a marginal effect on *ONECUT1* transcription, the disease-associated SNP substantially
331 reduced TF binding to *ONECUT1e-664kb*.
332

333 **Discussion**

334 Pancreatic developmental enhancers have been found in mice by integrating pre-existing
335 knowledge of important pancreatic genes with annotations such as evolutionary conservation and
336 DNA accessibility. Using this approach, several murine *PDX1* enhancers were identified, and

337 knockout mice had a spectrum of pancreatic phenotypes depending on the specific region that
338 was removed[86,87]. A similar approach also identified a putative *GATA4* enhancer, which was
339 subsequently verified by a transgenic reporter assay in mouse endoderm and endoderm-derived
340 tissues including the pancreas and duodenum[88]. These studies set the stage for integration of
341 biochemical epigenetic marks into the enhancer discovery framework, which led to the discovery
342 of a *PTF1a* enhanceropathy responsible for pancreatic agenesis and neonatal diabetes in
343 humans[12,13]. The studies also highlight the value of developmental enhancer discovery for
344 elucidating mechanisms of organogenesis and disease diagnosis. Building upon these findings,
345 our study expanded the scope of enhancer discovery by targeting developmentally accessible
346 chromatin regions for large-scale functional interrogation using dCas9-KRAB. Our study not
347 only identified the human *PDX1* and *GATA4* enhancers corresponding to the aforementioned
348 mouse enhancers (Sup. Table 1), but also uncovered additional enhancers that had not been
349 previously described. These newly discovered enhancers can serve as a diagnostic resource not
350 only for the large number of monogenic diabetes patients with no known genetic cause[14], but
351 also for comprehending variants identified through population-level association studies.
352
353 From all the discovered enhancers, this work contributes *ONECUT1e-664kb* to a select group of
354 well-documented long-range regulatory sequences that activate transcription over genomic
355 intervals exceeding 500 kb[89-97]. These enhancers are valuable models for further investigation
356 into long-range transcriptional regulation and its potential implications in human disease. We
357 speculate that gene regulation by long-range enhancers is more common than appreciated, and
358 our findings underscore the value of unbiased discovery. From a technical standpoint, the
359 substantial genomic distance between *ONECUT1e-664kb* and its target gene presented a

360 challenge for definitively demonstrating the *cis* impact of enhancer deletion. In addition to
361 conducting Hi-C to support enhancer-promoter interaction, we phased the variants in the
362 *ONECUT1* locus and employed high-sensitivity variant-specific ddPCR assays in heterozygous
363 lines. As a result, we uncovered two mechanisms beyond transcriptional control by which
364 regulatory regions could influence disease traits: loss of promoter antisense transcription and
365 H3K27ac in a genomic locus. We also demonstrated a near-complete, allele-specific loss of
366 *ONECUT1* transcription resulting from the deletion of *ONECUT1e-664kb*. Considering results
367 from *ONECUT1* knockout studies in mice[55] and hPSC pancreatic differentiation[28], we infer
368 ONECUT1 protein loss to be the driver of the pancreatic differentiation phenotype in our study.
369 Our findings are also in concordance with the reduced number of PDX1 expressing cells during
370 pancreatic budding in *ONECUT1* null mice that contributes to pancreatic hypoplasia[55]. We
371 predict that loss of *ONECUT1e-664kb* in an individual would phenocopy the severe pancreatic
372 hypoplasia observed in cases of *ONECUT1* protein-coding mutations in clinical settings[28,29].
373
374 The discovery of *ONECUT1e-664kb* also helped us prioritize a disease risk variant for functional
375 interrogation. In previous GWAS, the T2D-associated variant rs528350911 is assigned to the
376 closest gene – *WDR72* [27]. Considering the strong effect of *ONECUT1e-664kb* deletion on
377 *ONECUT1* expression and the known association of *ONECUT1* with diabetes, it is likely this
378 variant confers disease susceptibility through its effect on ONECUT1 expression. Our study
379 demonstrates how the integration of functional screening, precise genome editing, deep learning
380 models, and genomic conformation information can improve the functional pairing of variants
381 with genes. The introduction of the disease variant into hPSCs did not have a strong effect on
382 *ONECUT1* transcription, but its robust effect on TF binding to *ONECUT1e-664kb* suggests that

383 assays other than gene expression could be valuable for characterizing the molecular
384 consequences of variants in endogenous loci. We speculate rs528350911 may confer a latent
385 disease vulnerability, which could be exposed by factors not fully modeled in our current
386 differentiation system such as age and environment[98,99,]. Given the considerable distance
387 between *ONECUT1e-664kb* and the *ONECUT1* gene, as well as the limited impact of the
388 rs528350911 risk variant on *ONECUT1* transcription, it is unlikely this variant would have been
389 prioritized for investigation in a high-throughput variant modeling screen that relies on prime or
390 base editing technologies. Assays with higher sensitivity are likely necessary to detect smaller
391 magnitude effects of point variants like rs528350911 compared to the more pronounced effects
392 of enhancer repression with dCas9-KRAB[100]. Our work demonstrates that integrating *in vitro*
393 functional enhancer discovery with population-level genomics provides a rational approach to
394 efficiently prioritize variants for deeper investigation. This framework holds promise for
395 understanding the exponentially growing set of disease-associated variants discovered through
396 population-wide GWAS and whole genome sequencing efforts.

397 **Limitations of the Study**

398 1. We were not able to identify the promoter hits for two of the enhancers, presumably due to
399 variations in gRNA efficiency and screen cutoff stringency. The investigation was also limited to
400 the embryonic pancreatic differentiation system. This system may not be optimal for
401 investigating disease-associated variants for complex diseases that develop over a person's adult
402 lifetime such as Type II Diabetes. Additionally, genes can have different functions in different
403 tissues, and it is possible there are additional non-developmental[72,73] or non-pancreatic[101]
404 mechanisms through which ONECUT1 variants could contribute to metabolic disorders.

405 2. There are long non-coding RNAs (lncRNAs) in the intergenic region between *ONECUT1* and
406 *WDR72* with the annotations differing substantially between transcriptome assemblies. In some
407 databases, rs528350911 has been linked to one of these lncRNAs, *LINC02490*. We hypothesize
408 this is the case because it is the nearest annotated element centromeric to rs528350911, even
409 though it is ~330kb away. However, very low expression makes it challenging to quantify the
410 impact of enhancer knockout on any individual lncRNA from our RNA-seq data. We chose to
411 evaluate them in aggregate in the PAS analysis, however we recognize it is possible individual
412 lncRNAs may have molecular consequences.

413 3. While we have generated molecular evidence for the role of rs528350911 in pancreatic
414 differentiation, it is still plausible the SNP may not directly cause metabolic disorders and further
415 studies are needed.

416

417 **Acknowledgments**

418 We thank Dr. Thomas Vierbuchen for valuable advice. We thank Hanuman Kale and Emily
419 DeBitetto for assisting with experiments not included in the manuscript and acknowledge the
420 assistance from the following Memorial Sloan Kettering Cancer Center (MSKCC) Cores:
421 Integrated Genomics Operation, Flow Cytometry, Gene Editing & Screening, Molecular
422 Cytogenetics, and Stem Cell Research. We thank Dr. Ralph Garippa and Sanjoy Mehta for
423 assistance with CRISPR library generation and sequencing, Cassidy Cobbs for providing advice
424 regarding next-generation sequencing experiments, Dr. Stephanie Chrysanthou for conducting
425 nanopore sequencing with adaptive sampling, and Dr. Andrea Farina for designing and executing
426 the ddPCR assays.

427 ## Author Contributions

428 Conceptualization: SJK, DH

429 Methodology: SJK, JY, WW, JP, HSC, RL, DM, KD, EA, CSL, DH

430 Investigation: SJK, JY, WW, JP, HSC, JLI, JK, JZ, QL, DM, RL

431 Visualization: SJK, WW, QL

432 Funding acquisition: DH

433 Project administration: DH

434 Supervision: SJK, DH

435 Writing (original draft): SJK, DH

436 Writing (review & editing): all authors

437 ## Funding

438 This study was funded in part by the National Institutes of Health grant U01DK128852 (C.S.,

439 D.H., E.A.), National Institutes of Health grant U01HG012051 (D.H.), National Institutes of

440 Health grant R01DK096239 (D.H.), Starr Tri-I Stem Cell Initiative #2019-001 (D.H. and E.A.),

441 NIH/NCI MSKCC Cancer Center Support Grant P30CA008748, and a National Institutes of

442 Health T32 training grant T32GM008539 (S.J.K.).

443 ## Competing interests

444 Authors declare that they have no competing interests.

445 # STAR★Methods

446 KEY RESOURCES TABLE

447 RESOURCE AVAILABILITY

448 - Lead contact

449 - Materials availability

450 - Data and code availability

451 EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

452 - Primary Cells

453 METHOD DETAILS

454 - hPSC culture

455 - hPSC directed pancreatic differentiation

456 - gRNA library assembly

457 - Lentiviral gRNA libraries

458 - CRISPR dCas9-KRAB screens

459 - gRNA sequencing and screen analysis

460 - Generation of *ONECUT1e-664kb* clonal knockout hPSC lines

461 - Generation of clonal hPSC lines with disease-associated variant

462 - *ONECUT1* locus SNP phasing

463 - *ONECUT1e-664kb* heterozygous knockout line phasing

464 - *ONECUT1e-664kb* heterozygous rs528350911 knock-in line phasing

465 - Immunofluorescence staining

466 - Flow cytometry analysis

467 - RNA-seq and analysis

468 - Droplet digital PCR (ddPCR) and analysis

469 - Hi-C and analysis

470 - ChIP and ChIP-seq assays

471 - H3K27ac ChIP-seq analysis

472 - chromBPNet variant modeling

473 QUANTIFICATION AND STATISTICAL ANALYSIS

474 **KEY RESOURCES TABLE**

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
SOX17	BD BIOSCIENCES	Cat#561591; RRID: AB_10717121
CXCR4	R&D Systems	Cat#FAB170A; RRID: AB_357073
PDX1	R&D Systems	Cat#AF2419; RRID: AB_355257
ONECUT1	Santa Cruz Biotechnology	Cat#sc-376308; RRID: AB_10988385
NKX6-1	DHSB	Cat#F55A12; RRID: AB_532379
Alexa Fluor 488	ThermoFisher Scientific	Cat#A-11055; RRID: AB_2534102
Alexa Fluor 647	ThermoFisher Scientific	Cat#A-31571; RRID: AB_162542
LIVE/DEAD™ Fixable Violet Dead Cell Stain Kit	ThermoFisher Scientific	Cat#L34964
FOXA2	Cell Signaling Technology	Cat#8186; RRID: AB_10891055
GATA4	Invitrogen	Cat#MA5-15532; RRID: AB_10989032
GATA6	Cell Signaling Technology	Cat#5851; RRID: AB_10989032
H3K27ac	EMD Millipore	Cat#MABE647; RRID: AB_2893037
Bacterial and virus strains		
NEB 5-alpha Competent E. coli	NEB	Cat#C2987
Chemicals, peptides, and recombinant proteins		
Complete Essential 8 (E8) medium	Thermo Fisher Scientific	A1517001
Vitronectin (VTN-N)	Thermo Fisher Scientific	A14700
EDTA	KD Medical	RGE-3130
TrypLE Select 1X	Thermo Fisher Scientific	12563-029
ROCK inhibitor Y-27632 2HCl	Selleck Chemicals	S1049
Doxycycline	Sigma	D9891
Lipofectamine RNAiMAX Transfection Reagent	Thermo Fisher Scientific	13778150
Opti-MEM	Thermo Fisher Scientific	31985070
MCDB 131	Thermo Fisher Scientific	10372019
Activin A	PeproTech	120-14E

CHIR 99021	Tocris	442310
L-Ascorbic acid (vitamin C)	Sigma-Aldrich	A4544
FGF7 (KGF)	PeproTech	100-19
SANT-1	Sigma	S4572
Retinoic acid	Sigma	R2625
LDN-193189	Reprocell	04-0074
TPB	EMD Millipore	565740
ITS-X (100X)	Thermo Fisher Scientific	51500-056
Dynabeads Protein G	Thermo Fisher Scientific	10004D
Critical commercial assays		
Herculase II Fusion DNA Polymerase	Agilent Technologies	600679
LongAmp Taq DNA Polymerase	NEB	M0323
eBioscience™ Foxp3 / Transcription Factor Staining Buffer Set	Thermo Fisher Scientific	00-5523-00
Arima-Hi-C kit	Arima	A510008
Monarch® HMW DNA Extraction Kit	NEB	T3050
TruSeq Stranded Total RNA	Illumina	20020597
Zero Blunt TOPO PCR Cloning Kit	Thermo Fisher Scientific	450245
Deposited data		
Raw and processed data	This paper	GSE267330
Raw and processed data	Lee et al., 2019	GSE114102
Raw and processed data	Yang et. al 2022	GSE181480
Raw and Processed Data	Luo et al., 2023	4DNFIKCEWMIA, 4DNFI7RGXYFY, 4DNFIRLG5UWL
Experimental models: Cell lines		
H1 dCas9-KRAB line	Pulecio et al., 2023	Derived from the parental line (NIHhESC-10-0043); RRID: CVCL_WS14
H1 iCas9 line	Shi et al., 2017	Derived from the parental line (NIHhESC-10-0043); RRID: CVCL_WS14
Oligonucleotides		
See Table S1 Sheet 3 for screen gRNA spacer sequences, Table S3 for all other sequences	This paper	N/A
Software and algorithms		
PRISM v10	GraphPad Software	https://www.graphpad.com/; RRID: SCR_002798
Flowjo-v10	FlowJo LLC	https://www.flowjo.com/; RRID: SCR_008520
R	4.4.0	https://cran.r-project.org; RRID: SCR_003005
Genrich	0.6.1	https://github.com/jsh58/Genrich; RRID: SCR_025320
CHOPCHOPv3		https://chopchop.cbu.uib.no/; RRID: SCR_015723
MAGeCK	0.5.9.5	https://sourceforge.net/p/mageck/wiki/Home/ ; RRID: SCR_025016
Minimap2	2.26	https://github.com/lh3/minimap2; RRID: SCR_018550
samtools	1.2	http://htslib.org/; RRID: SCR_002105
WhatsHap	2.2	https://whatshap.readthedocs.io/; RRID: SCR_025319

longshot	0.4.5	https://github.com/pjedge/longshot; RRID: SCR_025318
STAR	2.7.11b	https://github.com/alexdobin/STAR; RRID: SCR_004463
Subread featureCounts	2.0.6	https://subread.sourceforge.net/; RRID: SCR_009803
QuantaSoft Analysis Pro	1.0.596	https://www.bio-rad.com; RRID: SCR_025321
HiC-Pro	3.1.0	https://github.com/nservant/HiC-Pro; RRID: SCR_017643
DESeq2	1.40.1	https://bioconductor.org/packages/release/bioc/html/DESeq2.html; RRID: SCR_015687
HiTC	1.34.0	http://www.bioconductor.org/packages//2.10/bioc/html/HiTC.html; RRID:SCR_013175
HiCDCPlus	1.5.2	https://www.bioconductor.org/packages/release/bioc/html/HiCDCPlus.html; RRID: SCR_025317
Bowtie2	2.5.1	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml; RRID: SCR_016368
MACS2	2.2.9.1	https://github.com/macs3-project/MACS; RRID: SCR_013291
IDR	2.0.3	https://github.com/nboley/idr; RRID: SCR_017237
IGV	2.17.4	http://www.broadinstitute.org/igv/; RRID:SCR_011793
GATK	4.0.1.2	https://software.broadinstitute.org/gatk/; RRID:SCR_001876
ChromBPNet	0.1.7	https://github.com/kundajelab/chrombpnet; RRID: SCR_024806
MEME Suite (FIMO)	5.5.5	http://meme-suite.org/; RRID:SCR_001783
bedtools	2.31.1	https://github.com/arq5x/bedtools2; RRID:SCR_006646
Other		

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to Danwei Huangfu (huangfud@mskcc.org).

Materials availability

All the hPSC cell lines and plasmids generated in this study are available upon reasonable request.

Data and code availability

- All the sequencing data generated in this study are available at GEO under accession code GSE267330. The Hi-C data from the PFG cells are also available in the 4DN Nucleome Data Portal (https://data.4dnucleome.org) under accession number 4DNFIQBPRZZD. Previously published data that were re-analyzed here are available at GEO under codes GSE114102 (ATAC-seq data used for screen design, and ATAC-seq and H3K27ac ChIP-seq visualization in Sup. Fig. 1E) and GATA4, GATA6, and FOXA2 ChIP-seq) and GSE181480 (HHEX and ONECUT1 ChIP-seq), and the 4DN Nucleome Data Portal, accession numbers: 4DNFIKCEWMIA, 4DNFI7RGXYFY, 4DNFIRLG5UWL. All genomic coordinates reported in manuscript are hg38 unless otherwise indicated.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Primary cells

The parental hESC line H1 (NIHhESC-10-0043) is available from WiCell under a material transfer agreement.

METHOD DETAILS

502 ## hPSC culture

503 Experiments were performed with H1 human embryonic stem cells (NIHhESC-10-0043), which
504 were regularly confirmed to be mycoplasma-free by the Memorial Sloan Kettering Cancer
505 Center (MSKCC) Antibody & Bioresource Core Facility. All experiments were conducted per
506 NIH guidelines and approved by the Tri-SCI Embryonic Stem Cell Research Oversight
507 (ESCRO) Committee. hPSCs were maintained in Essential 8 (E8) medium (Thermo Fisher
508 Scientific, A1517001) on vitronectin (Thermo Fisher Scientific, A14700) pre-coated plates at
509 37°C with 5% CO_2. The Rho-associated protein kinase (ROCK) inhibitor Y-276325 (5 μM;
510 Selleck Chemicals, S1049) was added to the E8 medium immediately after passaging or thawing
511 of hPSCs for 24 hours.

512 ## hPSC directed pancreatic differentiation

513 hPSCs were maintained in E8 medium for 2 days to reach ~80% confluence. Cells were washed
514 with PBS and differentiated to DE, GT, PFG, and PP2 stages following previously described
515 protocols[17,102]. hPSCs were rinsed with PBS and first differentiated into DE using S1/2 medium
516 supplemented with 100 ng/ml Activin A (Bon Opus Biosciences) for 3 days and CHIR99021
517 (Stemgent, 04-0004-10) for 2 days (first day, 5 μM; second day, 0.5 μM). DE cells were rinsed
518 with PBS and then exposed to S1/2 medium supplemented with 50 ng/ml KGF (FGF7)
519 (PeproTech, 100-19) and 0.25 mM vitamin C (VitC) (Sigma-Aldrich, A4544) for 2 days to reach
520 GT stage. GT cells were then switched to S3/4 medium supplemented with 50 ng/ml FGF7,
521 0.25 mM VitC and 1 μM retinoic acid (RA) (Sigma-Aldrich, R2625) for 2 days to reach PFG
522 stage. PFG cells were then switched to S3/4 medium supplemented with 2 ng/ml FGF7, 0.25 mM
523 VitC, 0.1 μM RA, 200 nM LDN, 0.25 μM SANT-1, 100 nM TPB and 1:200 ITS-X for 4 days to
524 reach PP2 stage.

525 ## gRNA library assembly

526 23 transcription factors expressed in pancreatic differentiation were selected based on known
527 clinical roles in metabolic disease and demonstrated pancreatic functional requirement from
528 CRISPR screens[29,56-59]. Around each transcription factor, loci within hg19 (lifted over in
529 supplemental tables to hg38) were defined to include at least the TAD containing the gene, with
530 TADs called on PFG-stage Hi-C data as previously done[67]. To select putative enhancers for
531 interrogation in these loci, Genrich (0.6.1)[103] was used to call peaks at an FDR-adjusted p-value
532 of 0.01 on previously published DE, GT, and PFG stage ATAC-seq data (Sup. Table 1)[25]. For
533 each region in the loci, and some additional regions, CHOPCHOPv3[104] was used to design
534 gRNAs. For regions <750 bp, the top 5 ranked gRNAs were taken, and for regions >750 bp, 1
535 gRNA/150bp of sequence was selected for interrogation. Finally, 1100 safe harbor[65] and 463
536 non-targeting gRNAs[66] were included in the library for a total of 38,747 gRNAs. Oligos were
537 synthesized, amplified, and restriction cloned into the lentiGuide-puro[105] (Addgene: 52963;
538 RRID: Addgene_52963) backbone by the MSKCC Gene Editing & Screening Core Facility.
539 Cloned plasmid libraries were PCR amplified to incorporate adapters for NGS. Samples were
540 purified and sequenced using Illumina HiSeq 2500 platform. FASTQ files were clipped by
541 position and reads were mapped back to the reference library file to quantify abundance of reads
542 per gRNA. The overall representation of the libraries was charted over a one-log fold change to
543 evaluate representation in the final library.

544 ## Lentiviral gRNA libraries

545 gRNA lentiviral library generation was performed as previously described[17]. In brief, a total of
546 9.45 µg of the library plasmid combined with 6.75 µg lentiviral packaging vector psPAX2 and
547 1.36 µg vesicular stomatitis virus G (VSV-G) envelope expressing plasmid pMD2.G (Addgene

548 plasmids 12260 and 12259) were transfected with the JetPRIME (VMR; 89137972) reagent into
549 10E6 293T cells in a 10cm plate to produce the lentiviral particles. Fresh medium was changed
550 24h after transfection and viral supernatant was collected at 48 and 72h after transfection, spun at
551 1000 rpm for 5 minutes, 0.45um filtered, and stored at −80°C.

552 ## CRISPR dCas9-KRAB screens

553 Two independent screens were performed following the same procedure. H1 hPSC (i)dCas9-
554 KRAB cells were used[67]. The lentiviral library was transduced into 12E6 hPSCs distributed in
555 three 10 cm plates following the conditions (MOI: 4, protamine sulfate: 6 µg/ml, Ri: 10 µM).
556 After one day of recovery, doxycycline (DOX, 2 µg/ml) was added, and cells were kept under
557 puromycin selection for 2 days. After two days, 30E6 cells were seeded into 6-well VTN-coated
558 plates (5E5 per well) and grown for 48 hours in E8 medium. After washing with Phosphate
559 Buffered Saline (PBS) w/o Ca2+ & Mg2+, cells were differentiated, with DOX added
560 throughout the seven-day differentiation process. At the PFG stage, cells were dissociated with
561 TrypLE, stained with LIVE/DEAD reagent, fixed, stained with the PDX1 antibody, and stained
562 with a fluorescent secondary antibody. They were then sorted by the MSKCC flow cytometry
563 facility based on 2ndary antibody fluorescence using FACSAria sorters (BD Biosciences).
564 Positive and negative cells were sorted aiming for a minimum 1000X representation (number of
565 cells sorted*MOI)/(unique gRNAs in library) per condition, with two independent sorts serving
566 as technical replicates for each biological replicate. For biological replicate 1, approximately
567 10E6 PDX1- and 10E6 PDX1+ cells were sorted. For replicate 2, approximately 20E6 PDX1-
568 and 20E6 PDX1+ cells were sorted. Cells were pelleted and kept at -80°C for downstream gRNA
569 sequencing.

570 ## gRNA sequencing and screen analysis

571 gRNA enrichment sequencing was performed by the MSKCC Gene Editing & Screening Core
572 Facility as previously described, but with modifications to enable DNA extraction from fixed
573 cells[17]. Cells were decrosslinked and treated with proteinase K (0.1 mg/ml) overnight followed
574 by genomic DNA extraction with QIAGEN Blood & Cell Culture DNA Maxi Kit (QIAGEN;
575 13362) and quantified by Qubit (Thermo-Scientific) following the manufacturer's guidelines. A
576 quantity of gDNA sufficient for 1000X representation of gRNAs was amplified with oligos
577 containing the Illumina adapters and multiplexing barcodes by PCR. Amplicons were quantified
578 by Qubit and Bioanalyzer (Agilent) and sequenced on the Illumina HiSeq 2500 platform. The
579 gRNA library sequences were used to align the sequencing reads and the counts for each gRNA
580 were determined with MAGeCK. Screens were analyzed with MAGeCK RRA (0.5.9.5), and the
581 two biological screen replicates were analyzed independently (Sup. Table 1). To calculate a
582 single value for the significance of effect, a ratio was taken between the positive and negative
583 MAGeCK scores and log-transformed. The top 36 hits associated with a PDX1 decrease had a
584 MAGeCK neg.FDR<0.1 in replicate 2 and caused a PDX1 decrease in both replicates (Sup.
585 Table 1). Hits overlapped with annotated promoter regions were designated as "Promoter Hits,"
586 with the rest designated as "Non-Promoter Hits." Promoter overlap was defined as overlap with
587 an interrogated region ± 1.5kb of Gencode (v18) TSS sites. For Figure 1F, if there were multiple
588 overlapped regions, the most significant hit was used for plotting. To link enhancers to the
589 putatively regulated genes, the nearest linearly adjacent "Promoter Hit" was inferred to be the
590 regulated gene of the "Non-Promoter Hit." To determine hit and non-hit overlap with H3K27ac
591 ChIP-seq peaks during differentiation, Genrich (0.6.1)[103] was used to call peaks at an FDR-
592 adjusted p-value of 0.01 on previously published ES, DE, GT, and PFG stage H3K27ac ChIP-seq
593 data[25]. Conservation scores of regions were determined by calculating the average of per

594 nucleotide conservation across all nucleotides within each region from PhyloP100 per-nucleotide
595 conservation scores[69].

596 Generation of *ONECUT1e-664kb* clonal knockout hPSC lines

597 H1 iCas9 hPSC lines[62] were used to generate three enhancer deletion lines carrying *ONECUT1e-*
598 *664kb* deletions. Two gRNAs targeting the flanking regions of the enhancer were designed, and
599 knockouts generated as previously described but with some modifications[8]. gRNAs and tracer
600 RNA were ordered from IDT (Alt-R CRISPR-Cas9 crRNA, 1072532) and added at a 15 nM final
601 concentration. In brief, gRNA/tracer RNA and Lipofectamine RNAiMAX (Thermo Fisher
602 Scientific, 13778030) were diluted separately in Opti-MEM (Invitrogen, 31985070), mixed
603 together, and incubated for 15 min at room temperature, and then added dropwise to freshly
604 seeded iCas9 hPSCs in a 24-well plate. DOX (2 µg/ml) was added the day before transfection,
605 the day of transfection, and 1 day after transfection to induce Cas9 expression. Three days after
606 transfection, hPSCs were dissociated into single cells and ~1-3E3 cells were plated into a 10 cm
607 tissue culture dish for colony formation. After ~10 days of expansion, single colonies were
608 picked. Genomic DNA from crude cell lysate was used for PCR genotyping (primers:
609 ONECUT1e_KOgenot_F, ONECUT1e_KOgenot_R). Heterozygous lines contained a band
610 corresponding to the WT band and a smaller band corresponding the knockout band, while
611 homozygous lines only had smaller bands. Further genotyping was done through TOPO cloning
612 of the genotyping amplicons which determined the exact genotypes of both alleles for all lines
613 (Sup. Table 2). gRNA target sequences and primers used for PCR and sequencing are listed in
614 Supplementary Table 3.

615 Generation of clonal hPSC lines with disease-associated variant

616 A process similar to enhancer deletion line generation was used to generate heterozygous knock-
617 in SNP lines carrying the rs528350911 variant. During transfection, a single gRNA proximal to
618 the variant site was used, and a ssDNA HDR template containing the disease-associated variant
619 was added to the mixture at the Lipofectamine transfection step (Sup. Table 3) at a concentration
620 of 20 nM. Genomic DNA from crude cell lysate was used for PCR (primers:
621 ONECUT1e_KOgenot_F, ONECUT1e_KOgenot_R). All picked clones were sequenced with
622 both primers, and heterozygous lines picked by the presence of a double peak on the spectrogram
623 that corresponded to the insertion of the disease-associated variant. For lines containing the
624 introduced variant, no indels or other changes from the reference genome were detected in the
625 584bp amplicon.

626 ## *ONECUT1* locus SNP phasing

627 To phase all SNPs in the *ONECUT1* locus, Nanopore sequencing with adaptive sampling of the
628 region chr15:52,707,645-53,586,740 (and controls regions) was conducted on an H1 iCas9 line
629 with the rs528350911 heterozygous variant (line N11). DNA was extracted using Monarch®
630 HMW DNA Extraction Kit for Cells & Blood (NEB, T3050), according to the manufacturer's
631 instructions. DNA quantity was measured using a QuantiT (Thermo Fisher Scientific), purity on
632 a NanoDrop (Thermo Fisher Scientific) and fragment-size distribution on a TapeStation
633 (Agilent). Prior to ONT library preparation, DNA was sheared to ~15-20 kb fragment size using
634 Covaris g-TUBE. Sequencing libraries were prepared from ~1–2ug of DNA using the native
635 library preparation kit SQK-LSK114 according to the manufacturer's instructions. Each library
636 was loaded onto a PromethION flow cell R10.4.1 and sequenced on an ONT PromethION P24
637 device. The sample was run for 72 h with 1 nuclease flush and reload performed during the run
638 to maximize sequencing yield.

639 Reads were aligned with minimap2 (2.26)[106], with samtools[107] coverage estimating ~138X
640 coverage across the region enriched with adaptive sampling. WhatsHap (2.2)[108] was used to
641 phase the variants in the locus as well as assemble a haplotype containing the *ONECUT1* gene,
642 enhancer, and surrounding regions (Sup. Fig. 2D). WhatsHap was also used to phase-tag reads
643 for visualization in Sup. Fig. D and E. From these results, a heterozygous SNP was identified in
644 the 3'UTR of the *ONECUT1* transcript, as well as the enhancer-proximal Genotyping SNP.

645 ## *ONECUT1e-664kb* heterozygous knockout line phasing

646 To determine which allele contained the knocked-out enhancer in the heterozygous enhancer
647 deletion lines (Het. 1, Het. 2), a forward primer was designed centromeric to the Genotyping
648 SNP and a reverse primer targeting the enhancer sequence that would be present on the non-
649 deletion allele. Since line designated as Het. 1 contained an inversion of the WT enhancer
650 sequence, two different reverse primers were needed. Thus, for Het. 1, and amplicon was
651 generated with the primer pair ONECUT1e_GenotSNP_F,
652 ONECUT1e_GenotSNP_R_WT_inv_e, and for Het.2, the pair ONECUT1e_GenotSNP_F,
653 ONECUT1e_GenotSNP_R_WT_e. The amplicon was sent for sanger sequencing with the
654 ONECUT1e_GenotSNP_F primer used as the sequencing primer. The Genotyping SNP from
655 this sequencing was used to determine the allele containing the deletion.

656 ## *ONECUT1e-664kb* heterozygous rs528350911 knock-in line phasing

657 A PCR amplifying an ~8.1kb fragment containing both the T2D-associated variant and the
658 Genotyping SNP (primers: ONECUT1e_GenotSNP_F, ONECUT1e_GenotSNP_R_WT_e) with
659 the LongAmp Taq polymerase kit was conducted. The amplicon was sent to Plasmidsaurus for
660 linear fragment sequencing. Reads were aligned with minimap2 (2.26), and variants were phased
661 with longshot (0.4.5)[109].

662 ## Immunofluorescence staining

663 PFG-stage pancreatic differentiated cells were fixed in 4% paraformaldehyde (Thermo Fisher
664 Scientific, 50980495) for 10 min at room temperature. After washing with PBST (PBS with
665 0.1% Triton X-100) three times, cells were blocked in 5% donkey serum in PBST buffer for
666 30 min at room temperature. Primary and second antibodies were diluted in the blocking solution
667 (Sup. Table 4). Cells were incubated with primary antibodies overnight at 4°C, followed by 1 hr
668 staining for secondary antibodies at room temperature. The cells were then stained with 4′,6-
669 diamidino-2-phenylindole (DAPI) for ~15 min at room temperature. Images were taken using a
670 confocal laser scanning platform (Leica TCS SP5).

671 ## Flow cytometry analysis

672 Cells were dissociated using TrypLE Select and resuspended in FACS buffer (5% FBS in PBS).
673 LIVE/DEAD Fixable Violet Dead cell stain (Invitrogen, L34955) was used to discriminate dead
674 cells from live cells. LIVE/DEAD staining was performed per manufacturer's instructions in
675 FACS buffer, followed by fixation and intracellular staining with a FOXP3 staining buffer set
676 (eBioscience, 00-5523-00) following the manufacturer's instructions. Permeabilization/fixation
677 was performed at room temperature for 30 minutes. Antibody staining was performed in
678 permeabilization buffer for 30 minutes. Antibodies for this study are listed in Supplementary
679 Table 8. Cells were then analyzed using BD LSRFortessa. Flow cytometry analysis and figures
680 were generated using FlowJo v.10.

681 ## RNA-seq and analysis

682 RNA was extracted from differentiated cells with the Zymo Quick-RNA MiniPrep kit (R1055).
683 TruSeq stranded total (Ribodepletion) RNA-seq was performed on three independent
684 differentiations by the MSKCC Integrated Genomics Operation (IGO) Core. Briefly, STAR

685 alignment was performed followed by featureCounts[110] quantification (Sup. Table 6) and
686 normalization with DESeq2 median of ratios prior to visualization in Fig. 3A, Fig. 3C, and Sup.
687 Fig. 3B. Differential gene expression was performed and adjusted p-values computed with
688 DESeq2. Strand-specific analysis of promoter antisense transcription analysis was performed
689 with featureCounts strand specific quantification of a region telomeric to *ONECUT1*
690 (chr15:52805588-53391232). For strand-specific transcript visualization, samtools was used to
691 extract sense and antisense alignments from .bam files from which bigwig coverage files were
692 generated.

693 ## Droplet digital PCR (ddPCR) and analysis

694 Assays specific for the detection of SNPs (Sup. Table 5) were designed (using Primer3Plus, if
695 needed) and ordered through Bio-Rad. Cycling conditions were tested to ensure optimal
696 annealing/extension temperature as well as optimal separation of positive from empty droplets.
697 Optimization was done with a known positive control. A constant amount of RNA or DNA was
698 used within each experiment. Reactions were partitioned into a median of ~17,500 droplets per
699 well using the QX200 droplet generator. Plates were read and files analyzed with the QuantaSoft
700 software to assess the number of positive droplets.
701 Procedure for RNA: after PicoGreen quantification of RNA, droplet generation was performed
702 on a QX200 ddPCR system (Bio-Rad catalog # 1864001) using cDNA generated from 0.1-2 ng
703 total RNA with the One-Step RT-ddPCR Advanced Kit for Probes (Bio-Rad catalog # 1864021)
704 according to the manufacturer's protocol with reverse transcription at 42°C and
705 annealing/extension at 60°C.
706 Procedure for DNA: after PicoGreen quantification of DNA, 0.25-9 ng gDNA of ChIP-ed DNA
707 was combined with locus-specific primers; FAM- and HEX-labeled probes; Hae III (DNA

708 extracted form hPSC SNP lines), Hind III (ChIP-ed DNA), or Mse I (DNA extracted
709 *ONECUT1e-664kb* hPSC knockout lines); and digital PCR Supermix for probes (no dUTP).
710 Emulsified PCRs were run on a 96-well thermal cycler using cycling conditions identified during
711 the optimization step (95°C 10'; 40 cycles of 94°C 30' and 53°C (rs528350911 assays only) or
712 60°C 1'; 98°C 10'; 4°C hold).

713 ## Hi-C and analysis

714 Hi-C was performed as previously described with cells differentiated to the PFG stage processed
715 with the Arima Hi-C kit (Arima, A510008)[8]. HiC-Pro (3.1.0)[111] was used to align the individual
716 Hi-C replicates to GRCh38, GCA_000001405.15, with alternative contigs removed, and then
717 final merged libraries were obtained by combining all *.allValidPairs files where duplicates
718 across biological replicates are maintained. DE libraries consisted of three biological replicates
719 (two HUES8, one H1 with accessions 4DNFIKCEWMIA.hic, 4DNFI7RGXYFY.hic, and
720 4DNFIRLG5UWL.hic respectively), and the PFG libraries consisted of two biological replicates
721 differentiated to the PFG stage (merged accession 4DNFIQBPRZZD.hic). For visualization, Hi-
722 C data was visualized at 25 KB resolution.
723 Virtual 4C counts for the regions of interest were extracted using utilities from plotGardener
724 (1.6.4)[112] and plotted as a one-dimensional signal. Hi-C matrix counts for the regions of interest
725 were extracted using plotGardener and plotted as a rectangular region. Observed counts are
726 library normalized prior to plotting so that the total number of reads between panels is
727 comparable. TADs that were previously identified were also overlaid on the data and can be
728 found in the Hi-C subseries of GEO data set published with this work. Counts within the vicinity
729 of the *ONECUT1* promoter (all contacts whose start or end bins sit within a 1.2Mb window
730 around the *ONECUT1* promoter), were extracted with strawr (0.0.91) and passed to DESeq2

731 (1.40.1)[113]. The Wald test was used to assess significance of all differential contacts within that
732 window, and significant hits anchored on the *ONECUT1* containing bin are reported at an FDR
733 of 0.1.
734 For calling topologically associating domains, merged Hi-C files were first ICE normalized
735 through HiTC (1.34.0)[114] and then topologically associating domains were identified with
736 HiCDCplus (1.5.2)[115] using its interface to TopDom at 50 kb. Window size for scanning was set
737 at 5.

738 ChIP and ChIP-seq assays

739 ChIP was performed as previously described[18]. For each sample, around 30E6 cells were
740 crosslinked in-plate with 1% formaldehyde for 10 minutes at 37°C and quenched with 0.125 M
741 glycine. Fixed cells were collected and washed in cold PBS buffer, snap frozen, and saved at -
742 80°C. On the day of immunoprecipitation, the cell pellet was thawed on ice, resuspended in
743 700uL SDS buffer (1% SDS, 10 mM EDTA, 50 mM Tris–HCl, pH 8), and incubated for 10min
744 on ice. Sonication was performed on a Branson Sonifier 150 set at 30% amplitude for 5.5min
745 (10s on/off pulsing). Supernatant was pre-cleared with Dynabeads Protein G (#10004D) and then
746 incubated overnight with antibodies (10 ug for transcription factors, 5 ug for H3K27ac, Sup.
747 Table 4) at 4°C.). On the next day, the ChIP samples were incubated with Dynabeads Protein G
748 for 6 hours at 4°C. Then the beads were pelleted and sequentially washed twice with low salt
749 (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris–HCl, pH 8, 150 mM NaCl), then high
750 salt (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris–HCl, pH 8, 500 mM NaCl), and
751 finally TE buffer (10 mM Tris–HCl, pH 8, 1 mM EDTA). The DNA was eluted from the beads
752 by incubating in elution buffer (1% SDS, 0.1 M NaHCO3) at 65°C for 15min and decrosslinked
753 with 190 mM NaCl at 65°C overnight. A total of 5 µl 0.5 M EDTA, 10 µl 1 M Tris–HCl, pH 6.5,

754 and 1 µl Proteinase K (20 mg/ml) were added to 260 uL of the de-crosslinked product and
755 incubated for 1 hr at 45°C. DNA was isolated by using QIAquick PCR purification kit (Qiagen,
756 28104; QIAGEN). H3K27ac ChIP samples were submitted to MSKCC Integrated Genomics
757 Operation core for NGS library preparation and sequencing. GATA4, GATA6, and FOXA2
758 ChIP samples were submitted for ddPCR.

759 ## H3K27ac ChIP-seq Analysis

760 Sequenced data were aligned to the hg38 reference genome using bowtie2 (2.5.1)[116]. Peak calling
761 was performed using MACS2 (2.2.9.1)[117] with the respective ChIP input as the control and the
762 default P value cutoff and default extension size. Irreproducible discovery rate (IDR)[118] with a
763 cutoff of 0.01 was used to filter reproducible peaks, where at least 2 pairwise comparisons of the
764 three replicates passed the cutoff. Peaks were filtered by ENCODE exclusion list version 2 for
765 hg38[119]. The quantification and differential signal intensity of the samples were calculated in
766 DESeq2[113]. The integer counts, log(fold-change), and adjusted P values are provided in
767 Supplementary Table 7. The replicates were combined, and the signal track was generated from
768 MACS2. MACS2 subtract was used to generate the difference of the signal track. All signal
769 tracks were visualized using Integrative Genomics Viewer (IGV)[120].
770 For allele-specific analysis, GATK[121] ASEReadCounter was used to quantify reads at
771 heterozygous positions phased by long-read sequencing and ratios were calculated for
772 heterozygous positions (score filters: PHASE_QUAL >= 140, and QUAL >= 200) with more
773 than 100 reads total across all WT and Het. 2 replicates. For Figure 4F, the "Phased Region" is
774 the haplotype block containing *ONECUT1* called by WhatsHap (chr15:50462878-54550826) and
775 visualized in Sup. Fig. 2I.

776 ## chromBPNet variant modeling

A ChromBPNet[81,82] model was trained to predict the local ATAC profile using bulk ATAC-seq data from the PFG stage[25]. Default hyperparameters were utilized with an input size of 2114 base pairs and an output size of 1000 base pairs for training both the bias model and the ChromBPNet model. For the bias model training, regions were selected that do not overlap with any ATAC-seq peaks in the four stages of pancreas differentiation[25], applying a bias threshold factor of 0.3 to filter out high-count regions in the sets. This resulted in a training set of 54957 regions. For ChromBPNet training, ATAC-seq peaks only from the PFG stage with IDR values greater than 830 were used, resulting in a training set of 52541 regions. During training for both models, peaks from chromosomes 15 and 13 for validation were held out. For each specific region of interest, the 2114 base pairs around the summit center was used as input and the DeepLift attribution score obtained using the trained ChromBPNet model. Specifically, the original attribution pipeline from the ChromBPNet GitHub repository was employed, which involves computing the DeepLift score over 20 reference inputs shuffled from the original sequence. To find the GATA motifs, FIMO[122] was used on our region of interest using the positional weight matrix (PWMs) of all GATA variants from the CIS-BP2.00 dataset[123], and motif hits with p-values < 0.05 were filtered out.

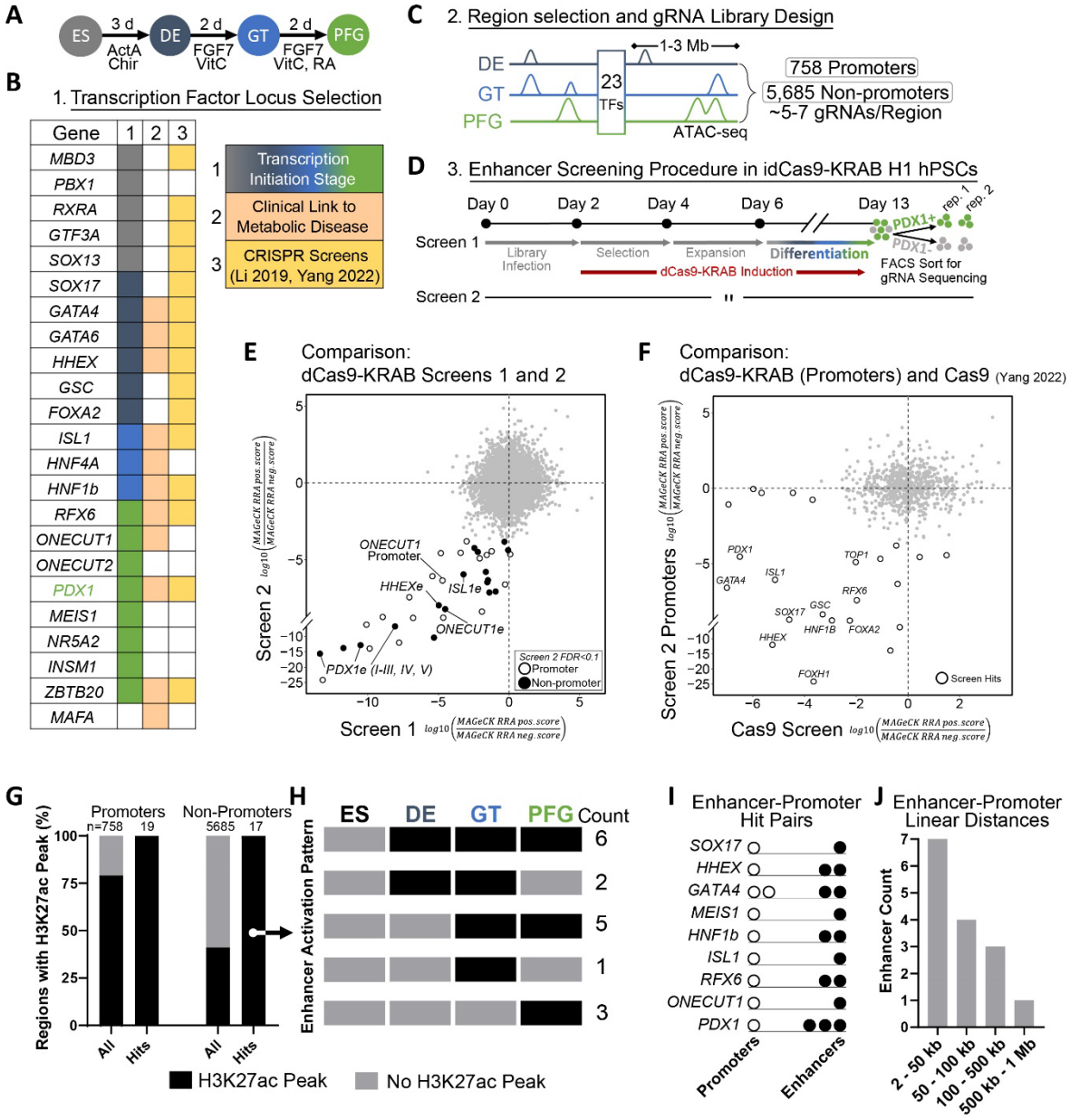
QUANTIFICATION AND STATISTICAL ANALYSIS

All datapoints refer to biological repeats. No statistical method was used to predetermine sample sizes. The investigators were not blinded to allocation during experiments and outcome assessment. No data were excluded from the analyses unless the differentiation experiment itself failed. The number of biological and technical replicates are reported in the legend of each figure. Flow cytometry analysis was derived from at least three independent experiments. For

800 ChIP-seq and bulk RNA-seq, quantification and statistics were derived from at least three
801 independent experiments. CRISPR dCas9-KRAB screening was performed twice. Quantification
802 is shown as the mean ± s.d. All the statistical analysis methods are indicated in the figure legends
803 and methods.
804

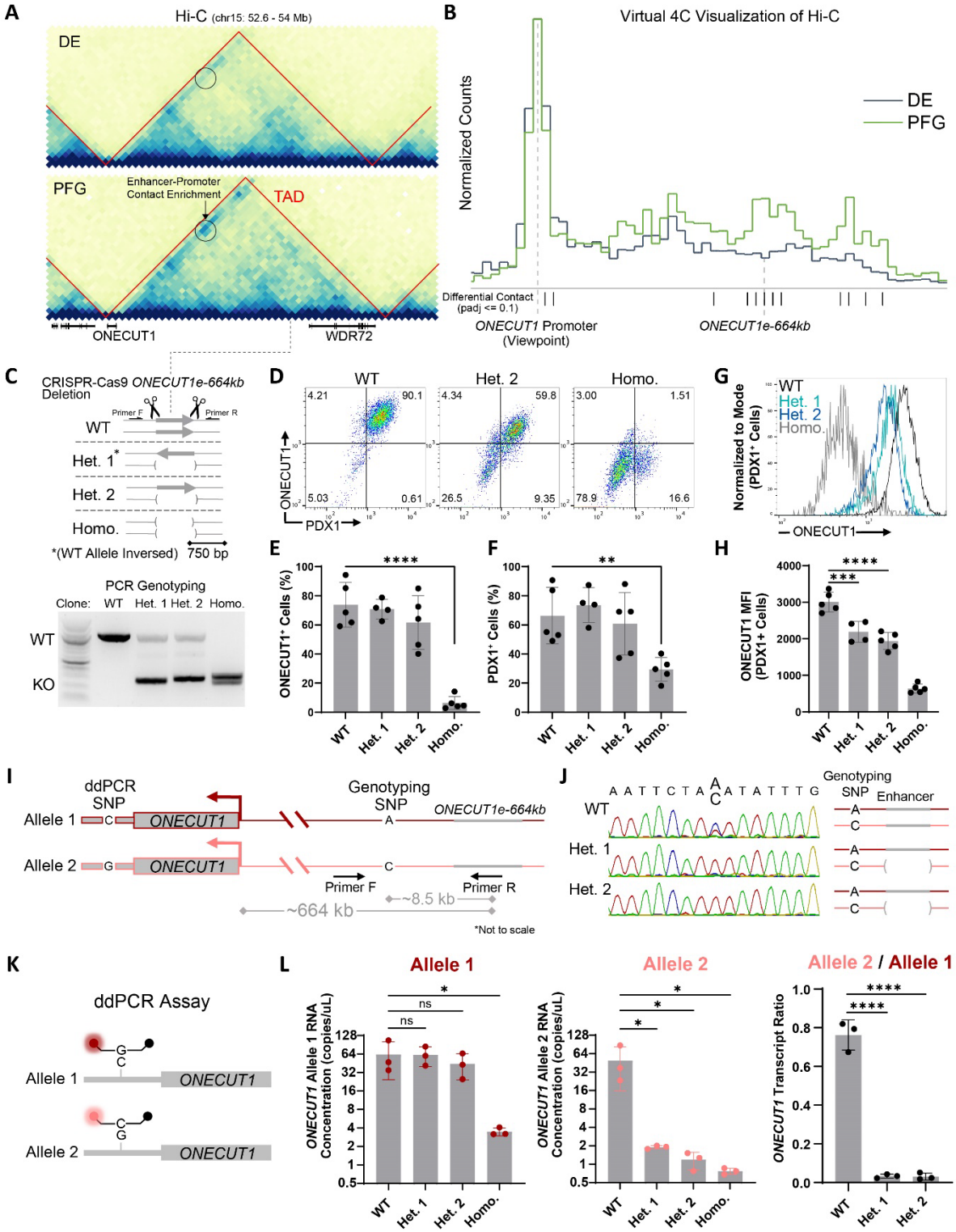
Main Figures:

Fig. 1



809 **Fig. 1: CRISPRi repression screen to discover pancreatic differentiation enhancers**
810 A) hPSC stepwise pancreatic differentiation protocol schematic. ActA, Activin A; CHIR,
811 CHIR99021; RA, retinoic acid; VitC, vitamin C.
812 B) Gene selection rationale for enhancer discovery screen.
813 C) Putative enhancer region selection and gRNA design schematic.
814 D) Screening procedure schematic.
815 E) Results from two screen replicates, each point represents a genomic region. Highlighted are
816 36 regions with a FDR < 0.1 in replicate 2, and have a decrease in PDX1 in both replicates.
817 F) Results from screen replicate 2 compared to whole genome PDX1 expression differentiation
818 screen. Genes showing enrichment in both screens are labeled.
819 G) Region overlap with H3K27ac peaks at any stage during differentiation (ES, DE, GT, PFG).
820 H) Differentiation stages at which non-promoter screen hits overlap with H3K27ac peaks, and
821 the number of hits with each activation pattern.
822 I) Enhancer-gene pair assignment.
823 J) Enhancer-gene pair linear distances.

824 **Fig. 2**

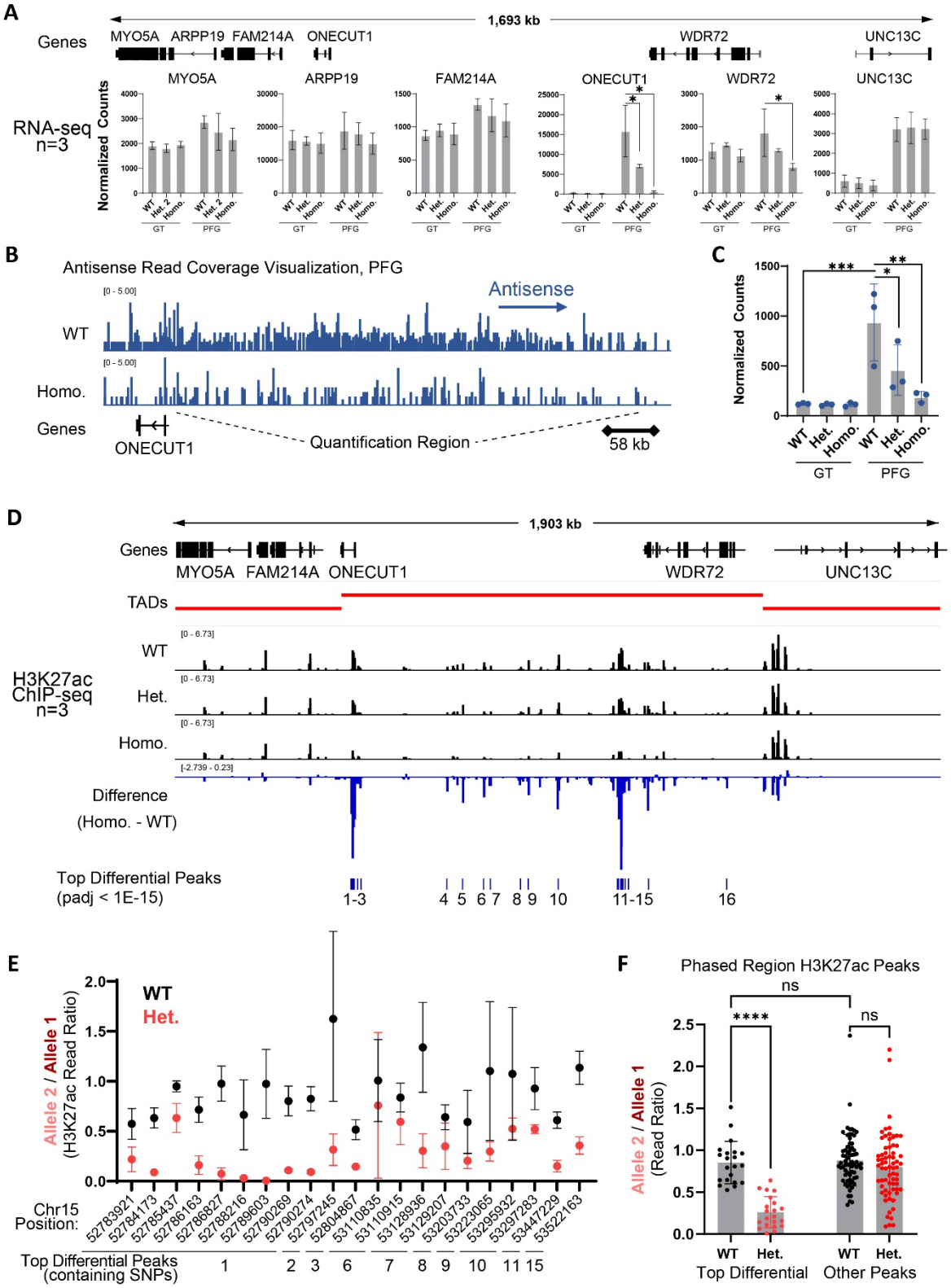


825

826

827 **Fig. 2: *ONECUT1e-664kb* deletion and pancreatic differentiation characterization**
828 A) Hi-C data of definitive endoderm and posterior foregut stage differentiated hPSCs are shown
829 around the ONECUT1 locus. Contact between ONECUT1e-664kb and ONECUT1 promoter
830 regions is circled to highlight the increase in contact frequency between the two loci. Red lines
831 demarcate TAD boundaries identified at 50 kb.
832 B) 1D slices of the 2D contact map were obtained with the ONECUT1 promoter set as the bin of
833 origin. Hi-C data was library scaled and shows the changes observed in (A). Significantly
834 different bins were determined by DESeq2.
835 C) hPSC enhancer deletion and genotyping PCR schematic.
836 D) Representative flow cytometry plots of WT and enhancer knockout cells differentiated to the
837 PFG stage.
838 E) Percentage of cells achieving a ONECUT1+ identity. For panels E, F, and H, each dot
839 represents one independent experiment (n = 4 or 5 independent experiments, 4 for Het. 1) and
840 data are presented as the mean ± s.d. One-way analysis of variance (ANOVA) followed by
841 Dunnett multiple comparisons test versus WT control.
842 F) Percentage of cells achieving a PDX1+ identity.
843 G) Representative histograms of ONECUT1 MFI of PDX1+ cells from flow cytometry plots of
844 unedited and enhancer knockout cells differentiated to the PFG stage.
845 H) Quantification and statistical comparison of ONECUT1 MFI of PDX1+ cells.
846 I) Schematic for identifying ddPCR SNP and Genotyping SNP.
847 J) Spectrogram of genotyping SNP amplicon sequencing and schematic showing allele that
848 contained the deletion in heterozygous lines.
849 K) ddPCR assay design schematic.
850 L) Allele-specific ONECUT1 transcript quantification with ddPCR on RNA extracted from PFG
851 stage cells. Each allele is shown in a separate plot, with the expression ratio shown in the last
852 plot. Each symbol represents one independent differentiation (n = 3 independent experiments),
853 with two averaged technical ddPCR replicates. Data are presented as the mean ± s.d. One-way
854 analysis of variance (ANOVA) followed by Dunnett multiple comparisons test versus WT
855 control.

856 **Fig. 3**


A
1,693 kb
Genes
MYO5A ARPP19 FAM214A ONECUT1
WDR72
UNC13C
RNA-seq
n=3
Normalized Counts
MYO5A
ARPP19
FAM214A
ONECUT1
WDR72
UNC13C
WT
Het.
Homo.
GT
PFG
B
Antisense Read Coverage Visualization, PFG
Antisense
WT
Homo.
Genes
ONECUT1
Quantification Region
58 kb
C
Normalized Counts
D
1,903 kb
Genes
MYO5A FAM214A ONECUT1
WDR72
UNC13C
TADs
H3K27ac
ChIP-seq
n=3
WT
Het.
Homo.
Difference
(Homo. - WT)
Top Differential Peaks
(padj < 1E-15)
1-3
4 5 6 7 8 9 10
11-15
16
E
WT
Het.
Allele 2 / Allele 1
(H3K27ac Read Ratio)
Chr15
Position:
52783921
52784173
52785437
52786163
52786827
52788216
52789603
52790269
52790274
52797245
52804867
53110835
53110915
53128936
53129207
53203733
53223065
53295932
53297283
53447229
53522163
Top Differential Peaks
(containing SNPs)
1 2 3 6 7 8 9 10 11 15
F
Phased Region H3K27ac Peaks
ns
****
ns
Allele 2 / Allele 1
(Read Ratio)
WT Het.
Top Differential
WT Het.
Other Peaks

857

858 **Fig. 3: Transcriptional and epigenetic consequences of *ONECUT1e-664kb* knockout**
859 A) RNA-seq normalized read counts for genes in ONECUT1 locus at GT and PFG stages from
860 WT, heterozygous (het. 2), and homozygous enhancer knockout hPSCs. * indicates padj<0.005
861 computed by DESeq2.
862 B) Visualization of antisense RNA relative to ONECUT1. Read coverage from strand-specific
863 RNA-seq of PFG-stage WT and homozygous enhancer knockout cells of representative
864 differentiation replicate.
865 C) Quantification of (B). Each dot represents one independent experiment (n = 3 independent
866 experiments) and data are presented as the mean ± s.d. One-way analysis of variance (ANOVA)
867 followed by Dunnett multiple comparisons test versus WT control.
868 D) H3K27ac ChIP-seq at PFG stage of WT, heterozygous, and homozygous enhancer knockout
869 cells. TADs called from PFG stage HiC data. Difference calculated via MACS2 subtract. Top
870 differential peaks between WT and homozygous enhancer knockout cells quantified with
871 DESeq2. ONECUT1e-664kb deletion region is within peak #12.
872 E) H3K27ac is decreased in an allele-specific manner within significantly affected H3K27ac
873 peaks.
874 F) Quantification of average H3K27ac read ratios at all heterozygous SNP positions in phased
875 locus in top significantly different peaks (WT vs Homo, padj<1E-15) and other peaks. Each
876 point represents a heterozygous SNP position, average of three replicates from independent
877 experiments. Ordinary one-way ANOVA with Šidák's multiple comparisons test, with a single
878 pooled variance.

Fig. 4:

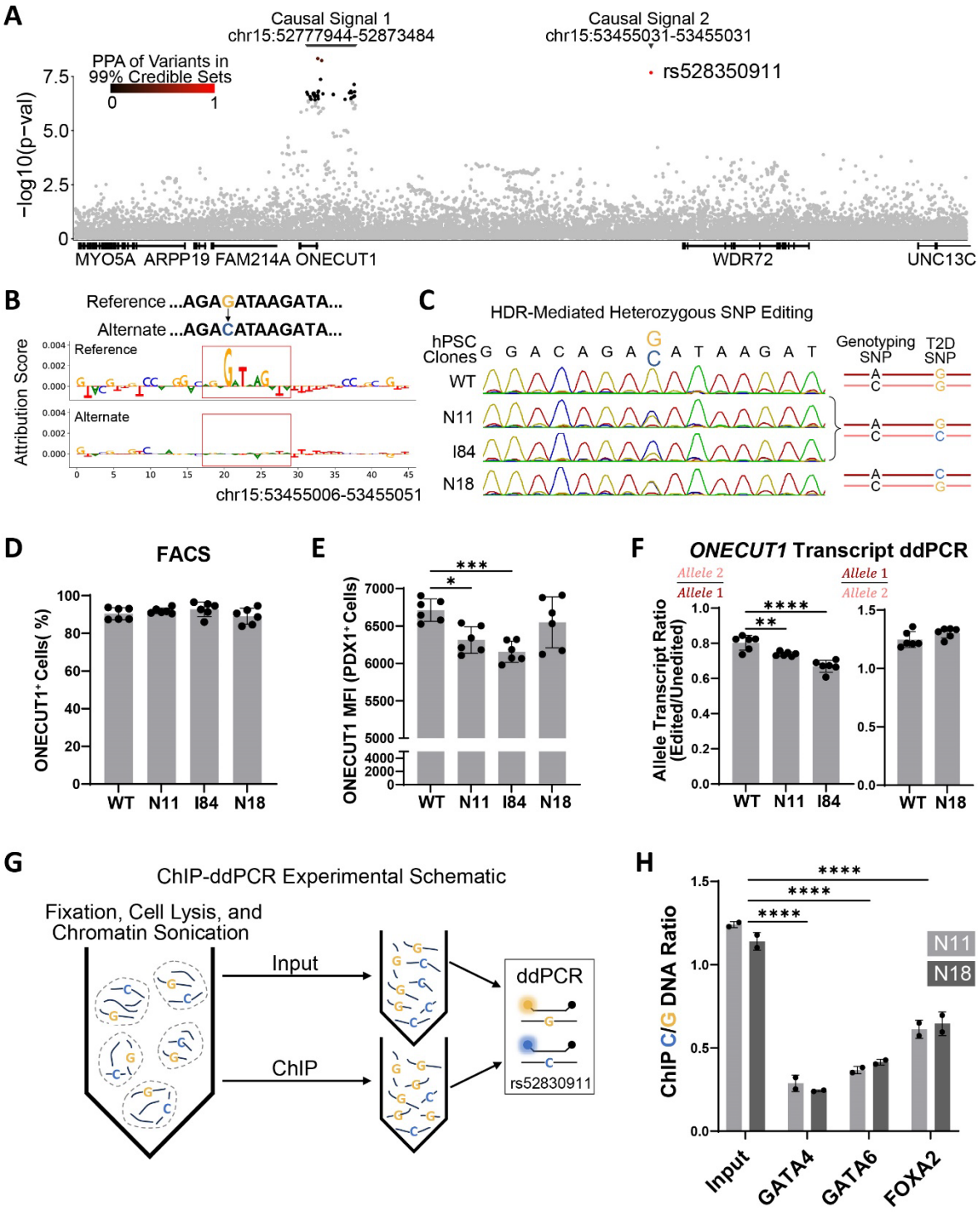
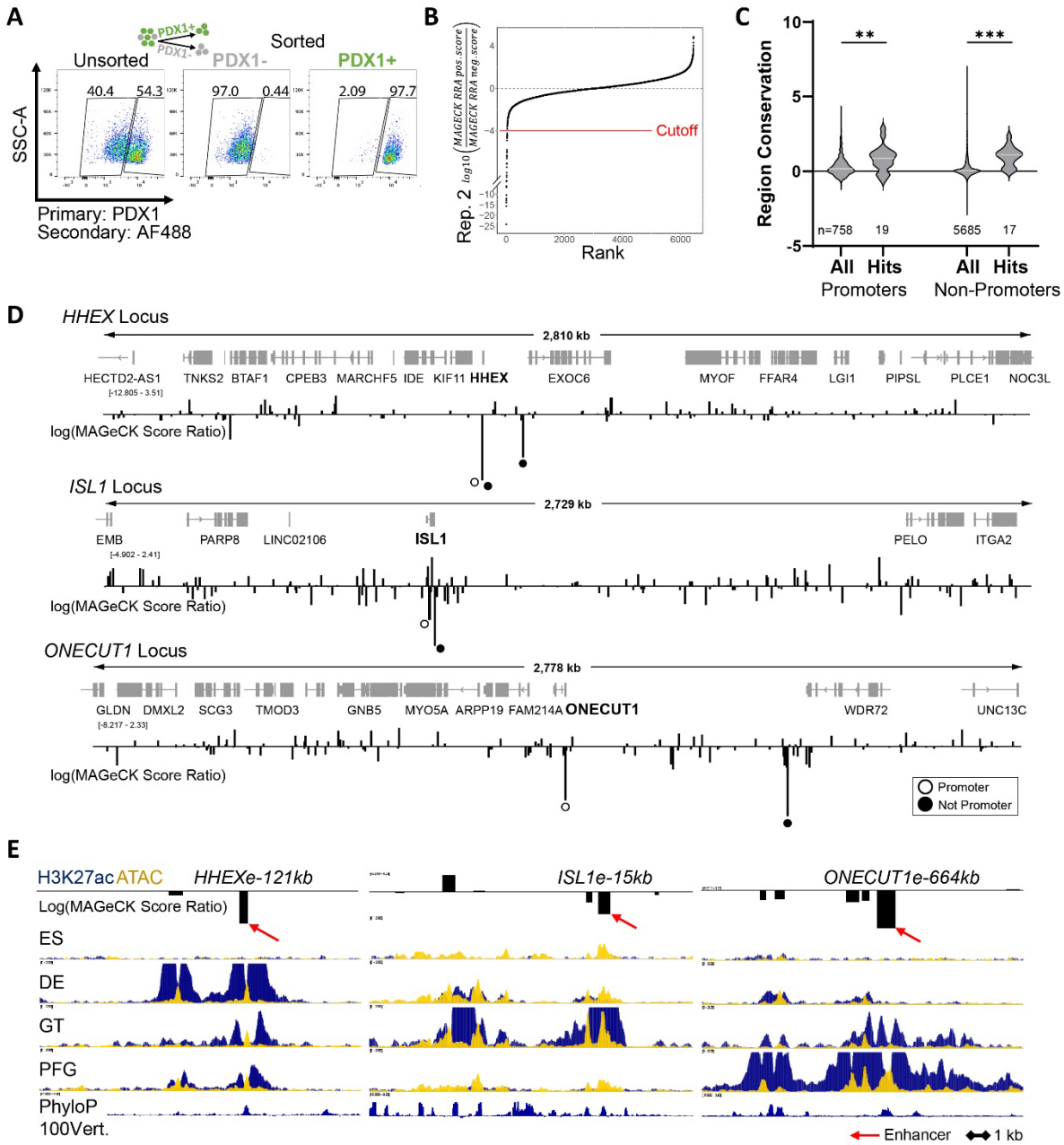

A
Causal Signal 1
chr15:52777944-52873484
Causal Signal 2
chr15:53455031-53455031
PPA of Variants in
99% Credible Sets
rs528350911
−log10(p−val)
MYO5A ARPP19 FAM214A ONECUT1
WDR72
UNC13C
B
Reference ...AGAGATAAGATA...
Alternate ...AGACATAAGATA...
Attribution Score
Reference
Alternate
chr15:53455006-53455051
C
HDR-Mediated Heterozygous SNP Editing
hPSC Clones
Genotyping SNP
T2D SNP
WT
N11
I84
N18
D
FACS
ONECUT1+ Cells(%)
WT N11 I84 N18
E
ONECUT1 MFI (PDX1+ Cells)
WT N11 I84 N18
F
ONECUT1 Transcript ddPCR
Allele 2
Allele 1
Allele Transcript Ratio (Edited/Unedited)
WT N11 I84
WT N18
G
ChIP-ddPCR Experimental Schematic
Fixation, Cell Lysis, and Chromatin Sonication
Input
ChIP
ddPCR
rs52830911
H
ChIP C/G DNA Ratio
N11
N18
Input
GATA4
GATA6
FOXA2

Fig. 4: T2D-associated SNP hPSC modeling and pancreatic differentiation characterization
A) LocusZoom plot of variants from Mahajan et al (2018b) T2D GWAS meta-analysis unadjusted for BMI. Variants within credible sets are colored by the PPA score.
B) Decreased DeepLift attribution score at GATA4 motif upon introduction of variant. Red box represents a GATA4/6 binding motif hit by FIMO ($p < 0.001$).
C) Spectrogram of disease-associated SNP amplicon sequencing and schematic demonstrating allele that contains the introduced SNP in hPSCs.
D) Percentage of ONECUT1+ cells at PFG stage. For panels D, E, and F, each symbol represents one independent experiment ($n = 6$ independent experiments) and data are presented as the mean ± s.d. One-way analysis of variance (ANOVA) followed by Dunnett multiple comparisons test versus WT control.
E) Statistical comparison of ONECUT1 MFI of PDX1+ cells.
F) ONECUT1 transcript ratio of edited allele divided by unedited allele. For each independent experiment shown, two averaged technical ddPCR replicates.
G) ChIP-ddPCR schematic.
H) ChIPed DNA ratio analysis. Each symbol represents one independent differentiation and ChIP experiment, two averaged technical ddPCR replicates. Two-way analysis of variance (ANOVA) followed by Tukey's multiple comparisons test versus WT control.

Supplemental Information

Supplementary Figures:

Sup. Fig. 1



Sup. Fig. 1, Related to Figure 1: Supplemental results and analysis for CRISPRi screens

A) Representative flow cytometry plot of PFG differentiated cells pre- and post-sort PDX1 sorting.

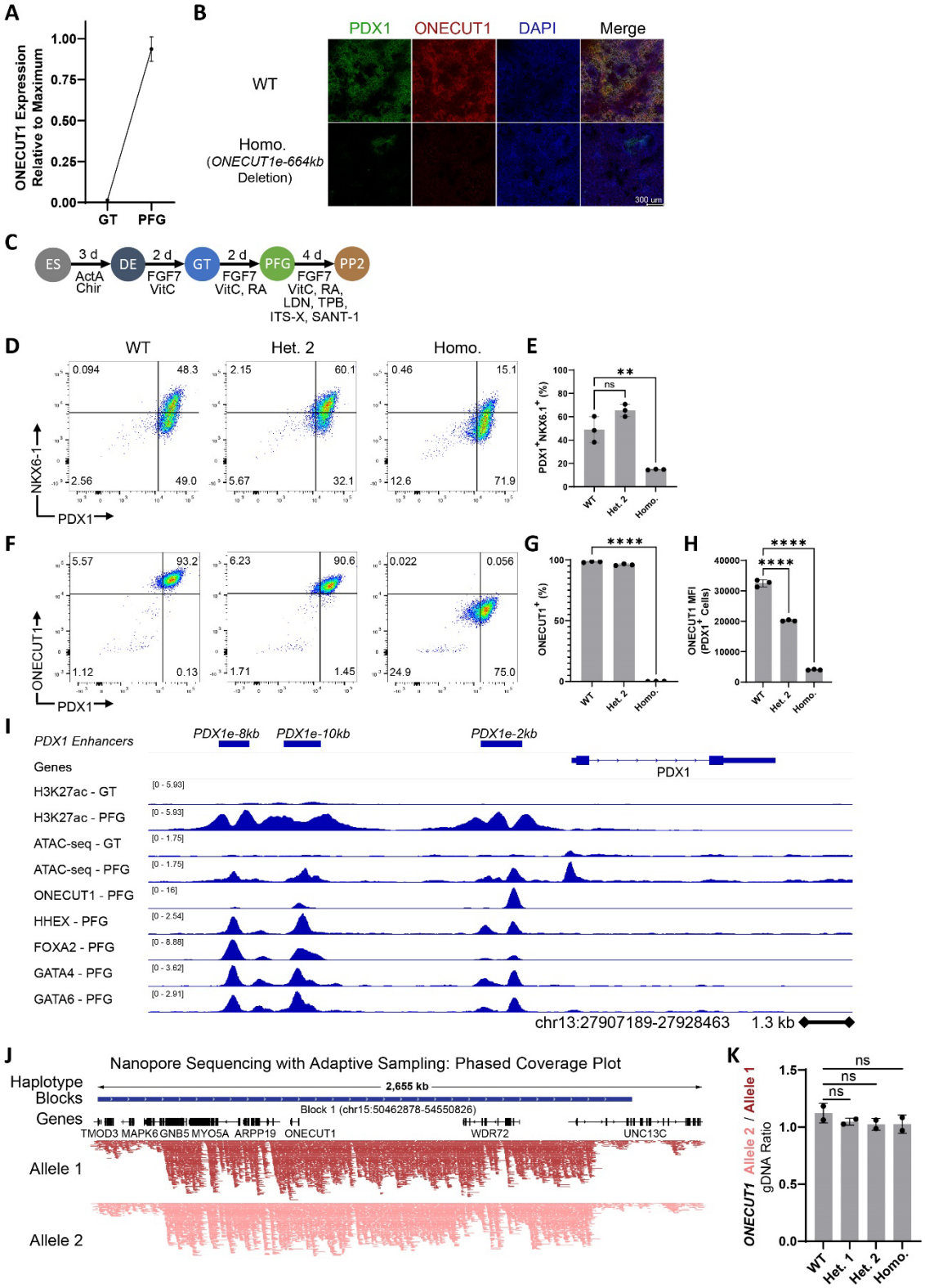
B) Rank plot of replicate 2 MAGeCK scores.

C) Calculated conservation scores for each region within indicated group. Ordinary one-way ANOVA with Šidák's multiple comparisons test, with a single pooled variance.

D) Representative IGV plots for enhancer-gene pair assignment rationale. Each bar represents a MAGeCK score ratio from replicate 2. A negative bar indicates region repression caused a decrease in the proportion of PDX1+ cells, with bar size indicating the magnitude of the change in PDX1 enrichment. Hollow black points represent promoter region hits, filled black points represent non-promoter region hits.

E) H3K27ac ChIP-seq, ATAC-seq, and phyloP100way vertebrate conservation at selected enhancer loci.

981 **Sup. Fig. 2**

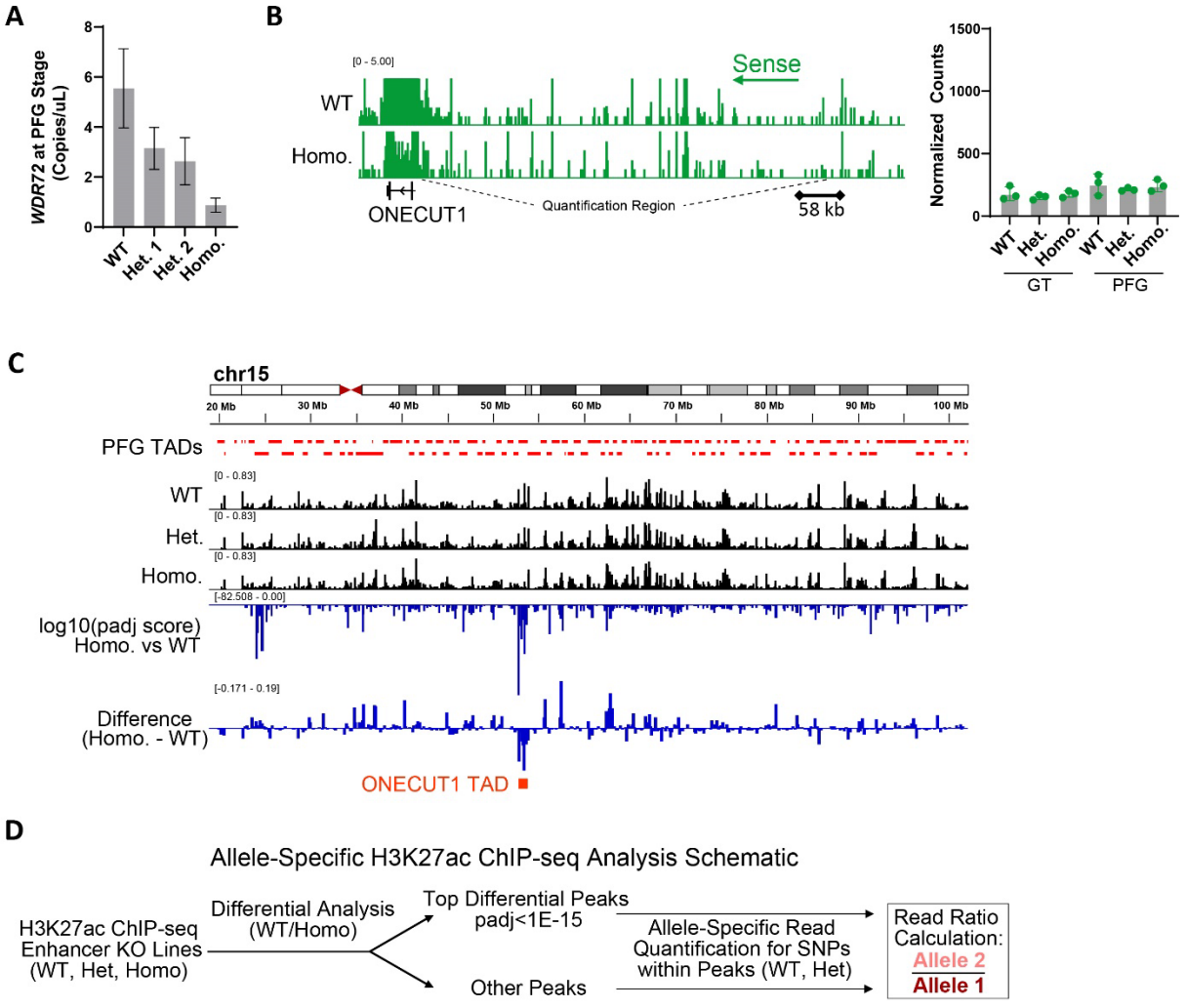


982

Sup. Fig. 2, Related to Figure 2: *ONECUT1e-664kb* deletion supplemental results and analysis

A) ONECUT1 transcript ddPCR on RNA from WT cells at GT and PFG stages.
B) PFG-stage immunofluorescence for PDX1 and ONECUT1 of WT and homozygous enhancer knockout cells.
C) ES to PP2 hPSC stepwise pancreatic differentiation protocol schematic.
D) Representative flow cytometry plots of WT and enhancer knockout cells differentiated to the PP2 stage.
E) Percentage of cells achieving a PDX1+NKX6.1+ identity. For panels E, G, and H, each dot represents one independent experiment (n = 3 independent experiments) and data are presented as the mean ± s.d. One-way analysis of variance (ANOVA) followed by Dunnett multiple comparisons test versus WT control.
F) Representative flow cytometry plots of WT and enhancer knockout cells differentiated to the PP2 stage.
G) Percentage of cells achieving a ONECUT1+ identity.
H) ONECUT1 MFI quantification from flow cytometry of unedited and enhancer knockout cells differentiated to the PP2 stage.
I) Transcription factor binding and epigenetic profile of the PDX1 locus in pancreatic differentiation. ONECUT1 and HHEX ChIP-seq from (Yang et al., 2022, GSE181480) and FOXA2, GATA4, and GATA6 ChIP-seq from (Lee et al., 2019, GSE114102).
J) Plot of phased reads and haplotype blocks from nanopore sequencing of H1 hPSCs with adaptive sampling.
K) ONECUT1 transcript ddPCR on DNA from WT and enhancer knockout cells. One-way analysis of variance (ANOVA) followed by Dunnett multiple comparisons test versus WT control.

1029 **Sup. Fig. 3**



1030

1031

1032 **Sup. Fig. 3, Related to Figure 3: Supplemental RNA-seq and H3K27ac ChIP-seq results**
1033 **and analysis**

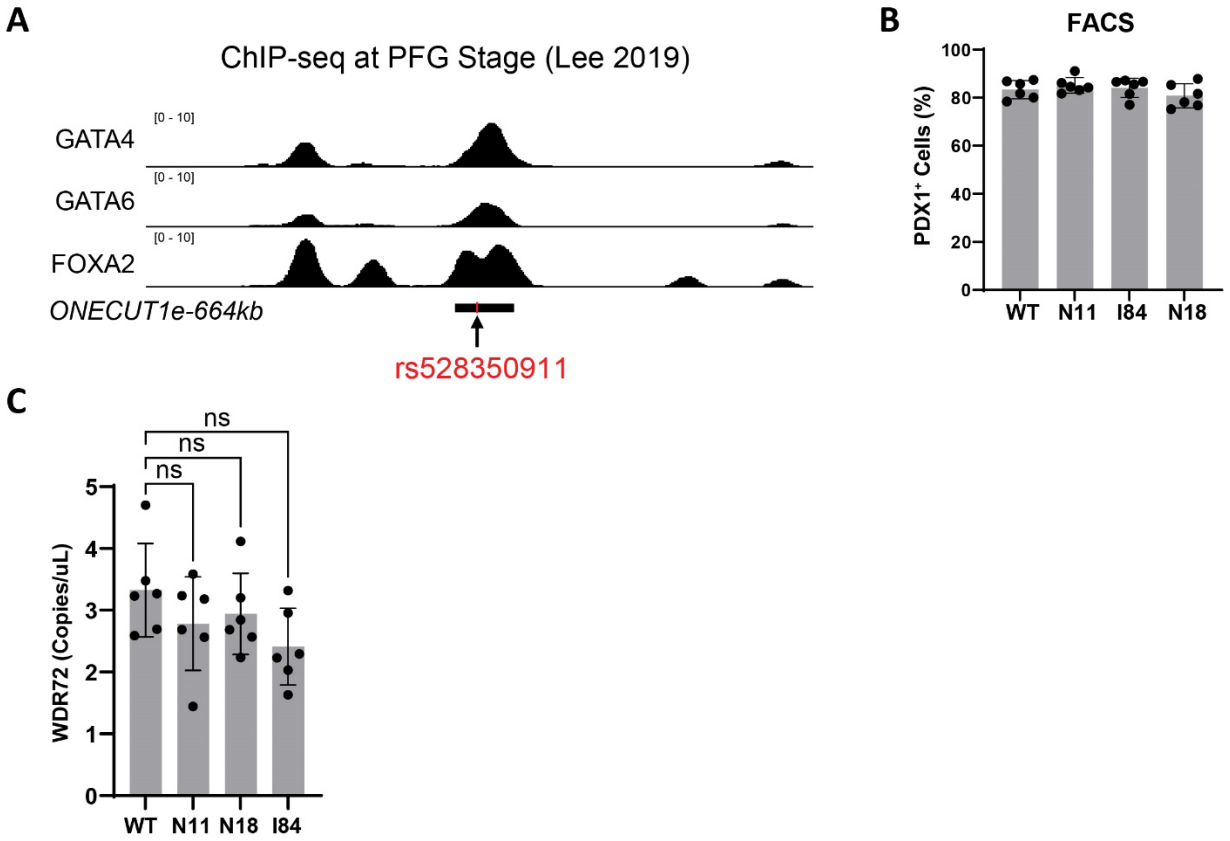
1034 A) Non allele-specific WDR72 ddPCR results of WT and enhancer knockout cells differentiated
1035 to the PFG stage.

1036 B) Visualization of sense read coverage (relative to ONECUT1) in region telomeric to
1037 ONECUT1 from strand-specific RNA-seq of PFG-stage WT and homozygous enhancer
1038 knockout cells (representative differentiation replicate). Each dot represents one independent
1039 experiment (n = 3 independent experiments) and data are presented as the mean ± s.d. One-way
1040 analysis of variance (ANOVA) followed by Dunnett multiple comparisons test versus WT
1041 control.

1042 C) Full chromosome view of H3K27ac in WT and enhancer knockout cells showing the
1043 ONECUT1 TAD to have the most significantly affected regions, and the regions with the
1044 greatest magnitude loss of H3K27ac.

1045 D) Allele-specific H3K27ac ChIP-seq analysis schematic.

1046 **Sup. Fig. 4**



1047

1048 **Sup. Fig. 4, Related to Figure 4: Supplemental results for T2D-associated SNP hPSC**
1049 **modeling**
1050 A) Visualization of ChIP-seq at ONECUT1e-664kb locus conducted at PFG stage. Black bar
1051 indicates ONECUT1e-664kb.
1052 B) Bar plots of percentage of cells achieving a PDX1+ identity. Each symbol represents one
1053 independent experiment (n = 6 independent experiments) and data are presented as the
1054 mean ± s.d. One-way analysis of variance (ANOVA) followed by Dunnett multiple comparisons
1055 test versus WT control.
1056 C) Non allele-specific WDR72 ddPCR results of WT and enhancer SNP cell lines differentiated
1057 to the PFG stage.
1058
1059

1060

1061

1062

1063

1064 **Supplementary Tables:**

1065 **Table S1:** Enhancer screen region selection and analysis results. Sheet 1: Transcription factor
1066 loci selected for CRISPRi interrogation. Sheet 2: ATAC-seq accessible regions selected for
1067 interrogation. Sheet 3: Read abundance quantification of each gRNA. Sheet 4: MAGeCK-RRA
1068 region-level analysis of two screens. Sheet 5: Top screen hits, FDR<0.1 in screen 2.

1069 **Table S2:** Engineered hPSC line allele genotypes.

1070 **Table S3:** List of gRNA target sequences and primer sequences for PCR genotyping.

1071 **Table S4:** Antibody target, company or lab that supplied antibody, catalog number, application,
1072 and quantity or dilution used are indicated. For ChIP experiments, the amount used for cells from
1073 around half of a 15cm plate (roughly 40-50 million cells) is indicated.

1074 **Table S5:** ddPCR assay details.

1075 **Table S6:** RNA-seq conducted on WT, heterozygous (het. 2) and homozygous *ONECUT1e-*
1076 *664kb* enhancer knockout cells. Sheet 1: DESeq2 normalized gene counts. Sheet 2: DESeq2
1077 differential analysis between WT and heterozygous enhancer knockout cells at PFG stage. Sheet
1078 3: DESeq2 differential analysis between WT and homozygous enhancer knockout cells at PFG
1079 stage.

1080 **Table S7:** H3K27ac ChIP-seq conducted on PFG-stage WT, heterozygous (het. 2), and
1081 homozygous *ONECUT1e-664kb* enhancer knockout cells. Sheet 1: Raw integer counts of
1082 H3K27ac reads. Sheet 2: DESeq2 calculated differential statistics between pairwise comparisons
1083 among WT, heterozygous, and homozygous enhancer knockout cells. Sheet 3: Phased read
1084 counts at indicated position, read ratios for each replicate, average read ratios for each genotype.
1085 Top_Differential_Peak indicates whether the position falls within a top differential peak (WT vs
1086 Homo, padj>1E-15).

1087 # References

1088 1. Katsanis, N. (2016). The continuum of causality in human genetic disorders. Genome Biology *17*.
1089 10.1186/s13059-016-1107-9.
1090 2. Lemelman, M.B., Letourneau, L., and Greeley, S.A.W. (2018). Neonatal Diabetes Mellitus: An
1091 Update on Diagnosis and Management. Clinics in Perinatology.
1092 3. Croucha, D.J.M., and Bodmer, W.F. (2020). Polygenic inheritance, GWAS, polygenic risk scores,
1093 and the search for functional variants. Proceedings of the National Academy of Sciences of the
1094 United States of America.
1095 4. Zhang, F., and Lupski, J.R. (2015). Non-coding genetic variants in human disease. Human
1096 Molecular Genetics.
1097 5. Edwards, S.L., Beesley, J., French, J.D., and Dunning, M. (2013). Beyond GWASs: Illuminating the
1098 dark road from association to function. American Journal of Human Genetics.
1099 6. Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill,
1100 R.E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing
1101 limb and fin and is associated with preaxial polydactyly. Human Molecular Genetics.
1102 10.1093/hmg/ddg180.
1103 7. Claringbould, A., and Zaugg, J.B. (2021). Enhancers in disease: molecular basis and emerging
1104 treatment strategies. Trends in Molecular Medicine.
1105 8. Luo, R., Yan, J., Oh, J.W., Xi, W., Shigaki, D., Wong, W., Cho, H.S., Murphy, D., Cutler, R., Rosen,
1106 B.P., et al. (2023). Dynamic network-guided CRISPRi screen identifies CTCF-loop-constrained
1107 nonlinear enhancer gene regulatory activity during cell state transitions. Nature Genetics *55*.
1108 10.1038/s41588-023-01450-7.
1109 9. Shukla, A., and Huangfu, D. (2018). Decoding the noncoding genome via large-scale CRISPR
1110 screens. Current Opinion in Genetics and Development.
1111 10. Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer–promoter contacts in gene
1112 expression control. Nature Reviews Genetics.
1113 11. Burgos, J.I., Vallier, L., and Rodríguez-Seguí, S.A. (2021). Monogenic Diabetes Modeling: In Vitro
1114 Pancreatic Differentiation From Human Pluripotent Stem Cells Gains Momentum. Frontiers in
1115 Endocrinology.
1116 12. Miguel-Escalada, I., Maestro, M.Á., Balboa, D., Elek, A., Bernal, A., Bernardo, E., Grau, V., García-
1117 Hurtado, J., Sebé-Pedrós, A., and Ferrer, J. (2022). Pancreas agenesis mutations disrupt a lead
1118 enhancer controlling a developmental enhancer cluster. Developmental Cell *57*.
1119 10.1016/j.devcel.2022.07.014.
1120 13. Weedon, M.N., Cebola, I., Patch, A.-M., Flanagan, S.E., Franco, E.D., Caswell, R., Rodríguez-Seguí,
1121 S.A., Shaw-Smith, C., Cho, C.H.H., Allen, H.L., et al. (2013). Recessive mutations in a distal PTF1A
1122 enhancer cause isolated pancreatic agenesis. Nature Genetics 2013 46:1 *46*, 61-64.
1123 10.1038/ng.2826.
1124 14. Franco, E.D., Flanagan, S.E., Houghton, J.A.L., Allen, H.L., MacKay, D.J.G., Temple, I.K., Ellard, S.,
1125 and Hattersley, A.T. (2015). The effect of early, comprehensive genomic testing on clinical care
1126 in neonatal diabetes: An international cohort study. The Lancet *386*. 10.1016/S0140-
1127 6736(15)60098-8.
1128 15. Gasperini, M., Tome, J.M., and Shendure, J. (2020). Towards a comprehensive catalogue of
1129 validated and target-linked human enhancers. Nature Reviews Genetics.
1130 16. Yan, J., and Huangfu, D. (2022). Epigenome rewiring in human pluripotent stem cells. Trends in
1131 Cell Biology.
1132 17. Yang, D., Cho, H., Tayyebi, Z., Shukla, A., Luo, R., Dixon, G., Ursu, V., Stransky, S., Tremmel, D.M.,
1133 Sackett, S.D., et al. (2022). CRISPR screening uncovers a central requirement for HHEX in

1134 pancreatic lineage commitment and plasticity restriction. Nature Cell Biology *24*.
1135 10.1038/s41556-022-00946-4.
1136 18. Li, Q.V., Dixon, G., Verma, N., Rosen, B.P., Gordillo, M., Luo, R., Xu, C., Wang, Q., Soh, C.L., Yang,
1137 D., et al. (2019). Genome-scale screens identify JNK–JUN signaling as a barrier for pluripotency
1138 exit and endoderm differentiation. Nature Genetics *51*. 10.1038/s41588-019-0408-9.
1139 19. Dixon, G., Pan, H., Yang, D., Rosen, B.P., Jashari, T., Verma, N., Pulecio, J., Caspi, I., Lee, K.,
1140 Stransky, S., et al. (2021). QSER1 protects DNA methylation valleys from de novo methylation.
1141 Science *372*. 10.1126/science.abd0875.
1142 20. Xu, X., Du, Y., Ma, L., Zhang, S., Shi, L., Chen, Z., Zhou, Z., Hui, Y., Liu, Y., Fang, Y., et al. (2021).
1143 Mapping germ-layer specification preventing genes in hPSCs via genome-scale CRISPR screening.
1144 iScience *24*. 10.1016/j.isci.2020.101926.
1145 21. Yilmaz, A., Braverman-Gross, C., Bialer-Tsypin, A., Peretz, M., and Benvenisty, N. (2020).
1146 Mapping Gene Circuits Essential for Germ Layer Differentiation via Loss-of-Function Screens in
1147 Haploid Human Embryonic Stem Cells. Cell Stem Cell *27*. 10.1016/j.stem.2020.06.023.
1148 22. Naxerova, K., Stefano, B.D., Makofske, J.L., Watson, E.V., de Kort, M.A., Martin, T.D., Dezfulian,
1149 M., Ricken, D., Wooten, E.C., Kuroda, M.I., et al. (2021). Integrated loss- And gain-of-function
1150 screens define a core network governing human embryonic stem cell behavior. Genes and
1151 Development *35*. 10.1101/GAD.349048.121.
1152 23. Armendariz, D.A., Goetsch, S.C., Sundarrajan, A., Sivakumar, S., Wang, Y., Xie, S., Munshi, N.V.,
1153 and Hon, G.C. (2023). CHD-associated enhancers shape human cardiomyocyte lineage
1154 commitment. eLife *12*. 10.7554/eLife.86206.
1155 24. Haswell, J.R., Mattioli, K., Gerhardinger, C., Maass, P.G., Foster, D.J., Peinado, P., Wang, X.,
1156 Medina, P.P., Rinn, J.L., and Slack, F.J. (2021). Genome-wide CRISPR interference screen
1157 identifies long non-coding RNA loci required for differentiation and pluripotency. PLoS ONE *16*.
1158 10.1371/journal.pone.0252848.
1159 25. Lee, K., Cho, H., Rickert, R.W., Li, Q.V., Pulecio, J., Leslie, C.S., and Huangfu, D. (2019). FOXA2 Is
1160 Required for Enhancer Priming during Pancreatic Differentiation. Cell Reports.
1161 10.1016/j.celrep.2019.06.034.
1162 26. Balboa, D., Iworima, D.G., and Kieffer, T.J. (2021). Human Pluripotent Stem Cells to Model Islet
1163 Defects in Diabetes. Frontiers in Endocrinology.
1164 27. Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J.,
1165 Steinthorsdottir, V., Scott, R.A., and Grarup, N. (2018). Fine-mapping type 2 diabetes loci to
1166 single-variant resolution using high-density imputation and islet-specific epigenome maps.
1167 Nature genetics *50*, 1505-1513.
1168 28. Philippi, A., Heller, S., Costa, I.G., Senée, V., Breunig, M., Li, Z., Kwon, G., Russell, R., Illing, A., Lin,
1169 Q., et al. (2021). Mutations and variants of ONECUT1 in diabetes. Nature Medicine *27*.
1170 10.1038/s41591-021-01502-7.
1171 29. Russ-Silsby, J., Patel, K.A., Laver, T.W., Hawkes, G., Johnson, M.B., Wakeling, M.N., Patil, P.P.,
1172 Hattersley, A.T., Flanagan, S.E., Weedon, M.N., and Franco, E.D. (2023). The Role of ONECUT1
1173 Variants in Monogenic and Type 2 Diabetes Mellitus. Diabetes *72*. 10.2337/db23-0498.
1174 30. Offield, M.F., Jetton, T.L., Labosky, P.A., Ray, M., Stein, R.W., Magnuson, M.A., Hogan, B.L.M.,
1175 and Wright, C.V.E. (1996). PDX-1 is required for pancreatic outgrowth and differentiation of the
1176 rostral duodenum. Development *122*. 10.1242/dev.122.3.983.
1177 31. Jonsson, J., Carlsson, L., Edlund, T., and Edlund, H. (1994). Insulin-promoter-factor 1 is required
1178 for pancreas development in mice. Nature *371*. 10.1038/371606a0.
1179 32. Zhu, Z., Li, Q.V., Lee, K., Rosen, B.P., González, F., Soh, C.L., and Huangfu, D. (2016). Genome
1180 Editing of Lineage Determinants in Human Pluripotent Stem Cells Reveals Mechanisms of
1181 Pancreatic Development and Diabetes. Cell Stem Cell. 10.1016/j.stem.2016.03.015.

1182 33. Wang, X., Sterr, M., Burtscher, I., Böttcher, A., Beckenbauer, J., Siehler, J., Meitinger, T., Häring,
1183 H.-U., Staiger, H., and Cernilogar, F.M. (2019). Point mutations in the PDX1 transactivation
1184 domain impair human β-cell development and function. Molecular metabolism *24*, 80-97.
1185 34. Stoffers, D.A., Zinkin, N.T., Stanojevic, V., Clarke, W.L., and Habener, J.F. (1997). Pancreatic
1186 agenesis attributable to a single nucleotide deletion in the human IPF1 gene coding sequence.
1187 Nature Genetics *15*. 10.1038/ng0197-106.
1188 35. Stoffers, D.A., Ferrer, J., Clarke, W.L., and Habener, J.F. (1997). Early-onset type-ll diabetes
1189 mellitus (MODY4) linked to IPF1. Nature genetics *17*, 138-139.
1190 36. Steinthorsdottir, V., Thorleifsson, G., Sulem, P., Helgason, H., Grarup, N., Sigurdsson, A.,
1191 Helgadottir, H.T., Johannsdottir, H., Magnusson, O.T., Gudjonsson, S.A., et al. (2014).
1192 Identification of low-frequency and rare sequence variants associated with elevated or reduced
1193 risk of type 2 diabetes. Nature Genetics *46*. 10.1038/ng.2882.
1194 37. Smith, S.B., Qu, H.-Q., Taleb, N., Kishimoto, N.Y., Scheel, D.W., Lu, Y., Patch, A.-M., Grabs, R.,
1195 Wang, J., and Lynn, F.C. (2010). Rfx6 directs islet formation and insulin production in mice and
1196 humans. Nature *463*, 775-780.
1197 38. Kropp, P.A., Zhu, X., and Gannon, M. (2019). Regulation of the Pancreatic Exocrine
1198 Differentiation Program and Morphogenesis by Onecut 1/Hnf6. CMGH *7*.
1199 10.1016/j.jcmgh.2019.02.004.
1200 39. Kim, S.K., Selleri, L., Lee, J.S., Zhang, A.Y., Gu, X., Jacobs, Y., and Cleary, M.L. (2002). Pbx1
1201 inactivation disrupts pancreas development and in Ipf1-deficient mice promotes diabetes
1202 mellitus. Nature Genetics *30*. 10.1038/ng860.
1203 40. Ahlgren, U., Pfaff, S.L., Jessell, T.M., Edlund, T., and Edlund, H. (1997). Independent requirement
1204 for ISL1 in formation of pancreatic mesenchyme and islet cells. Nature *385*. 10.1038/385257a0.
1205 41. Lee, C.S., Sund, N.J., Vatamaniuk, M.Z., Matschinsky, F.M., Stoffers, D.A., and Kaestner, K.H.
1206 (2002). Foxa2 controls Pdx1 gene expression in pancreatic β-cells in vivo. Diabetes *51*.
1207 10.2337/diabetes.51.8.2546.
1208 42. Barbera, J.P.M., Clements, M., Thomas, P., Rodriguez, T., Meloy, D., Kioussis, D., and
1209 Beddington, R.S.P. (2000). The homeobox gene Hex is required in definitive endodermal tissues
1210 for normal forebrain, liver and thyroid formation. Development *127*. 10.1242/dev.127.11.2433.
1211 43. Bort, R., Martinez-Barbera, J.P., Beddington, R.S.P., and Zaret, K.S. (2004). Hex homeobox gene-
1212 dependent tissue positioning is required for organogenesis of the ventral pancreas.
1213 Development *131*. 10.1242/dev.00965.
1214 44. Decker, K., Goldman, D.C., Grasch, C.L., and Sussel, L. (2006). Gata6 is an important regulator of
1215 mouse pancreas development. Developmental Biology *298*. 10.1016/j.ydbio.2006.06.046.
1216 45. Xuan, S., Borok, M.J., Decker, K.J., Battle, M.A., Duncan, S.A., Hale, M.A., Macdonald, R.J., and
1217 Sussel, L. (2012). Pancreas-specific deletion of mouse Gata4 and Gata6 causes pancreatic
1218 agenesis. Journal of Clinical Investigation *122*. 10.1172/JCI63352.
1219 46. Carrasco, M., Delgado, I., Soria, B., Martín, F., and Rojas, A. (2012). GATA4 and GATA6 control
1220 mouse pancreas organogenesis. Journal of Clinical Investigation *122*. 10.1172/JCI63240.
1221 47. Watt, A.J., Zhao, R., Li, J., and Duncan, S.A. (2007). Development of the mammalian liver and
1222 ventral pancreas is dependent on GATA4. BMC Developmental Biology *7*. 10.1186/1471-213X-7-
1223 37.
1224 48. Spence, J.R., Lange, A.W., Lin, S.C.J., Kaestner, K.H., Lowy, A.M., Kim, I., Whitsett, J.A., and Wells,
1225 J.M. (2009). Sox17 Regulates Organ Lineage Segregation of Ventral Foregut Progenitor Cells.
1226 Developmental Cell *17*. 10.1016/j.devcel.2009.05.012.
1227 49. Lorberbaum, D.S., Kishore, S., Rosselot, C., Sarbaugh, D., Brooks, E.P., Aragon, E., Xuan, S.,
1228 Simon, O., Ghosh, D., Mendelsohn, C., et al. (2020). Retinoic acid signaling within pancreatic

1229 endocrine progenitors regulates mouse and human β cell specification. Development
1230 (Cambridge) *147*. 10.1242/dev.189977.
1231 50. Haumaitre, C., Barbacci, E., Jenny, M., Ott, M.O., Gradwohl, G., and Cereghini, S. (2005). Lack of
1232 TCF2/vHNF1 in mice leads to pancreas agenesis. Proceedings of the National Academy of
1233 Sciences of the United States of America *102*. 10.1073/pnas.0405776102.
1234 51. Vanhorenbeeck, V., Jenny, M., Cornut, J.F., Gradwohl, G., Lemaigre, F.P., Rousseau, G.G., and
1235 Jacquemin, P. (2007). Role of the Onecut transcription factors in pancreas morphogenesis and in
1236 pancreatic and enteric endocrine differentiation. Developmental Biology *305*.
1237 10.1016/j.ydbio.2007.02.027.
1238 52. Nishimura, W., Kondo, T., Salameh, T., Khattabi, I.E., Dodge, R., Bonner-Weir, S., and Sharma, A.
1239 (2006). A switch from MafB to MafA expression accompanies differentiation to pancreatic β-
1240 cells. Developmental Biology *293*. 10.1016/j.ydbio.2006.02.028.
1241 53. Osipovich, A.B., Long, Q., Manduchi, E., Gangula, R., Hipkens, S.B., Schneider, J., Okubo, T.,
1242 Stoeckert, C.J., Takada, S., and Magnuson, M.A. (2014). Insm1 promotes endocrine cell
1243 differentiation by modulating the expression of a network of genes that includes Neurog3 and
1244 Ripply3. Development (Cambridge) *141*. 10.1242/dev.104810.
1245 54. Nissim, S., Weeks, O., Talbot, J.C., Hedgepeth, J.W., Wucherpfennig, J., Schatzman-Bone, S.,
1246 Swinburne, I., Cortes, M., Alexa, K., Megason, S., et al. (2016). Iterative use of nuclear receptor
1247 Nr5a2 regulates multiple stages of liver and pancreas development. Developmental Biology *418*.
1248 10.1016/j.ydbio.2016.07.019.
1249 55. Jacquemin, P., Lemaigre, F.P., and Rousseau, G.G. (2003). The Onecut transcription factor HNF-6
1250 (OC-1) is required for timely specification of the pancreas and acts upstream of Pdx-1 in the
1251 specification cascade. Developmental Biology *258*. 10.1016/S0012-1606(03)00115-5.
1252 56. PanelApp (2024). Genomics England PanelApp;https://panelapp.genomicsengland.co.uk (date
1253 accessed), Monogenic diabetes (Version 2.54).
1254 57. Horikoshi, M., Hara, K., Ito, C., Shojima, N., Nagai, R., Ueki, K., Froguel, P., and Kadowaki, T.
1255 (2007). Variations in the HHEX gene are associated with increased risk of type 2 diabetes in the
1256 Japanese population. Diabetologia *50*. 10.1007/s00125-007-0827-5.
1257 58. Shimomura, H., Sanke, T., Hanabusa, T., Tsunoda, K., Furuta, H., and Nanjo, K. (2000). Nonsense
1258 mutation of Islet-1 (Q310X) found in a type 2 diabetic patient with a strong family history.
1259 Diabetes *49*. 10.2337/diabetes.49.9.1597.
1260 59. Iacovazzo, D., Flanagan, S.E., Walker, E., Quezado, R., Barros, F.A.D.S., Caswell, R., Johnson, M.B.,
1261 Wakeling, M., Brändle, M., Guo, M., et al. (2018). MAFA missense mutation causes familial
1262 insulinomatosis and diabetes mellitus. Proceedings of the National Academy of Sciences of the
1263 United States of America *115*. 10.1073/pnas.1712262115.
1264 60. Ng, N.H.J., Jasmen, J.B., Lim, C.S., Lau, H.H., Krishnan, V.G., Kadiwala, J., Kulkarni, R.N., Ræder,
1265 H., Vallier, L., Hoon, S., and Teo, A.K.K. (2019). HNF4A Haploinsufficiency in MODY1 Abrogates
1266 Liver and Pancreas Differentiation from Patient-Derived Induced Pluripotent Stem Cells. iScience
1267 *16*. 10.1016/j.isci.2019.05.032.
1268 61. Haldorsen, I.S., Vesterhus, M., Ræder, H., Jensen, D.K., Søvik, O., Molven, A., and Njølstad, P.R.
1269 (2008). Lack of pancreatic body and tail in HNF1B mutation carriers. Diabetic Medicine *25*.
1270 10.1111/j.1464-5491.2008.02460.x.
1271 62. Shi, Z.D., Lee, K., Yang, D., Amin, S., Verma, N., Li, Q.V., Zhu, Z., Soh, C.L., Kumar, R., Evans, T., et
1272 al. (2017). Genome Editing in hPSCs Reveals GATA6 Haploinsufficiency and a Genetic Interaction
1273 with GATA4 in Human Pancreatic Development. Cell Stem Cell *20*. 10.1016/j.stem.2017.01.001.
1274 63. Trott, J., Alpagu, Y., Tan, E.K., Shboul, M., Dawood, Y., Elsy, M., Wollmann, H., Tano, V., Bonnard,
1275 C., Eng, S., et al. (2020). Mitchell-Riley syndrome iPSCs exhibit reduced pancreatic endoderm
1276 differentiation due to a mutation in RFX6. Development (Cambridge) *147*. 10.1242/dev.194878.

1277 64. Bolt, C.C., Lopez-Delisle, L., Hintermann, A., Mascrez, B., Rauseo, A., Andrey, G., and Duboule, D.
1278 (2022). Context-dependent enhancer function revealed by targeted inter-TAD relocation. Nature
1279 Communications *13*. 10.1038/s41467-022-31241-3.
1280 65. Morgens, D.W., Wainberg, M., Boyle, E.A., Ursu, O., Araya, C.L., Tsui, C.K., Haney, M.S., Hess,
1281 G.T., Han, K., Jeng, E.E., et al. (2017). Genome-scale measurement of off-target activity using
1282 Cas9 toxicity in high-throughput screens. Nature Communications *8*. 10.1038/ncomms15178.
1283 66. Sanson, K.R., Hanna, R.E., Hegde, M., Donovan, K.F., Strand, C., Sullender, M.E., Vaimberg, E.W.,
1284 Goodale, A., Root, D.E., Piccioni, F., and Doench, J.G. (2018). Optimized libraries for CRISPR-Cas9
1285 genetic screens with multiple modalities. Nature Communications *9*. 10.1038/s41467-018-
1286 07901-8.
1287 67. Pulecio, J., Tayyebi, Z., Liu, D., Wong, W., Luo, R., Damodaran, J.R., Kaplan, S., Cho, H., Yan, J.,
1288 Murphy, D., et al. (2023). Discovery of Competent Chromatin Regions in Human Embryonic Stem
1289 Cells. bioRxiv.
1290 68. Li, W., Xu, H., Xiao, T., Cong, L., Love, M.I., Zhang, F., Irizarry, R.A., Liu, J.S., Brown, M., and Liu,
1291 X.S. (2014). MAGeCK enables robust identification of essential genes from genome-scale
1292 CRISPR/Cas9 knockout screens. Genome biology *15*, 1-12.
1293 69. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral
1294 substitution rates on mammalian phylogenies. Genome Research. 10.1101/gr.097857.109.
1295 70. Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J.,
1296 Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from
1297 poised enhancers and predicts developmental state. Proceedings of the National Academy of
1298 Sciences of the United States of America. 10.1073/pnas.1016071107.
1299 71. Jacquemin, P., Durviaux, S.M., Jensen, J., Godfraind, C., Gradwohl, G., Guillemot, F., Madsen,
1300 O.D., Carmeliet, P., Dewerchin, M., Collen, D., et al. (2000). Transcription Factor Hepatocyte
1301 Nuclear Factor 6 Regulates Pancreatic Endocrine Cell Differentiation and Controls Expression of
1302 the Proendocrine Gene ngn3. Molecular and Cellular Biology *20*. 10.1128/mcb.20.12.4445-
1303 4454.2000.
1304 72. Pierreux, C.E., Poll, A.V., Kemp, C.R., Clotman, F., Maestro, M.A., Cordi, S., Ferrer, J., Leyns, L.,
1305 Rousseau, G.G., and Lemaigre, F.P. (2006). The Transcription Factor Hepatocyte Nuclear Factor-6
1306 Controls the Development of Pancreatic Ducts in the Mouse. Gastroenterology *130*.
1307 10.1053/j.gastro.2005.12.005.
1308 73. Zhang, H., Ables, E.T., Pope, C.F., Washington, M.K., Hipkens, S., Means, A.L., Path, G., Seufert, J.,
1309 Costa, R.H., Leiter, A.B., et al. (2009). Multiple, temporal-specific roles for HNF6 in pancreatic
1310 endocrine and ductal differentiation. Mechanisms of Development *126*.
1311 10.1016/j.mod.2009.09.006.
1312 74. Uyehara, C.M., and Apostolou, E. (2023). 3D enhancer-promoter interactions and multi-
1313 connected hubs: Organizational principles and functional roles. Cell Reports.
1314 75. Yang, F. (2022). Promoter antisense RNAs: beyond transcription by-products of active
1315 promoters. RNA biology *19*, 533-540.
1316 76. Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nature
1317 Structural and Molecular Biology.
1318 77. Mills, J.D., Kawahara, Y., and Janitz, M. (2013). Strand-Specific RNA-Seq Provides Greater
1319 Resolution of Transcriptome Profiling. Current Genomics *14*. 10.2174/1389202911314030003.
1320 78. Slieker, R.C., Donnelly, L.A., Fitipaldi, H., Bouland, G.A., Giordano, G.N., Åkerlund, M., Gerl, M.J.,
1321 Ahlqvist, E., Ali, A., Dragan, I., et al. (2021). Distinct Molecular Signatures of Clinical Clusters in
1322 People With Type 2 Diabetes: An IMI-RHAPSODY Study. Diabetes *70*. 10.2337/DB20-1281.
1323 79. Torres, J.M., Abdalla, M., Payne, A., Fernandez-Tajes, J., Thurner, M., Nylander, V., Gloyn, A.L.,
1324 Mahajan, A., and McCarthy, M.I. (2020). A Multi-omic Integrative Scheme Characterizes Tissues

1325 of Action at Loci Associated with Type 2 Diabetes. American Journal of Human Genetics *107*.
1326 10.1016/j.ajhg.2020.10.009.
1327 80. Zou, Y., Carbonetto, P., Wang, G., and Stephens, M. (2022). Fine-mapping from summary data
1328 with the "Sum of Single Effects" model. PLoS Genetics *18*, e1010299.
1329 81. Brennan, K.J., Weilert, M., Krueger, S., Pampari, A., yun Liu, H., Yang, A.W.H., Morrison, J.A.,
1330 Hughes, T.R., Rushlow, C.A., Kundaje, A., and Zeitlinger, J. (2023). Chromatin accessibility in the
1331 Drosophila embryo is determined by transcription factor pioneering and enhancer activation.
1332 Developmental Cell *58*. 10.1016/j.devcel.2023.07.007.
1333 82. Pampari, A., Shcherbina, A., Nair, S., Schreiber, J., Patel, A., Wang, A., Kundu, S., Shrikumar, A.,
1334 and Kundaje, A. (2023). Bias factorized, base-resolution deep learning models of chromatin
1335 accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory
1336 variants.
1337 83. Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through
1338 propagating activation differences.
1339 84. Shaw-Smith, C., Franco, E.D., Allen, H.L., Batlle, M., Flanagan, S.E., Borowiec, M., Taplin, C.E.,
1340 Velden, J.V.A.-V.D., Cruz-Rojo, J., Nanclares, G.P.D., et al. (2014). GATA4 mutations are a cause
1341 of neonatal and childhood-onset diabetes. Diabetes *63*. 10.2337/db14-0061.
1342 85. Allen, H.L., Flanagan, S.E., Shaw-Smith, C., De Franco, E., Akerman, I., Caswell, R., Consortium,
1343 I.P.A., Ferrer, J., Hattersley, A.T., and Ellard, S. (2012). GATA6 haploinsufficiency causes
1344 pancreatic agenesis in humans. Nature genetics *44*, 20-22.
1345 86. Yang, Y.P., Magnuson, M.A., Stein, R., and Wright, C.V.E. (2017). The mammal-specific Pdx1 area
1346 II enhancer has multiple essential functions in early endocrine cell specification and postnatal β-
1347 cell maturation. Development (Cambridge). 10.1242/dev.143123.
1348 87. Fujitani, Y., Fujitani, S., Boyer, D.F., Gannon, M., Kawaguchi, Y., Ray, M., Shiota, M., Stein, R.W.,
1349 Magnuson, M.A., and Wright, C.V.E. (2006). Targeted deletion of a cis-regulatory region reveals
1350 differential gene dosage requirements for Pdx1 in foregut organ differentiation and pancreas
1351 formation. Genes and Development. 10.1101/gad.1360106.
1352 88. Rojas, A., Schachterle, W., Xu, S.M., and Black, B.L. (2009). An endoderm-specific transcriptional
1353 enhancer from the mouse Gata4 gene requires GATA and homeodomain protein-binding sites
1354 for function in vivo. Developmental dynamics : an official publication of the American
1355 Association of Anatomists *238*. 10.1002/dvdy.22091.
1356 89. Kelly, M.R., Wisniewska, K., Regner, M.J., Lewis, M.W., Perreault, A.A., Davis, E.S., Phanstiel,
1357 D.H., Parker, J.S., and Franco, H.L. (2022). A multi-omic dissection of super-enhancer driven
1358 oncogenic gene expression programs in ovarian cancer. Nature Communications *13*.
1359 10.1038/s41467-022-31919-8.
1360 90. Long, H.K., Osterwalder, M., Welsh, I.C., Hansen, K., Davies, J.O.J., Liu, Y.E., Koska, M., Adams,
1361 A.T., Aho, R., Arora, N., et al. (2020). Loss of Extreme Long-Range Enhancers in Human Neural
1362 Crest Drives a Craniofacial Disorder. Cell Stem Cell *27*. 10.1016/j.stem.2020.09.001.
1363 91. Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., and Shiroishi, T. (2005). Elimination of a long-
1364 range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation
1365 of the mouse limb. Development *132*. 10.1242/dev.01613.
1366 92. Uslu, V.V., Petretich, M., Ruf, S., Langenfeld, K., Fonseca, N.A., Marioni, J.C., and Spitz, F. (2014).
1367 Long-range enhancers regulating Myc expression are required for normal facial morphogenesis.
1368 Nature Genetics *46*. 10.1038/ng.2971.
1369 93. Isoda, T., Moore, A.J., He, Z., Chandra, V., Aida, M., Denholtz, M., van Hamburg, J.P., Fisch, K.M.,
1370 Chang, A.N., Fahl, S.P., et al. (2017). Non-coding Transcription Instructs Chromatin Folding and
1371 Compartmentalization to Dictate Enhancer-Promoter Communication and T Cell Fate. Cell *171*.
1372 10.1016/j.cell.2017.09.001.

1373 94. Osterwalder, M., Barozzi, I., Tissiéres, V., Fukuda-Yuzawa, Y., Mannion, B.J., Afzal, S.Y., Lee, E.A.,
1374 Zhu, Y., Plajzer-Frick, I., Pickle, C.S., et al. (2018). Enhancer redundancy provides phenotypic
1375 robustness in mammalian development. Nature *554*. 10.1038/nature25461.
1376 95. Choi, H.I., Chai, J.C., Lee, Y.S., Jung, K.H., and Chai, Y.G. (2021). Targeting MYC-inducing
1377 enhancer-associated noncoding (MYC-IEANC) RNAs inhibits the proliferation of HCC cells.
1378 96. Luo, X., Liu, Y., Dang, D., Hu, T., Hou, Y., Meng, X., Zhang, F., Li, T., Wang, C., Li, M., et al. (2021).
1379 3D Genome of macaque fetal brain reveals evolutionary innovations during primate
1380 corticogenesis. Cell *184*. 10.1016/j.cell.2021.01.001.
1381 97. Gonen, N., Futtner, C.R., Wood, S., Garcia-Moreno, S.A., Salamone, I.M., Samson, S.C., Sekido,
1382 R., Poulat, F., Maatouk, D.M., and Lovell-Badge, R. (2018). Sex reversal following deletion of a
1383 single distal enhancer of Sox9. Science *360*. 10.1126/science.aas9408.
1384 98. Lascar, N., Brown, J., Pattison, H., Barnett, A.H., Bailey, C.J., and Bellary, S. (2018). Type 2
1385 diabetes in adolescents and young adults. The Lancet Diabetes and Endocrinology.
1386 99. Franks, P.W., Pearson, E., and Florez, J.C. (2013). Gene-environment and gene-treatment
1387 interactions in type 2 diabetes: Progress, pitfalls, and prospects. Diabetes Care.
1388 100. Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N.,
1389 Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-mediated modular RNA-
1390 guided regulation of transcription in eukaryotes. Cell *154*, 442-442. 10.1016/j.cell.2013.06.044.
1391 101. Beaudry, J.-B., Pierreux, C.E., Hayhurst, G.P., Plumb-Rudewiez, N., Weiss, M.C., Rousseau, G.G.,
1392 and Lemaigre, F.P. (2006). Threshold levels of hepatocyte nuclear factor 6 (HNF-6) acting in
1393 synergy with HNF-4 and PGC-1α are required for time-specific gene expression during liver
1394 development. Molecular and cellular biology *26*, 6037-6046.
1395 102. Rezania, A., Bruin, J.E., Arora, P., Rubin, A., Batushansky, I., Asadi, A., O'Dwyer, S., Quiskamp, N.,
1396 Mojibian, M., Albrecht, T., et al. (2014). Reversal of diabetes with insulin-producing cells derived
1397 in vitro from human pluripotent stem cells. Nature Biotechnology. 10.1038/nbt.3033.
1398 103. Gaspar, J.M. (2018). Genrich: detecting sites of genomic enrichment.
1399 104. Labun, K., Montague, T.G., Krause, M., Cleuren, Y.N.T., Tjeldnes, H., and Valen, E. (2019).
1400 CHOPCHOP v3: Expanding the CRISPR web toolbox beyond genome editing. Nucleic Acids
1401 Research *47*. 10.1093/nar/gkz365.
1402 105. Sanjana, N.E., Shalem, O., and Zhang, F. (2014). Improved vectors and genome-wide libraries for
1403 CRISPR screening. Nature Methods.
1404 106. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094-
1405 3100.
1406 107. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane,
1407 T., McCarthy, S.A., and Davies, R.M. (2021). Twelve years of SAMtools and BCFtools. GigaScience
1408 *10*. 10.1093/gigascience/giab008.
1409 108. Martin, M., Patterson, M., Garg, S., Fischer, S.O., Pisanti, N., Klau, G.W., Schöenhuth, A., and
1410 Marschall, T. (2016). WhatsHap: fast and accurate read-based phasing. bioRxiv.
1411 109. Edge, P., and Bansal, V. (2019). Longshot enables accurate variant calling in diploid genomes
1412 from single-molecule long read sequencing. Nature Communications *10*. 10.1038/s41467-019-
1413 12493-y.
1414 110. Liao, Y., Smyth, G.K., and Shi, W. (2014). FeatureCounts: An efficient general purpose program
1415 for assigning sequence reads to genomic features. Bioinformatics *30*.
1416 10.1093/bioinformatics/btt656.
1417 111. Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., and
1418 Barillot, E. (2015). HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. Genome
1419 Biology *16*. 10.1186/s13059-015-0831-x.

1420 112. Kramer, N.E., Davis, E.S., Wenger, C.D., Deoudes, E.M., Parker, S.M., Love, M.I., and Phanstiel,
1421 D.H. (2022). Plotgardener: Cultivating precise multi-panel figures in R. Bioinformatics *38*.
1422 10.1093/bioinformatics/btac057.
1423 113. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and
1424 dispersion for RNA-seq data with DESeq2. Genome Biology *15*. 10.1186/s13059-014-0550-8.
1425 114. Servant, N., Lajoie, B.R., Nora, E.P., Giorgetti, L., Chen, C.J., Heard, E., Dekker, J., and Barillot, E.
1426 (2012). HiTC: Exploration of high-throughput 'C' experiments. Bioinformatics *28*.
1427 10.1093/bioinformatics/bts521.
1428 115. Sahin, M., Wong, W., Zhan, Y., Deynze, K.V., Koche, R., and Leslie, C.S. (2021). HiC-DC+ enables
1429 systematic 3D interaction calls and differential analysis for Hi-C and HiChIP. Nature
1430 Communications *12*. 10.1038/s41467-021-23749-x.
1431 116. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient
1432 alignment of short DNA sequences to the human genome. Genome Biology *10*. 10.1186/gb-
1433 2009-10-3-r25.
1434 117. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers,
1435 R.M., Brown, M., and Li, W. (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology
1436 *9*, 1-9.
1437 118. Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-
1438 throughput experiments. Annals of Applied Statistics *5*. 10.1214/11-AOAS466.
1439 119. Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of
1440 Problematic Regions of the Genome. Scientific Reports *9*. 10.1038/s41598-019-45839-z.
1441 120. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and
1442 Mesirov, J.P. (2011). Integrative genomics viewer. Nature Biotechnology.
1443 121. Van der Auwera, G.A., and O'Connor, B.D. (2020). Genomics in the cloud: using Docker, GATK,
1444 and WDL in Terra (O'Reilly Media).
1445 122. Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: Scanning for occurrences of a given motif.
1446 Bioinformatics *27*. 10.1093/bioinformatics/btr064.
1447 123. Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi,
1448 H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic
1449 transcription factor sequence specificity. Cell *158*. 10.1016/j.cell.2014.08.009.

1450