

1 **High-throughput Screening and CRISPR-Cas9 Modeling of Causal Lipid-associated**
2 **Expression Quantitative Trait Locus Variants**

3

4 Avanthi Raghavan¹, Xiao Wang², Peter Rogov^{3,#a}, Li Wang³, Xiaolan Zhang³, Tarjei S.
5 Mikkelsen^{3,#b}, Kiran Musunuru^{2,3*}

6

7 ¹ Harvard Medical School, Boston, Massachusetts, United States of America

8

9 ² Cardiovascular Institute, Department of Medicine, and Department of Genetics, Perelman
10 School of Medicine at The University of Pennsylvania, Philadelphia, Pennsylvania, United States
11 of America

12

13 ³ Broad Institute, Cambridge, Massachusetts, United States of America

14

15 ^{#a} Current Address: Neon Therapeutics, Cambridge, Massachusetts, United States of America

16

17 ^{#b} Current Address: 10X Genomics, Pleasanton, California, United States of America

18

19 * Corresponding author

20 E-mail: kiranmusunuru@gmail.com (KM)

21

22 **Abstract**

23

24 Genome-wide association studies have identified a number of novel genetic loci linked to serum
25 cholesterol and triglyceride levels. The causal DNA variants at these loci and the mechanisms by
26 which they influence phenotype and disease risk remain largely unexplored. Expression
27 quantitative trait locus analyses of patient liver and fat biopsies indicate that many lipid-
28 associated variants influence gene expression in a *cis*-regulatory manner. However, linkage
29 disequilibrium among neighboring SNPs at a genome-wide association study-implicated locus
30 makes it challenging to pinpoint the actual variant underlying an association signal. We used a
31 methodological framework for causal variant discovery that involves high-throughput
32 identification of putative disease-causal loci through a functional reporter-based screen, the
33 massively parallel reporter assay, followed by validation of prioritized variants in genome-edited
34 human pluripotent stem cell models generated with CRISPR-Cas9. We complemented the stem
35 cell models with CRISPR interference experiments *in vitro* and in knock-in mice *in vivo*. We
36 provide validation for two high-priority SNPs, rs2277862 and rs10889356, being causal for lipid-
37 associated expression quantitative trait loci. We also highlight the challenges inherent in
38 modeling common genetic variation with these experimental approaches.

39

40 **Author Summary**

41

42 Genome-wide association studies have identified numerous loci linked to a variety of clinical
43 phenotypes. It remains a challenge to identify and validate the causal DNA variants in these loci.
44 We describe the use of a high-throughput technique called the massively parallel reporter assay

45 to analyze thousands of candidate causal DNA variants for their potential effects on gene
46 expression. We use a combination of genome editing in human pluripotent stem cells, “CRISPR
47 interference” experiments in other cultured human cell lines, and genetically modified mice to
48 analyze the two highest-priority candidate DNA variants to emerge from the massively parallel
49 reporter assay, and we confirm the relevance of the variants to nearby gene expression. These
50 findings highlight a methodological framework with which to identify and functionally validate
51 causal DNA variants.

52

53 **Introduction**

54

55 Genome-wide association studies (GWASs) have emerged as a powerful unbiased tool to
56 identify single nucleotide polymorphisms (SNPs) associated with incidence of a particular
57 phenotype or disease [1]. Interestingly, only a small fraction of GWAS lead variants lie within
58 coding sequence and thus directly implicate a causal gene at a locus. The vast majority of
59 implicated SNPs fall in noncoding sequence, including introns and gene deserts, suggesting they
60 may play a regulatory role in gene expression. Moreover, most of these SNPs are not themselves
61 causal but exist in linkage disequilibrium (LD) with the true functional variants. The causal gene
62 driving an association signal is often not immediately apparent, unless the locus harbors a gene
63 with a known connection to the phenotype of interest. Although reports of GWASs typically
64 label each associated SNP with the name of the nearest annotated gene or most plausible
65 biological candidate at that locus, experiments in biological models are often necessary to
66 identify the true causal gene at the locus [2–4].

67

68 Reassuringly, GWASs for lipid traits have identified variants in loci harboring genes that have
69 previously been implicated in Mendelian disorders of lipoprotein metabolism (i.e. *LDLR*,
70 *PCSK9*, *ABCA1*, etc). Moreover, GWASs have uncovered a plethora of loci with no prior
71 connection to lipid metabolism. In an extensive GWAS for blood lipids, the Global Lipids
72 Genetics Consortium conducted a meta-analysis of 46 prior lipid GWASs comprising >100,000
73 individuals of European descent, and identified 95 loci associated with total cholesterol (TC),
74 LDL-C, HDL-C and/or TG [2]. Of these loci, 36 had previously been reported by smaller-scale
75 lipid GWASs at genome-wide significance, while the other 59 were previously unpublished. In
76 principle, these novel loci may offer new insights into lipoprotein metabolism and promising
77 targets for therapeutic intervention.

78
79 Although such GWASs have identified a host of associated loci, the pace of functional validation
80 has lagged far behind. For the vast majority of GWAS loci, the causal DNA variants and genes
81 remain unexplored, largely due to the difficult and time-intensive nature of functional follow-up.
82 Even after fine mapping within a locus, as a result of LD tens to hundreds of variants can
83 demonstrate indistinguishably strong associations with the phenotype, suggesting that genetic
84 epidemiology alone is an insufficient means of causal variant discovery.

85
86 Because many disease-associated variants are believed to modulate gene expression, expression
87 quantitative trait locus (eQTL) studies may illuminate potential downstream targets of the causal
88 variant. An eQTL is a genomic region that influences gene expression either in *cis* or *trans*;
89 however, GWAS-implicated variants are predominantly believed to function in *cis*-acting
90 manner. Importantly, eQTL studies have identified differentially regulated transcripts hundreds

91 of kilobases away from the genotyped variant, implicating long-range chromatin looping
92 interactions in the mechanisms of some eQTLs [5,6]. These differentially regulated genes then
93 become candidates for experimental manipulation (for example, through overexpression and
94 knockout of the orthologous genes in murine models) to ascertain their relevance to the
95 phenotype of interest [2]. However, identifying the causal variants underlying eQTLs through
96 functional studies is more challenging due to the poor conservation of noncoding DNA across
97 species.

98

99 Further insight into eQTL mechanisms can be gleaned by integrating risk-associated variants
100 with annotated maps of regulatory elements, such as DNase I hypersensitivity sites, chromatin
101 immunoprecipitation sequencing (ChIP-seq) peaks, and histone modifications. Candidate SNPs
102 that fall in transcriptionally active regions can then be prioritized for functional investigation.
103 Experimental approaches such as reporter assays, electrophoretic mobility shift assays (EMSA),
104 and ChIP can be employed to investigate allele-specific regulatory activity at each variant, as
105 well as determinants of differential transcription factor binding and function [7–9]. However,
106 such efforts have been laborious, requiring a significant amount of dedicated effort to
107 functionally validate each candidate SNP.

108

109 Two recently emergent technologies make it feasible to interrogate risk-associated variants in
110 eQTL loci in a much higher throughput fashion. The massively parallel reporter assay (MPRA)
111 allows investigators to generate high-complexity pools of reporter constructs where each
112 regulatory element or variant of interest is linked to a synthetic reporter gene that carries an
113 identifying barcode [10,11]. The reporter construct pools are introduced into relevant populations

114 of cultured cells, and the relative transcriptional activities of the individual elements or variants
115 are measured by sequencing the transcribed reporter mRNAs and counting their specific
116 barcodes. This approach can be used to rapidly profile the regulatory activity of thousands of
117 variants linked to eQTLs for specific phenotypes of interest.

118
119 Advances in genome editing technologies—from first-generation zinc finger nucleases (ZFNs)
120 to, more recently, transcription activator-like effector nucleases (TALENs) and clustered
121 regularly interspaced short palindromic repeats (CRISPR)-CRISPR-associated 9 (Cas9)
122 systems—have opened up unprecedented avenues by which to rigorously assess the functional
123 impact of novel genetic variants [12]. All three of these genome-editing tools can be used to
124 introduce targeted alterations into mammalian cells and model organisms. However, CRISPR-
125 Cas9 offers an optimal combination of high targeting efficiency, ease of use, and scalability [13].
126 The most widely used CRISPR-Cas9 system employs the *Streptococcus pyogenes* Cas9 nuclease,
127 which complexes with a synthetic guide RNA (gRNA) encoding a site-specific 20-nt protospacer
128 sequence that hybridizes an N₂₀NGG target DNA sequence. Once Cas9 induces a double-strand
129 break three nucleotides upstream of the NGG sequence, or protospacer adjacent motif (PAM),
130 the cell employs error-prone non-homologous end joining (NHEJ) to repair the break, often
131 leading to the introduction of an insertion or deletion that may disrupt gene function. If a single-
132 strand oligonucleotide (ssODN) or double-strand DNA vector is introduced, the cell can utilize it
133 as a donor template for homology-directed repair (HDR), enabling knock-in of specific
134 mutations.

135

136 With a goal of better understanding the role of human genetic variation in influencing blood lipid

137 levels, we employed a combination of MPRA and CRISPR-Cas9 genome editing to high-
138 throughput screen, identify, and validate causal variants at eQTL loci that have been linked to
139 blood lipids in humans.

140

141 **Results**

142

143 **Massively parallel reporter assays define two lipid-associated sites with significant** 144 **transcriptional activity**

145

146 The Global Lipids Genetics Consortium interrogated lead SNPs at 95 loci for blood lipids against
147 transcript abundance of local genes in biopsy samples of human liver, subcutaneous fat, and
148 omental fat [2]. This analysis identified 57 total eQTLs, suggesting that many lipid-associated
149 causal SNPs influence gene expression in a *cis*-regulatory manner. These eQTLs may highlight
150 candidate lipid-modulating genes, possibly located up to hundreds of kilobases away from the
151 eQTL tag SNPs, that underlie the GWAS association signals at these loci. However, LD among
152 neighboring SNPs at any given eQTL locus makes it challenging to pinpoint the causal variant
153 underlying an association signal.

154

155 To address this issue, we performed an MPRA experiment to rapidly profile the regulatory
156 activity of 1,837 variants linked to the 16 eQTL lead SNPs identified for subcutaneous fat and/or
157 omental fat (**S1 Table**; see Supplementary Tables 9 and 10 in ref. 2 for more information on
158 eQTLs) and thus prioritize potentially causal candidate SNPs. We used a pool of reporter
159 constructs in which every plausible regulatory variant (i.e., all SNPs with $r^2 \geq 0.5$ relative to the

160 16 eQTL lead SNPs) was embedded within a 144-bp genomic “tile” in six versions—major or
161 minor allele in the center, towards the 5’ end (right-shifted), or towards the 3’ end (left-shifted).
162 Each fragment was coupled to a reporter gene with a unique barcode identifier in the 3’
163 untranslated region. To minimize barcode- and amplification-associated biases, each fragment
164 was coupled to ~22 distinct barcodes. The pool was transfected into cultured mouse 3T3-L1
165 adipocytes or pre-adipocytes, and the copy number of each barcode was determined by RNAseq
166 and normalized to the amount of corresponding reporter DNA plasmid that entered the cells [10].
167 From the MPRA data (**S2 Table**), we prioritized variants that displayed significant allele-specific
168 enhancer activity, as measured by reporter expression, in 3T3-L1 adipocytes. We selected the
169 two variants with the strongest evidence, rs2277862 and rs10889356, for further investigation
170 (**Fig 1A**).

171

172 **Fig 1. Results of massively parallel reporter assays.** (A) MPRA identified rs2277862 and
173 rs10889356 as the SNPs with highest allele-specific regulatory activity in mouse 3T3-L1
174 adipocytes. For this experiment, each candidate SNP was represented on a 144-bp tile that was
175 either centered, left-shifted, or right-shifted relative to the SNP, in order to increase the
176 probability of capturing the correct regulatory context for that SNP. For each tile, the individual
177 signals for the major and minor alleles are shown for a representative experiment (where signal
178 refers to the log of median barcode counts for the given tile divided by median barcode counts
179 for all tiles). A positive signal implies enhancer activity, while a negative signal implies
180 repressor activity. In the final two columns, a log-ratio for major allele signal over minor allele
181 signal is calculated, along with a *P*-value (by Mann-Whitney *U* test) for the null hypothesis that
182 the major and minor alleles generate equal signals. (B) MPRA-based single-hit saturation
183 mutagenesis experiment with the rs2277862 right-shifted tile, either with the major allele of the
184 SNP (top) or the minor allele (bottom). Red bars indicate a significant change from the original
185 tile’s activity (Mann-Whitney *U* test, 5% FDR); blue bars, not significant. (C) MPRA-based
186 single-hit saturation mutagenesis experiment with the rs10889356 left-shifted tile, either with the
187 major allele of the SNP (top) or the minor allele (bottom).
188

189 The Global Lipids Genetics Consortium found rs2277862 to be the lead SNP for total cholesterol
190 (TC) at chromosome 20q11 ($P = 4 \times 10^{-10}$), with the minor allele associated with a 1.19 mg/dL

191 decrease in TC [2]. This locus harbors several genes with no prior connection to lipid
192 metabolism, including *ERGIC3*, *CPNE1*, and *CEP250*. rs10889356 is tightly linked to the lead
193 SNP for triglycerides (TG) ($P = 9 \times 10^{-43}$), low-density-lipoprotein cholesterol (LDL-C), and TC
194 at chromosome 1p31, which harbors *DOCK7* and *ANGPTL3*, the latter a well-established
195 modulator of blood lipids [14]. We hypothesized that rs2277862 and rs10889356 are causal for
196 their respective eQTLs and that each lies within a transcriptional regulatory site that influences
197 local gene expression in an allele-specific manner.

198

199 To better define the transcriptional activity at the sites of each of the two SNPs, we performed
200 single-hit saturation mutagenesis experiments with MPRA in 3T3-L1 adipocytes using pools of
201 reporter constructs harboring 144-bp tiles around each SNP with mutagenesis at each position in
202 the tiles. This enabled us to generate “footprint” representations (**Figs 1B** and **1C**) for each site
203 [10]. For rs2277862, when the minor allele (T) was present, there was a signal indicative of a
204 factor binding directly to the site of the SNP (CAAATA[T]GGCGA) and another signal
205 indicative of a second factor binding upstream of the first factor (ACGAGGTCA), with no signal
206 observed downstream of the site of the SNP (**Fig 1B**). When the major allele (C) was present, all
207 signal disappeared, suggesting that this allele abolishes all of the binding interactions. For
208 rs10889356, when the major allele (G) was present, there was a signal indicative of a factor
209 binding directly to the site of the SNP (AACTTCCT[G]T), as well as several other signals
210 indicative of multiple other factors binding on either side of the first factor (**Fig 1C**). When the
211 minor allele (A) was present, almost all signal disappeared. Of note, there was one downstream
212 nucleotide that showed a very strong signal regardless of allelic variation at rs10889356,
213 suggesting it contributes to transcriptional regulation in a manner independent of the SNP. These

214 data suggest that, for both rs2277862 and rs10889356, complexes of factors that are in part
215 anchored at the site of the SNP are responsible for modulation of local gene expression.

216

217 **Genome editing of the rs2277862 site in human pluripotent stem cells**

218

219 The MPRA data from 3T3-L1 adipocytes indicate that the minor allele (T) of rs2277862 (minor
220 allele frequency in Europeans = 0.15) increases transcriptional activity relative to the major allele
221 (C) (**Fig 1**). This would suggest that the minor allele functions as an enhancer, the major allele
222 functions as a repressor, or both. The rs2277862 locus harbors a number of genes with no prior
223 connection to lipid metabolism, although only three—*CEP250*, *CPNE1* and *ERGIC3*—show
224 evidence of differential regulation in human liver, subcutaneous fat, and omental fat biopsies.

225 The biology and relevant sites of action of all three of these genes are poorly understood.

226 Interestingly, rs2277862 is located over 50 kb away from two of these eQTL genes, *CEP250* and
227 *CPNE1*, suggesting that it may function as a long-range enhancer or repressor (**Fig 2A**).

228

229 **Fig 2. CRISPR-Cas9 genome editing at the rs2277862 locus in hPSCs.** (A) Schematic of
230 human chromosome 20q11 locus showing the relative positions of rs2277862, *CEP250*, *CPNE1*,
231 and *ERGIC3*. (B) Heterozygous rs2277862 minor allele knock-in was generated on the HUES 8
232 background (homozygous major at rs2277862) at 0.15% frequency using a gRNA that cuts 3 bp
233 upstream from the SNP and an exogenous ssODN. Representative indels are also shown. The
234 guide RNA protospacer is underlined, the PAM is bolded, and the SNP position is indicated in
235 red. (C) Gene expression in undifferentiated HUES 8 cells ($N = 2$ wild-type clones and 1 knock-
236 in clone; 6 wells per clone), normalized to mean expression level in wild-type clones. (D) Gene
237 expression in differentiated HUES 8 adipocytes ($N = 2$ wild-type clones and 1 knock-in clone; 6
238 wells per clone) (E, F) Individual data points for each clone shown in (C) and (D). Data are
239 displayed as means and s.e.m. (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ relative to control, calculated by
240 two-tailed unpaired Student's t -test).

241

242 To define the contribution of each allele to gene expression, we sought to alter the putative
243 regulatory element encompassing rs2277862 in two hPSC lines with different genotypes at this
244 SNP: HUES 8 (C/C, major/major) and H7 (T/T, minor/minor). Although the precise regulatory
245 elements at this locus remain to be determined, the saturation mutagenesis MPRA data for the
246 site suggests that rs2277862 lies directly within such an element. We began by attempting to
247 knock in the minor allele of rs2277862 (T) onto the HUES 8 (C/C) background via HDR. Using
248 CRISPR-Cas9 with a single gRNA with a predicted cleavage site 3 bp upstream from the SNP
249 along with a 67-nt ssODN in HUES 8 cells, we obtained a single recombinant heterozygote at a
250 frequency of 0.15% (1 out of 672 clones screened) (**Fig 2B**).

251
252 Because rs2277862 has an eQTL in three developmentally distinct tissues—liver, subcutaneous
253 fat, and omental fat—we reasoned that it may function as a global, non-cell-type-restricted
254 regulator of gene expression and thus modulate transcription in undifferentiated hPSCs as well.
255 We analyzed gene expression and observed no difference in *CEP250*, *CPNE1*, or *ERGIC3*
256 expression between wild-type clones (C/C) and the recombinant heterozygote clone (C/T) ($n = 2$
257 wild-type clones and 1 knock-in clone; 6 wells per clone) (**Fig 2C**). We then differentiated the
258 three clones into adipocytes, the cell type in which the MPRA was originally performed. We
259 reasoned that differentiated adipocytes might have greater transcriptional activity at the
260 rs2277862 locus, perhaps by recruitment of additional transcriptional machinery to the site, and
261 thereby exhibit more pronounced gene expression differences. However, hPSC-derived knock-in
262 adipocytes ($n = 2$ wild-type clones and 1 knock-in clone; 6 wells per clone) demonstrated altered
263 expression of only one of the 20q11 genes, *CPNE1* (down 6.9%) (**Fig 2D**).

264

265 Analysis of the data revealed that the differentiation protocol had introduced a great deal of
266 stochastic variation in gene expression, not only among genetically identical clones but also
267 among independent wells of the same clone. This point is emphasized in **Figs 2E** and **2F**, where
268 we have plotted the individual data points for the clones analyzed in **Figs 2C** and **2D** (with each
269 point representing a different well of the indicated clone). hPSCs demonstrated minimal intra-
270 clonal well-to-well variability, as well as minimal inter-clonal variability, based on comparison
271 of the two wild-type clones. However, when the same clones were differentiated to adipocytes,
272 the two wild-type clones displayed fairly distinct “setpoints” in gene expression, presumably due
273 to variations in differentiation efficiency. Thus, this variability between genetically equivalent
274 clones within a group could be confounding the ability to discern true differences secondary to a
275 genetic modification.

276

277 In light of the inefficiency of the HDR knock-in at the rs2277862 locus, we adopted an
278 alternative approach. Reasoning that small deletions encompassing the site of the SNP should
279 have effects on gene expression, since transcription factor binding sites are typically 8-10
280 nucleotides long, we used CRISPR-Cas9 to generate numerous deletion mutants. One
281 disadvantage of using NHEJ to introduce indels at a genomic site with a single gRNA is that a
282 wide variety of insertions and deletions are generated, and different indels may have different
283 effects on transcriptional regulatory activity. We therefore utilized a different strategy with two
284 gRNAs with cleavage sites flanking the SNP, 38 bp apart. Through this multiplexing NHEJ
285 approach we efficiently generated small deletions encompassing the candidate SNP at a
286 frequency of 59% (168 out of 285 clones screened) (**Fig 3A**). The use of dual gRNAs facilitated
287 efficient generation of many hPSC clones harboring predictable, and often homozygous,

288 deletions. Due to the efficacy of the multiplexing strategy, we were easily able to generate a
289 large number of homozygous deletion (“knockout”) cell clones in both the HUES 8 (C/C) and
290 H7 (T/T) cell lines.

291

292 **Fig 3. Genetic deletion of rs2277862 site in hPSCs alters gene expression at the 20q11 locus.**
293 (A) Homozygous 38-bp deletions (“knockout”) encompassing rs2277862 were generated on the
294 HUES 8 (homozygous major) and H7 (homozygous minor) backgrounds using a dual gRNA
295 approach. A representative agarose gel of PCR amplicons is shown. The guide RNA
296 protospacers are underlined, the PAMs are bolded, and the SNP position is indicated in red. (B)
297 Gene expression in undifferentiated HUES 8 cells ($N = 10$ wild-type and 10 knockout clones, 3
298 wells per clone), normalized to mean expression level in wild-type clones. (C) Gene expression
299 in undifferentiated H7 cells ($N = 8$ wild-type and 6 knockout clones, 3 wells per clone). Data are
300 displayed as means and s.e.m. (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ relative to control, calculated by
301 two-tailed unpaired Student’s t -test).

302

303 By quantitative reverse transcriptase-polymerase chain reaction (qRT-PCR), we observed

304 statistically significant changes in the expression of 20q11 genes in both hPSC lines.

305 Homozygous disruption of the major allele in HUES 8 cells ($n = 10$ wild-type and 10 knockout

306 clones, 3 wells per clone) substantially decreased expression of *CEP250* (down 25.7%), *CPNE1*

307 (down 31.2%) and *ERGIC3* (down 20.0%) (**Fig 3B**). Homozygous disruption of the minor allele

308 in H7 cells ($n = 8$ wild-type and 6 knockout clones, 3 wells per clone) subtly decreased

309 expression of *CEP250* (down 8.8%) and *ERGIC3* (down 10.2%) (**Fig 3C**).

310

311 In combination, the data from genome editing of the rs2277862 site in hPSCs suggest that the

312 major allele of rs2277862 functions as a transcriptional enhancer in hPSCs, and the minor allele

313 may have minimal enhancer activity as well. Intriguingly, these data are at odds with the MPRA

314 data, which suggest that the minor allele increases transcriptional activity relative to the major

315 allele. Nonetheless, the MPRA and genome-editing data agree that rs2277862 is a causal variant
316 with respect to transcriptional regulation.

317

318 **Genome editing of the rs10889356 site in human pluripotent stem cells**

319

320 rs10889356 is tightly linked to rs2131925 ($r^2 = 0.90$), the lead SNP for TG, TC and LDL-C at
321 the 1p31 locus, and possession of the minor allele at rs2131925 (MAF = 0.32) is associated with
322 a 4.9 mg/dL decrease in TG [2]. rs10889356 is situated in the promoter of the *DOCK7* gene,
323 which encodes a guanine nucleotide exchange factor that has not previously been implicated in
324 lipid metabolism. *ANGPTL3*, the probable causal gene at this locus, lies within an intron of
325 *DOCK7* and encodes a liver-specific secreted protein that inhibits endothelial lipase and
326 lipoprotein lipase, thereby increasing circulating levels of TG and HDL-C (**Fig 4A**) [14].

327 Curiously, eQTL data for rs2131925 suggest that expression levels of *DOCK7* and *ANGPTL3* are
328 inversely related in human liver samples, implying that the causal variant variably upregulates or
329 downregulates transcription of different genes at this locus through an unknown mechanism [2].

330

331 **Fig 4. Genetic deletion of rs10889356 site in hPSCs alters gene expression at the 1p31 locus.**
332 (A) Schematic of human chromosome 1p31 locus showing the relative positions of rs10889356,
333 *DOCK7*, and *ANGPTL3*. (B) Homozygous 36- to 39-bp deletions (“knockout”) encompassing
334 rs10889356 were generated on the H7 (homozygous major) background using a dual gRNA
335 approach. The guide RNA protospacers are underlined, the PAMs are bolded, and the SNP
336 position is indicated in red. (C) *DOCK7* expression in undifferentiated H7 cells ($N = 12$ wild-
337 type and 8 knockout clones, 3 wells per clone), normalized to mean expression level in wild-type
338 clones. (D) Gene expression in differentiated H7 hepatocyte-like cells ($N = 4$ wild-type and 4
339 knockout clones, 6 wells per clone). Data are displayed as means and s.e.m. (* $P < 0.05$, ** $P < 0.01$,
340 *** $P < 0.001$ relative to control, calculated by two-tailed unpaired Student’s *t*-test).

341

342 The MPRA data indicate that the major allele (G) of rs10889356 has enhancer activity relative to
343 the minor allele (A). Using dual gRNAs, we efficiently generated homozygous deletion mutants

344 for rs10889356 in H7 cells, which are homozygous major (G/G) at this SNP. We observed some
345 heterogeneity in deletion size, presumably because one of the gRNAs did not always induce a
346 DSB exactly at the predicted location 3 bp upstream from the PAM (**Fig 4B**). For gene
347 expression studies, we utilized hPSC clones harboring a range of 36- to 39-bp homozygous
348 deletions as we were not able to obtain a sizeable number of deletion mutants of one particular
349 genotype. In undifferentiated hPSCs ($n = 12$ wild-type and 8 knockout clones, 3 wells per clone),
350 disruption of the major allele diminished expression of *DOCK7* (down 8.3%), suggesting that the
351 major allele indeed confers enhancer activity (**Fig 4C**). *ANGPTL3* expression was too low to be
352 reliably measured in undifferentiated hPSCs, consistent with its being a liver-specific gene.
353
354 We then differentiated a subset of clones ($n = 4$ wild-type and 4 knockout clones, 6 wells per
355 clone) to hepatocyte-like cells (HLCs). The HLCs were characterized by expression of the liver-
356 specific markers *ALB* and *SERPINA1*, and the average levels of both these transcripts were
357 roughly equivalent between wild-type and knockout clones. Expression of *DOCK7* was
358 decreased in knockout HLCs, while expression of the liver-specific gene *ANGPTL3* was
359 increased (**Fig 4D**). However, these differences did not achieve statistical significance due to
360 marked variation in differentiation efficiency, not only among clones but also among different
361 wells from the same clone. Indeed, while average *ALB* expression levels were equivalent
362 between groups, the dynamic range of *ALB* expression within each group exceeded an order of
363 magnitude (data not shown), highlighting the tremendous variability of the HLC differentiation
364 protocol, paralleling our findings with the adipocyte differentiation protocol used in studying the
365 rs2277862 heterozygous knock-in clone.
366

367 **CRISPR interference at rs2277862 and rs10889356 sites**

368

369 One possible complementary approach to validating a candidate causal SNP is to harness the
370 sequence-dependent targeting specificity of CRISPR-Cas9 to direct a catalytically dead Cas9
371 mutant (dCas9) to the SNP site. The rationale is that if a variant is truly causal and lies within a
372 transcriptional regulatory element, then the presence of the bulky dCas9 protein at the site could
373 sterically hinder recruitment of native transcription factors to the regulatory site and thus
374 interfere with transcriptional regulation, i.e., CRISPR interference (CRISPRi) [15].

375

376 We generated CRISPRi constructs that co-expressed dCas9 with enhanced green fluorescent
377 protein (EGFP) and expressed each of the three gRNAs shown in **Fig 5A**. We transiently
378 transfected HEK 293T cells, which are homozygous for the major allele (C/C) at rs2277862,
379 with the dCas9 and gRNA constructs, either singly or in combination. With at least two of the
380 gRNAs, expression of *CEP250* and *CPNE1* were diminished relative to control cells, which
381 received the dCas9 construct without an accompanying gRNA (**Fig 5B**), suggesting that
382 dCas9/gRNA complexes had obstructed binding or function of a transcriptional enhancer at this
383 site. This result is directionally consistent with the data from the isogenic HUES 8 rs2277862
384 deletion mutants, which also indicated that the major allele (C) has enhancer activity, though at
385 odds with the MPRA data that suggested the minor allele (T) has enhancer activity.

386

387 **Fig 5. CRISPR interference modulates gene expression at the 20q11 and 1p31 loci.** (A)
388 Guide RNAs used at the rs2277862 site. The guide RNA protospacers are in colors, the PAMs
389 are underlined, and the SNP position is indicated in bold. (B) Gene expression, normalized to
390 mean expression level in control cells, in HEK 293T cells (homozygous major at rs2277862)
391 transfected with catalytically dead Cas9 (dCas9) with various gRNAs targeting the rs2277862
392 site (either singly or in combination, 3 wells per condition). Control cells were transfected with

393 dCas9 without an accompanying gRNA. (C) Guide RNAs used at the rs10889356 site. (D) Gene
394 expression in HepG2 hepatoma cells (homozygous major at rs10889356) transfected with dCas9
395 with various gRNAs targeting the rs10889356 site (3 wells per condition). Data are displayed as
396 means and s.e.m. (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ relative to control, calculated by two-tailed
397 unpaired Student's t -test).
398

399 We validated rs10889356 as a causal variant with CRISPRi constructs targeting the rs10889356
400 site in HepG2 cultured hepatoma cells, which are homozygous major (G/G) at this SNP (**Fig**
401 **5C**). We chose HepG2 cells so that we could assess not just *DOCK7* but also *ANGPTL3*, the
402 latter being liver-specific. In light of the low transfection efficiency of HepG2 cells, positive
403 transfectants were isolated by fluorescence-activated cell sorting (FACS) prior to gene
404 expression analysis. With three different gRNAs, singly or in combination, expression of
405 *DOCK7* was significantly decreased and expression of *ANGPTL3* was significantly increased
406 relative to control cells, which received the dCas9 construct without an accompanying gRNA
407 (**Fig 5D**). These results are directionally consistent with the data from the isogenic H7 HLC
408 rs10889356 deletion mutant experiment, which was also performed on a homozygous major
409 background. Additionally, both experiments recapitulated the inverse relationship between
410 *DOCK7* and *ANGPTL3* expression levels revealed by the human eQTL data for the locus lead
411 SNP rs2131925 [2].
412

413 **Interrogation of an rs2277862 knock-in mouse**

414
415 Because independent hPSC clones displayed variable propensities for directed differentiation,
416 thereby introducing confounding variability in gene expression studies, we sought an alternative
417 approach to faithfully model the effect of regulatory variation in primary tissues of interest,
418 namely liver and fat. The noncoding region encompassing rs2277862 is well conserved in mouse

419 (Fig 6A). Remarkably, the orthologous nucleotide in mouse also displays naturally occurring
420 variation and has been previously cataloged as rs27324996 on chromosome 2, with the same two
421 alleles (C and T) as in humans. As documented in dbSNP, rs27324996 has a MAF of 43% based
422 on genotyping analyses of 14 different inbred strains of mice. All three human eQTL genes have
423 a murine ortholog at this locus, and the orientation of these genes relative to the putative
424 regulatory variant is conserved between mouse and human (Fig 6A).

425

426 **Fig 6. Gene expression in an rs2277862 knock-in mouse.** (A) Schematic of mouse
427 chromosome 2qH1 locus showing the relative positions of rs27324996, *Cep250*, *Cpne1*, and
428 *Ergic3*. The architecture closely matches the orthologous human chromosome 20q11 locus. (B)
429 The noncoding region encompassing rs2277862 is well conserved in mouse, including allelic
430 variants of the SNP itself, with the murine equivalent being rs27324996. The SNP position is
431 indicated in blue bold, non-conserved nucleotides are indicated in red, and the guide RNA
432 protospacer used to generate the knock-in mouse is boxed. The electropherogram is from a
433 mouse in which the minor allele of rs2277862/rs27324996 (T) has been knocked into one
434 chromosome, along with the four non-conserved nucleotides that “humanize” the site. (C) Gene
435 expression in liver and fat tissues from littermates of the C57BL/6J background ($N = 18$ wild-
436 type mice and 10 homozygous knock-in mice), normalized to mean expression level in wild-type
437 mice. Data are displayed as means and s.e.m. (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ relative to
438 control, calculated by two-tailed unpaired Student’s t -test).

439

440 Because MPRA identified a transcriptional role for rs2277862 when performed in 3T3-L1
441 adipocytes, which are of murine origin, we reasoned that the regulatory machinery at this site
442 may be present in mouse tissues *in vivo*. Since HDR is far more efficient in mouse embryos than
443 in hPSCs [16], we used CRISPR-Cas9 to generate a knock-in mouse with a minor allele on the
444 C57BL/6J background, which is homozygous major (C/C) at rs27324996. Compared to genome
445 editing of hPSCs, one clear advantage of a mouse model is that even if only a single positive
446 founder mouse is obtained with genome editing, the knock-in allele can be bred to homozygosity
447 in a matter of months, and a large number of wild-type and knock-in mice can be generated for
448 well-powered gene expression studies in liver, subcutaneous fat, and omental fat. In principle,

449 this study design could enable replication of human eQTL association data while avoiding the
450 disadvantages of cultured human transformed cell lines (such as HepG2) and of suboptimal and
451 variable hPSC differentiation.

452

453 One-cell mouse embryos were injected with Cas9 mRNA, a gRNA targeting the site of
454 rs27324996, and a ssDNA donor template carrying the minor SNP allele as well as four
455 additional nucleotide changes that served to both “humanize” the sequence (i.e., make a perfect
456 match to the orthologous human sequence) and prevent re-cleavage of the knock-in allele by
457 CRISPR-Cas9. Out of 37 founder mice, there was one mouse bearing the knock-in allele (**Fig**
458 **6B**). We bred this mouse through two generations to obtain homozygous minor allele knock-in
459 mice.

460

461 We compared *Cep250*, *Cpne1*, and *Ergic3* gene expression in primary liver, subcutaneous fat,
462 and omental fat from wild-type (C/C) and homozygous knock-in (T/T) C57BL/6J littermates (n
463 = 18 wild-type and 10 knock-in mice). In liver, there was significantly decreased expression of
464 *Cep250* (down 28.2%), *Cpne1* (down 37.2%) and *Ergic3* (down 34.1%) in the homozygous
465 knock-in mice (**Fig 6C**). Notably, these changes are quite concordant with the directionality and
466 magnitudes of the changes observed between wild-type (C/C) and homozygous knockout hPSCs
467 (compare with **Fig 3B**). In fat tissues from knock-in mice, no statistically significant differences
468 were apparent, due to a large degree of variance in the gene expression measurements (**Fig 6C**).
469 Although no conclusions can be drawn from the fat data, the liver data from the knock-in mice in
470 combination with the genome-edited hPSC data and the CRISPRi data suggest that the site of
471 rs2277862/rs27324996 has transcriptional regulatory activity in both human and mouse.

472

473 **Discussion**

474

475 One of the principal challenges of the post-GWAS era has been cataloging the allelic spectrum
476 of causal variants underlying complex trait susceptibility. In this work, we describe a
477 methodological framework for causal variant discovery that involves high-throughput
478 identification of putative disease-causal loci through a functional reporter-based screen, MPRA,
479 followed by validation of prioritized variants in genome-edited cellular models. As a proof-of-
480 concept, we focused our experimental efforts on validating two top-ranked MPRA variants that
481 are potentially causal for lipid phenotypes.

482

483 We sought to rigorously demonstrate causality for rs2277862 and rs10889356 through the
484 generation of isogenic wild-type and mutant hPSCs. Consistent with prior studies, we found the
485 generation of knock-in clones via HDR to be very inefficient compared to the generation of
486 clones with defined deletions via multiplexed NHEJ. The difference in efficiency proved to be
487 crucial to the relative success of the alternative study designs. We were only able to generate a
488 single rs2277862 heterozygous knock-in clone despite screening hundreds of clones. We found
489 that clone-to-clone variability was sufficiently high to swamp out the expected small differences
490 in gene expression—on the order of 10% to 20% differences, based on eQTLs from human tissue
491 studies—and, accordingly, little informative data was obtained from the knock-in hPSC clone.
492 The situation was even worse when using cells differentiated from the hPSCs, given the
493 increased clone-to-clone variability that resulted.

494

495 It was feasible to obtain more precise data only when numerous wild-type and mutant clones
496 were used in the experiments (i.e., large n), which in turn was only possible when using NHEJ to
497 generate knockout clones with high efficiency. The knockout study design proved fruitful for
498 both rs2277862 and rs10889356, with knockout clones displaying statistically significant, albeit
499 small, differences from wild-type clones. Even with the use of a multitude of clones, however,
500 differentiation of the hPSCs greatly increased the clone-to-clone variability and obscured any
501 differences between wild-type and knockout clones.

502
503 This is not to say that hPSCs are a poor platform for functional genetic studies. hPSCs offer
504 many attractive advantages over existing model systems, including genomic stability,
505 pluripotency, and renewability. However, it seems prudent to consider the estimated “signal”
506 from a putative disease-causal variant relative to the estimated “noise” from a directed
507 differentiation protocol prior to pursuing a disease-modeling experiment in hPSCs. For highly
508 penetrant variants with large effect sizes, hPSCs have been shown to yield informative insights
509 into human genetics. For example, Ding et al. were able to characterize the rare gain-of-function
510 variant E17K in the gene *AKT2*, which is associated with hypoglycemia, hypoinsulinemia, and
511 increased body fat secondary to dysregulated insulin signaling [17]. hPSC-derived *AKT2* E17K
512 knock-in HLCs displayed significantly decreased glucose production, while hPSC-derived
513 adipocytes had increased TG content and increased glucose uptake compared to matched wild-
514 type controls. Through disease modeling in hPSCs, Ding et al. unequivocally established a
515 dominant activating role for the *AKT2* E17K mutation. However, GWAS-implicated common
516 variants would be expected to display much smaller effect sizes, suggesting that hPSC-derived
517 cell types may not be the ideal model system in which to investigate common genetic variation,

518 despite the touted benefits of isogenic disease modeling.

519

520 We explored two alternative model systems. The first involved CRISPR interference, a term that
521 is typically used to describe the repression of transcription via the positioning of dCas9 to a gene
522 promoter or coding sequence to block transcriptional initiation or elongation via steric

523 interference [15]; alternatively, dCas9 can be attached to a repressor domain such as Krüppel

524 associated box (KRAB) to effect transcriptional silencing via epigenetic chromatin modification

525 [18]. We opted to use the unadorned dCas9 protein rather than attaching an extra domain, since

526 the regulatory elements in the vicinity of the two SNPs are undefined and it was unclear whether

527 each SNP allele acted as an enhancer or repressor or was neutral. Instead, we relied on steric

528 interference to reverse any effect of the SNP allele. For both rs2277862 and rs10889356,

529 CRISPRi yielded results that were concordant with the effects observed in knockout hPSCs and

530 differentiated cells—decreased gene expression vis-à-vis the major allele of rs2277862, and

531 decreased *DOCK7* and increased *ANGPTL3* expression vis-à-vis the major allele of rs10889356.

532

533 Of note, we performed the CRISPRi experiments in cultured human transformed cells rather than

534 hPSCs, making the experiments far easier to perform and, in the case of HepG2 cells, allowing

535 us to use cells with inherent hepatocyte properties without having to undertake the differentiation

536 of hPSCs into HLCs. Indeed, our findings suggest that CRISPRi may be better suited to the high-

537 throughput interrogation of candidate causal variants (whether nominated by MPRA or another

538 technique) than genome editing of cells, although the latter should be regarded as the gold

539 standard, especially when it entails the knock-in of a specific allelic variant. Furthermore, a

540 limitation of CRISPRi is that it relies on the major allele of a SNP having enhancer or repressor

541 activity (rather than being neutral), since it may not always be possible to identify commonly
542 used human cultured transformed cell lines with minor alleles.
543
544 The second alternative model system we attempted to use was a knock-in mouse. This is not a
545 generalizable strategy, as it depends on the human SNP sequence and the orthologous sequence
546 in the mouse genome being very closely matched, which is not often the case with noncoding
547 regions. (One possible means to overcome this limitation would be the introduction of the
548 entirety of the orthologous human locus into the mouse genome, via bacterial artificial
549 chromosome transgenesis or another approach.) Fortuitously, the human rs2277862 sequence
550 matches closely with mouse, to the extent that the mouse genome has a perfectly corresponding
551 SNP, rs27324996. This allowed us to generate a “humanized” minor allele homozygous mouse
552 to compare with the wild-type major allele homozygous mouse. The theoretical advantage of this
553 study design is that it allows for the analysis of authentic primary tissues such as liver and fat
554 without the need for directed differentiation *in vitro* or reliance on transformed cell lines. In
555 practice, while we observed statistically significant trends in mouse liver that were concordant
556 with our findings in genome-edited hPSCs, we found that there was substantial mouse-to-mouse
557 variability with respect to gene expression of *Cep250*, *Cpne1*, and *Ergic3* in fat. Thus, the
558 challenges inherent in studying common DNA variants with small effect sizes may apply to both
559 mice and hPSC models. One difference between the model systems, of course, is that it would be
560 much easier to generate a very large number of isogenic mice, through breeding, than it would be
561 generate a very large number of independent isogenic hPSC clones in order to increase the power
562 of a study.
563

564 Finally, we note the discordance among some of the results obtained from the MPRA
565 experiments and the genome editing/CRISPRi experiments. While the results for rs10889356
566 were concordant—all three types of experiments agreed that the major allele (G) has enhancer
567 activity—with respect to rs2277862, the MPRA data suggest that the minor allele (T) has more
568 enhancer activity on a closely linked reporter gene, while the genome editing/CRISPRi data
569 suggest that the major allele (C) has more enhancer activity on a more distant endogenous gene.
570 A salient difference between the two types of experiments is that the MPRA experiments
571 assessed heterologously expressed 144-nt DNA fragments shorn of their genomic context, which
572 could substantially alter their effects on gene expression. In contrast, both the genome editing
573 and CRISPRi experiments targeted endogenous sequences. In light of this distinction, MPRA
574 experiments should be considered exploratory only, whereas genome editing and CRISPRi
575 experiments should be considered more representative of regulatory effects in endogenous loci,
576 and accordingly, requisite confirmation for any results obtained by MPRA.

577
578 In summary, these studies have highlighted the utility of high-throughput functional genomics
579 approaches to prioritize putative causal GWAS variants, and they provide validation of
580 rs2277862 and rs10889356 as casual eQTL variants for two lipid-associated loci. These studies
581 also illustrate the challenges inherent in modeling common genetic variation in available model
582 systems.

583

584 **Materials and Methods**

585

586 **Massively parallel reporter assays**

587
588 MPRA experiments were performed as previously described [10]. Approximately 240,000 144-
589 bp oligonucleotides representing ~11,000 distinct “tiles” with the major or minor alleles of 1,837
590 candidate causal variants in the center, left-shifted, or right-shifted positions and coupled to
591 distinguishing barcodes were generated by microarray-based DNA synthesis (Agilent). The tiles
592 and barcodes were separated by two common restriction sites. The oligonucleotides were PCR
593 amplified using universal primer sites and directionally cloned into a pMPRA1 (Addgene
594 plasmid #49349) backbone using Gibson assembly. A minimal promoter-firefly luciferase
595 segment from pMPRA_{donor2} (Addgene plasmid #49353) was inserted between the tiles and
596 barcodes via double digestion and directional ligation. The resulting reporter plasmid pools were
597 co-transfected into either undifferentiated 3T3-L1 cells (pre-adipocytes) or differentiated 3T3-L1
598 adipocytes using FuGENE 6 (Promega). Two biological replicate MPRA experiments were
599 performed in each cell type. The relative enhancer activities of the different tiles were calculated
600 by comparing the corresponding barcodes from the cellular mRNA and the transfected plasmid
601 pool. The highest-priority “hits” were tiles for which there was the greatest difference (in
602 magnitude and statistical significance) in enhancer activity between the major and minor alleles
603 in multiple positions (center, left-shifted, right-shifted) in replicate experiments in both
604 adipocytes and pre-adipocytes.

605
606 For the single-hit saturation mutagenesis MPRA experiments [10], all possible single
607 substitutions were introduced at all positions (except the position of the SNP) within the
608 rs2277862 right-shifted tile (with either the major or minor allele) or the rs10889356 left-shifted
609 tile (with either the major or minor allele). The pools of variant tiles were generated, introduced

610 into reporter constructs, transfected into 3T3-L1 adipocytes, and analyzed as described above.

611 The plots in **Fig 1** were generated as previously described [10].

612

613 **CRISPR-Cas9 plasmid construction**

614

615 Guide RNAs were designed by manual inspection of the genomic sequence flanking rs2277862

616 and rs10889356 and evaluated for potential off-target activity using the CRISPR Design tool at

617 <http://crispr.mit.edu>. Protospacers were cloned into the BbsI site of pGuide (Addgene plasmid

618 #64711) via the oligonucleotide annealing method, and, if not already present, a G was added to

619 the 5' end to facilitate U6 polymerase transcription. Genome editing was performed using

620 pCas9_GFP (Addgene plasmid #44719), which co-expresses a human codon-optimized Cas9

621 nuclease and GFP via a viral 2A sequence.

622

623 For CRISPRi studies, pAC154-dual-dCas9VP160- sgExpression (Dr. Rudolph Jaenisch,

624 Addgene plasmid #48240), a dual expression construct that expresses dCas9-VP160 and sgRNA

625 from separate promoters, was modified by PCR-based methods to remove the VP160 domain

626 and include a viral 2A sequence and GFP after dCas9. Additionally, the gRNA sequence was

627 modified to include a 5-bp hairpin extension, which improves Cas9-gRNA interaction, and a

628 single base pair substitution (A-U flip) that removes a putative Pol III terminator sequence, as

629 described previously [19].

630

631 **Cultured cell line maintenance and transfection**

632

633 All cell lines were maintained in a humidified 37°C incubator with 5% CO₂. HEK 293T,
634 HepG2, and 3T3-L1 cells were cultured in high glucose DMEM supplemented with 10% FBS
635 and 1% penicillin/streptomycin. For MPRA experiments, 3T3-L1 cells were differentiated into
636 adipocytes with the addition of 0.5 mM IBMX, 1 µM dexamethasone, and 10 µg/mL insulin to
637 the media for 3 days, followed by addition of only 10 µg/mL insulin for another 3 days. For
638 CRISPRi experiments, HEK 293T and HepG2 cells were seeded into 6-well plates and
639 transfected 24 hours later using Lipofectamine 3000 (Life Technologies) according to the
640 manufacturer's instructions.

641

642 **hPSC culture and CRISPR-Cas9 targeting**

643

644 HUES 8 and H7 cells were grown under feeder-free conditions on Geltrex (Life Technologies)-
645 coated plates in chemically defined mTeSR1 medium (STEMCELL Technologies) supplemented
646 with 1% penicillin/streptomycin and 5 µg/mL Plasmocin (InvivoGen). Medium was changed
647 every 24 hours. For electroporation, cells in a 60%-70% confluent 10-cm plate were dissociated
648 into single cells with Accutase (Life Technologies), resuspended in PBS, and combined with 25
649 µg pCas9_GFP and 25 µg gRNA plasmid (or 12.5 µg of two different gRNA plasmids, for
650 multiplexed targeting) in a 0.4 cm cuvette. For knock-in, 15 µg pCas9_GFP, 15 µg gRNA
651 plasmid, and 30 µg ssODN (5'-

652 GGTCGTCAGAACCCACGAGGTCATGATCAAATATGGCGACCGTCAGCTCCGT

653 CTCAGCTGGGAGAGA-3') were used instead. A single pulse was delivered at 250 V/500 µF

654 (Bio-Rad Gene Pulser), and the cells were recovered and plated in mTeSR1 with 0.4 µM ROCK

655 inhibitor (Y-27632, Cayman Chemical). Cells were dissociated with Accutase 48 hours post-

656 electroporation, and GFP-positive cells were isolated by FACS (FACSAriaII, BD Biosciences)
657 and replated onto 10-cm Geltrex-coated plates (15,000 cells/plate) with conditioned medium and
658 0.4 μ M ROCK inhibitor to facilitate recovery.

659

660 **Isolation and screening of clonal hPSC populations**

661

662 Following FACS, single cells were permitted to expand for 10-14 days to establish clonal
663 populations. Colonies were manually picked and replated into individual wells of a 96-well plate.
664 Once the wells reached 80-90% confluence, cells were dissociated with Accutase and split at a
665 1:3 ratio to create a frozen stock and two working stocks that were maintained in culture. For
666 genomic DNA isolation, cells from one of the working stocks were lysed in 50 μ L lysis buffer
667 (10 mM Tris pH 7.5, 10 mM EDTA, 10 mM NaCl, 0.5% Sarcosyl) with 40 μ g/mL Proteinase K
668 for 1-2 hours in a humidified incubator at 56°C. Genomic DNA was precipitated by addition of
669 100 μ L 95% ethanol with 75 mM NaCl, followed by incubation at -20°C for 2 hours.
670 Precipitated DNA was washed three times with 70% ethanol, resuspended in 30-50 μ L TE buffer
671 with 0.1 mg/mL RNase A, and allowed to dissolve at room temperature overnight.

672

673 hPSC clones were screened by PCR amplification of a small region surrounding the targeted site
674 using BioReady rTaq DNA Polymerase (Bulldog Bio) and the following cycling conditions:
675 94°C for 5 min, [94°C for 30 sec, 54-56.5°C for 30 sec, 72°C for 30 sec] \times 40 cycles, 72°C for 5
676 min. The following primer pairs were used: for rs2277862, F: 5'-
677 TGCTGGACCCACACTTCATA-3' and R: 5'-CTCAGTCCCTCTCCCTCCTT-3'; for
678 rs10889356, F: 5'-CCATTAGGTCAGTGGCCAGA-3' and R: 5'-

679 ACAGGGGGATTCTGTCTAAAA-3'. PCR amplicons were separated on a high-percentage
680 agarose gel, and clones with indels were identified based on size shifts relative to the wild-type
681 band. Suspected mutant clones were confirmed by Sanger sequencing of the PCR products.

682

683 Multiple mutant clones were retrieved from the frozen stock, or if possible, from the second
684 working stock and expanded for experiments. Additionally, several clones that underwent the
685 targeting procedure but remained genetically wild-type at the intended site were expanded as
686 controls.

687

688 **Differentiation of hPSCs into white adipocytes**

689

690 Differentiation of HUES 8 cells to white adipocytes was performed as previously described [20].
691 To induce embryoid body formation, wild-type and mutant hPSCs were pre-treated overnight
692 with 2% DMSO; dissociated into small clumps with Accutase; resuspended in growth medium
693 containing DMEM, 10% knockout serum replacement (Life Technologies), 2 mM GlutaMAX
694 (Life Technologies), 1% non-essential amino acids, 1% penicillin/streptomycin, and 0.1 mM
695 beta-mercaptoethanol; and transferred to low-attachment 6-well plates (Costar Ultra Low
696 Attachment; Corning Life Sciences). After one week in culture, embryoid bodies were collected
697 and replated onto gelatin-coated plates in MPC medium containing DMEM, 10% FBS, 1%
698 penicillin/streptomycin, and 2.5 ng/mL bFGF (Aldevron). Cells were serially passaged at a 1:3
699 ratio to obtain a homogenous population of MPCs by passage 3-4.

700

701 Recombinant lentivirus was produced using a third-generation, Tat-free packaging system.
702 Lentiviral vectors encoding either doxycycline-inducible *PPARG2* or rtTA were transfected into
703 HEK 293T cells by the calcium phosphate method, along with the packaging plasmids pMDL
704 and pREV and a capsid plasmid encoding VSV-G. Viral supernatant was harvested at 48 and 72
705 hours post-transfection and filtered through a 0.45 μm membrane. One day before transduction,
706 MPCs were plated at 1×10^6 cells per 10-cm plate. The following day, MPCs were transduced
707 with 5 mL lenti-*PPARG2* and 5 mL lenti-rtTA and incubated at 37°C for 16 hours. After the
708 viral supernatant was aspirated, the cells were washed with PBS and allowed to grow to
709 confluence. Transduced MPCs were split into 6-well dishes prior to initiating white adipocyte
710 differentiation.

711
712 Differentiation was induced by the addition of adipogenic media containing DMEM, 7.5%
713 knockout serum replacement, 7.5% human plasmanate (Grifols), 0.5% non-essential amino
714 acids, 1% penicillin/streptomycin, 0.1 μM dexamethasone (Sigma), 10 $\mu\text{g}/\text{mL}$ insulin (Sigma),
715 and 0.5 μM rosiglitazone (Santa Cruz). The differentiation medium was supplemented with 700
716 ng/mL doxycycline from day 0 to 16. Doxycycline was then removed from the culture medium
717 until day 21, at which point the differentiated cells were used for experiments.

718

719 **Differentiation of hPSCs into hepatocyte-like cells**

720

721 Differentiation of H7 cells into HLCs was performed as previously described [17]. One day
722 before differentiation, hPSCs at 60% confluence were split at a 1:3 ratio into 6-well dishes with
723 mTeSR1 plus 0.4 μM ROCK inhibitor. Cells were serially cultured in (1) RPMI-B27 (RPMI-

724 1640 from Sigma; B27 supplement minus Vitamin A from Life Technologies) supplemented
725 with 100 ng/mL Activin A (PeproTech) and 3 μ M CHIR99021 (Cayman Chemical), a glycogen
726 synthase kinase 3 inhibitor, for 3 days to obtain definite endoderm; (2) RPMI- B27 supplemented
727 with 5 ng/mL bFGF (Millipore), 20 ng/mL BMP4 (PeproTech), and 0.5% DMSO for 5 days to
728 obtain hepatic endoderm; (3) RPMI-B27 supplemented with 20 ng/mL HGF (PeproTech) and
729 0.5% DMSO for 5 days to obtain hepatoblasts; and (4) Hepatocyte Culture Medium (Lonza)
730 supplemented with 20 ng/mL HGF, 20 ng/mL Oncostatin M (PeproTech), 100 nM
731 dexamethasone (Sigma), and 0.5% DMSO for 10 to 12 days to obtain HLCs.

732

733 **CRISPR knock-in mice**

734

735 Four candidate guide RNAs with a cut site near rs27324996 were designed by manual
736 inspection, and the corresponding protospacers were cloned into the pGuide plasmid as described
737 above. Each gRNA plasmid was co-transfected with pCas9_GFP into mouse 3T3-L1 cells using
738 TransIT-2020 Reagent (Mirus Bio) according to the manufacturer's instructions. Two days post-
739 transfection, GFP-positive cells were isolated by FACS, and genomic DNA was isolated using
740 the DNeasy Blood and Tissue Kit (Qiagen). The region flanking rs27324996 was PCR amplified
741 (F: 5'-TGGGAATGGCTTCTTAGGGC-3' and R: 5'-CATCCCCAAGCAACTCAACC-3')

742 using AccuPrime Taq DNA Polymerase (Life Technologies) with the following cycling
743 conditions: 94°C for 2 min, [94°C for 30 sec, 55°C for 30 sec, 68°C for 30 sec] \times 40 cycles,
744 68°C for 5 min. PCR products were purified using the DNA Clean and Concentrator kit (Zymo
745 Research) and analyzed for the presence of indels using the Surveyor Mutation Detection Kit
746 (IDT) according to the manufacturer's instructions. Cel-I nuclease-treated PCR products were

747 resolved on a 1.5% agarose gel to detect mutagenesis activity. The gRNA sequence exhibiting
748 the highest mutation rate was PCR amplified, and the purified PCR product was used as a
749 template for *in vitro* transcription using the MEGAscript T7 kit (Life Technologies). The
750 transcribed RNA was purified by phenol/chloroform extraction, ethanol precipitated, and
751 resuspended in injection buffer (5 mM Tris-HCl pH 7.6, 0.1 mM EDTA).

752

753 All animal procedures described here were reviewed and approved by the Harvard University
754 Institutional Animal Care and Use Committee (protocol #14-05-202). Euthanasia in all instances
755 was via terminal inhalation of carbon dioxide, consistent with the 2013 AVMA Guidelines on
756 Euthanasia.

757

758 One-cell embryo injections were performed by the Genome Modification Facility at Harvard
759 University. Superovulated C57BL/6J females were mated with C57BL/6J males, and fertilized
760 embryos were harvested from the oviducts. One-cell embryos were injected with a mixture of
761 100 ng/ μ L Cas9 mRNA (TriLink BioTechnologies), 50 ng/ μ L gRNA, and 100 ng/ μ L ssODN
762 (5'-AGCCACAGTTGGCTCTGTGGTGGCTATAGAATCTGTTTTCCAGGTCAATGTG
763 GGTCTCCCCGATGAGGTCATCTGAACCCACGAGGTCATGATCAAATATGGCG
764 ACCGTCAGCTCTGGCTGGGCTGGGAGGGAGACGCTCAGCTCCAGGACCCTGG
765 GCAGGAAGGGAAATTGACTAACCACAGCTCCATGCCCTCAGAG-3'). Injected embryos
766 were implanted into the uteri of pseudopregnant foster mothers.

767

768 DNA was prepared from tail biopsies of 3-week-old founder mice by the hot hydroxide method,
769 and genotyping was performed with the same PCR primers and cycling conditions used for the

770 Cel-I nuclease assay. Positive founders were identified by Sanger sequencing of PCR products.
771 The single positive founder was bred to a wild-type C57BL/6J mouse (Jackson Laboratories),
772 and the resulting progeny were intercrossed for one to two generations to breed the knock-in
773 allele to homozygosity. Genotyping of progeny was performed in the same manner. Wild-type
774 and homozygous knock-in littermates from several litters, ~12 weeks of age, were used for gene
775 expression analyses.

776

777 **Quantitative reverse transcriptase-polymerase chain reaction**

778

779 Wells of cells were washed with ice-cold PBS and lysed directly in TRIzol Reagent (Thermo
780 Fisher Scientific), and primary liver and fat samples from mice were homogenized in TRIzol
781 Reagent. RNA was isolated according to the manufacturer's instructions and reverse transcribed
782 using SuperScript III Reverse Transcriptase (Thermo Fisher Scientific) with an equimolar
783 mixture of random hexamers and oligo-dT. Gene expression was measured using the following
784 TaqMan assays (Applied Biosystems): Hs00898245_m1 for *CEP250*, Hs00537765_m1 for
785 *CPNE1*, Hs00211070_m1 for *ERGIC3*, Hs00205581_m1 for *ANGPTL3*, Hs00290630_m1 for
786 *DOCK7*, Hs00910225_m1 for *ALB*, Hs01097800_m1 for *SERPINA1*, Mm00623502_m1 for
787 *Cep250*, Mm00467970_m1 for *Cpne1*, and Mm00499400_m1 for *Ergic3*. Human *B2M* (Assay
788 ID 4326319E) or mouse *Actb* (Assay ID 4352341E) was used as the reference gene. Each 10 μ L
789 qPCR reaction contained 1 μ L cDNA (diluted 1:3 with water) and was performed in technical
790 duplicate or triplicate. Reactions were carried out on a ViiA 7 Real-Time PCR system (Applied
791 Biosystems), and relative expression differences were quantitated by the $\Delta\Delta C_t$ method.

792

793 **Acknowledgments**

794

795 We are grateful to Qiurong Ding, Alexandra Chadwick, and Alanna Strong for critical reading of
796 the manuscript.

797

798 **References**

- 799 1. Manolio TA. Genomewide association studies and assessment of the risk of disease. *N Engl J*
800 *Med.* 2010; 363: 166–1176. doi: 10.1056/NEJMra0905980 PMID: 20647212
- 801 2. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al.
802 Biological, clinical and population relevance of 95 loci for blood lipids. *Nature.* 2010; 466:
803 707–713. doi: 10.1038/nature09270 PMID: 20686565
- 804 3. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, et al. From
805 noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature.* 2010; 466:
806 714–719. doi: 10.1038/nature09266 PMID: 20686566
- 807 4. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gómez-Marín C, et al. Obesity-
808 associated variants within FTO form long-range functional connections with IRX3. *Nature.*
809 2014; 507: 371–375. doi: 10.1038/nature13138 PMID: 24646999
- 810 5. Battle A, Montgomery SB. Determining causality and consequence of expression
811 quantitative trait loci. *Hum Genet.* 2014; 133: 727–735. doi: 10.1007/s00439-014-1446-0
812 PMID: 24770875
- 813 6. Gupta RM, Musunuru K. Mapping novel pathways in cardiovascular disease using eQTL
814 data: the past, present, and future of gene expression analysis. *Front Genet.* 2013; 3: 232. doi:
815 10.3389/fgene.2012.00232 PMID: 23755065

- 816 7. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a
817 genetic journey. *Hum Mol Genet.* 2008; 17: R156–R165. doi: 10.1093/hmg/ddn289 PMID:
818 18852205
- 819 8. Raychaudhuri S. Mapping rare and common causal alleles for complex human diseases. *Cell.*
820 2011; 147: 57–69. doi: 10.1016/j.cell.2011.09.011 PMID: 21962507
- 821 9. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark
822 road from association to function. *Am J Hum Genet.* 2013; 93: 779–797 doi:
823 10.1016/j.ajhg.2013.10.012 PMID: 24210251
- 824 10. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic
825 dissection and optimization of inducible enhancers in human cells using a massively parallel
826 reporter assay. *Nat Biotechnol.* 2012; 30: 271–277. doi: 10.1038/nbt.2137 PMID: 22371084
- 827 11. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, et al. Massively parallel
828 functional dissection of mammalian enhancers in vivo. *Nat Biotechnol.* 2012; 30: 265–270.
829 doi: 10.1038/nbt.2136 PMID: 22371081
- 830 12. Musunuru K. Genome editing of human pluripotent stem cells to generate human cellular
831 disease models. *Dis Model Mech.* 2013; 6: 896–904. doi: 10.1242/dmm.012054 PMID:
832 23751357
- 833 13. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome
834 engineering. *Cell.* 2014; 157: 1262–1278. doi: 10.1016/j.cell.2014.05.010 PMID: 24906146
- 835 14. Musunuru K, Pirruccello JP, Do R, Peloso GM, Guiducci C, Sougnez C, et al. Exome
836 sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N Engl J Med.*
837 2010; 363: 2220–2227. doi: 10.1056/NEJMoa1002926 PMID: 20942659

- 838 15. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, et al. Repurposing
839 CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*.
840 2013; 152: 1173–1183. doi: 10.1016/j.cell.2013.02.022 PMID: 23452860
- 841 16. Yang H, Wang H, Shivalila CS, Cheng AW, Shi L, Jaenisch R. One-step generation of mice
842 carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering.
843 *Cell*. 2013; 154: 1370–1379. doi: 10.1016/j.cell.2013.08.022 PMID: 23992847
- 844 17. Ding Q, Lee YK, Schaefer EA, Peters DT, Veres A, Kim K, et al. A TALEN genome-editing
845 system for generating human stem cell-based disease models. *Cell Stem Cell*. 2013; 12: 238–
846 251. doi: 10.1016/j.stem.2012.11.011 PMID: 23246482
- 847 18. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, et al. CRISPR-mediated
848 modular RNA-guided regulation of transcription in eukaryotes. *Cell*. 2013; 154: 442–451.
849 doi: 10.1016/j.cell.2013.06.044 PMID: 23849981
- 850 19. Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, Li GW, et al. Dynamic imaging
851 of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell*. 2013; 155:
852 1479–1491. doi: 10.1016/j.cell.2013.12.001. PMID: 24360272
- 853 20. Ahfeldt T, Schinzel RT, Lee YK, Hendrickson D, Kaplan A, Lum DH, et al. Programming
854 human pluripotent stem cells into white and brown adipocytes. *Nat Cell Biol*. 2012; 14: 209–
855 219. doi: 10.1038/ncb2411. PMID: 22246346

856

857 **Supporting Information Captions**

858

859 **S1 Table. Candidate SNPs in adipose lipid-associated eQTLs that were interrogated in the**
860 **MPRA experiments.**

861

862 **S2 Table. MPRA experimental data.** Each row corresponds to a 144-bp tile either centered (C),

863 right-shifted (R), or left-shifted (L) relative to a SNP, with the two alleles indicated by the last

864 two letters in the label for each row (penultimate is allele 1, last is allele 2). Each allele on each

865 tile was repeated in the design with 22 different barcodes, although due to synthesis and cloning

866 biases, some have too many dropouts to give robust results (marked by NaN). Four independent

867 transfection experiments were performed: L1Ad_R1 (3T3-L1 adipocytes #1), L1Ad_R2

868 (adipocytes #2), L1Pre_R1 (3T3-L1 pre-adipocytes #1), and L1Pre_R2 (pre-adipocytes #2). For

869 each SNP and transfection there are 6 columns: *_Sig1 = signal from allele 1 (log of median

870 barcode counts for this tile divided by median barcode counts for all tiles, with a high positive

871 value meaning enhancer activity, and a negative value meaning repressor/silencer activity);

872 *_Sig1_P = Mann-Whitney *U*-test *P*-value for the signal from allele 1 being the same as the

873 median signal from all tiles (not corrected for multiple testing); *_Sig2 = signal from allele 2;

874 *_Sig2_P = *P*-value for allele 2; *_Var = log-ratio of signal from allele 1 over signal from allele

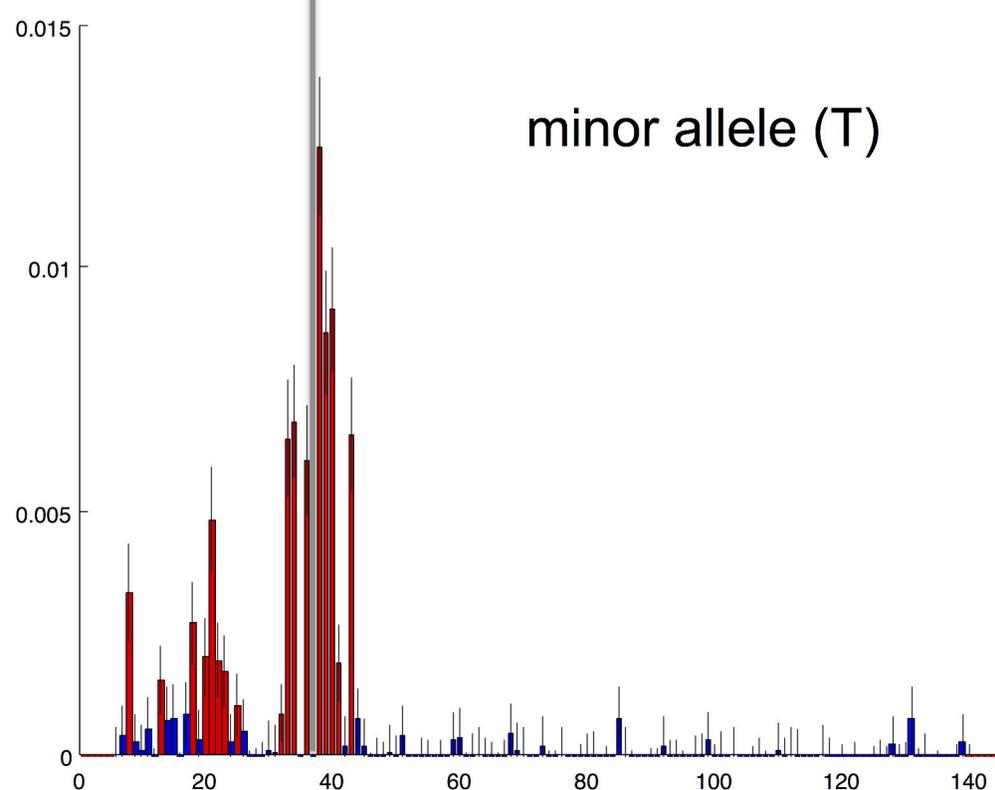
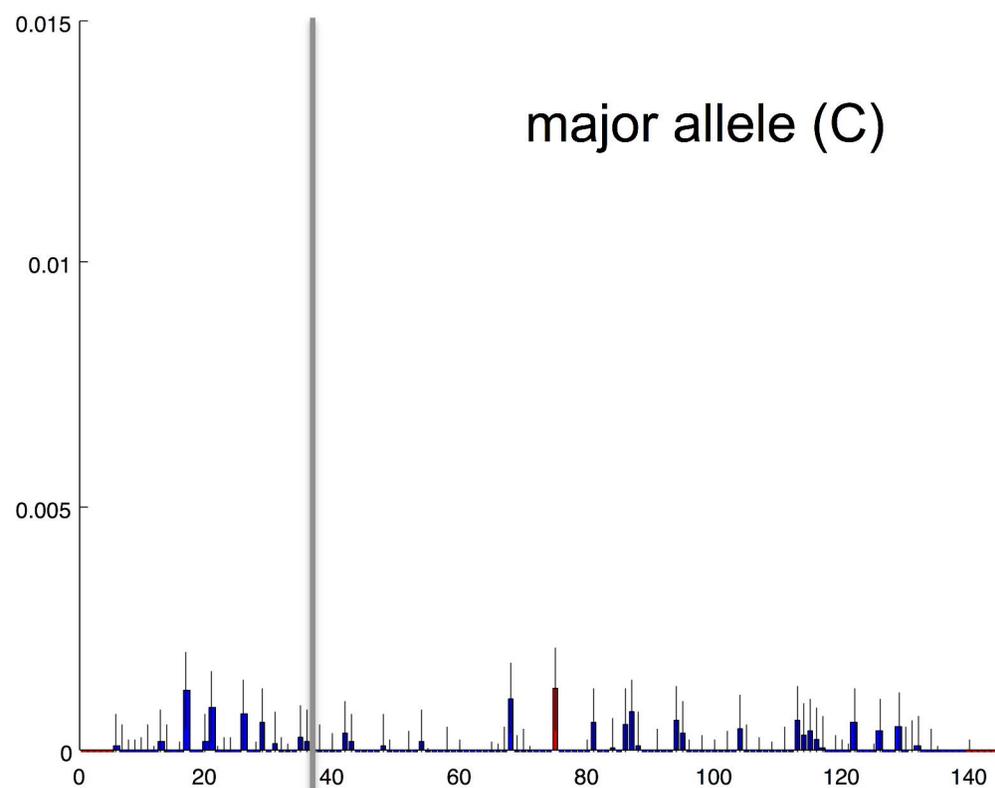
875 2; *_Var_P = *P*-value for the signal from alleles 1 and 2 being the same.

A

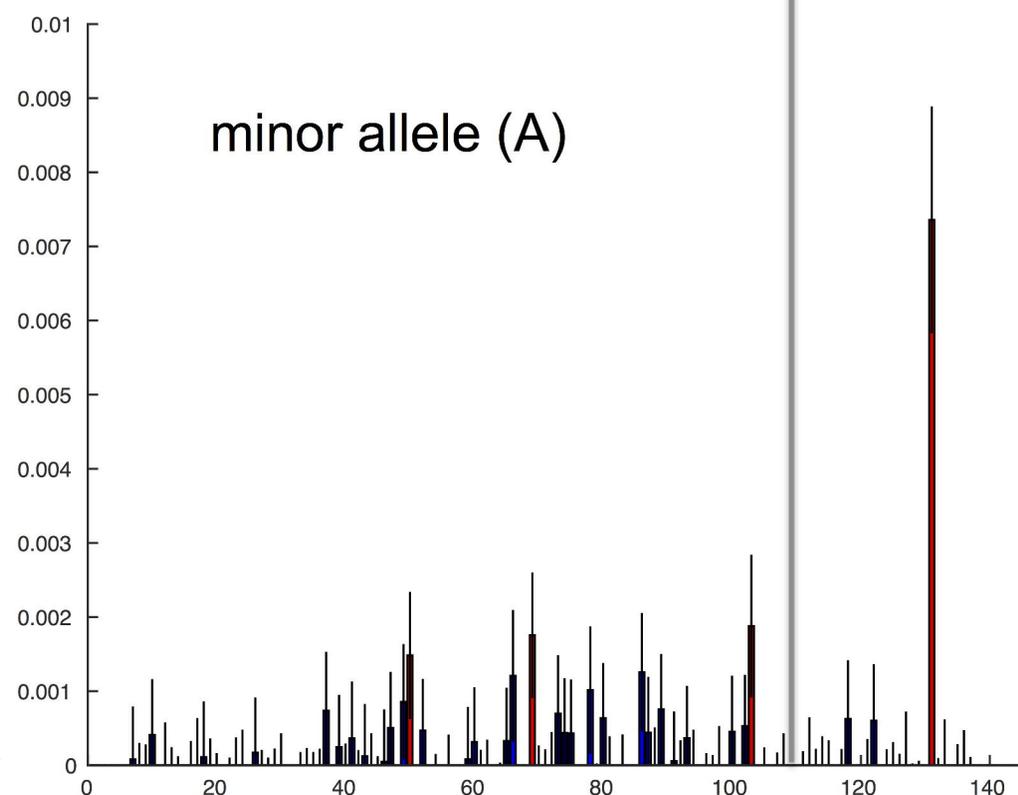
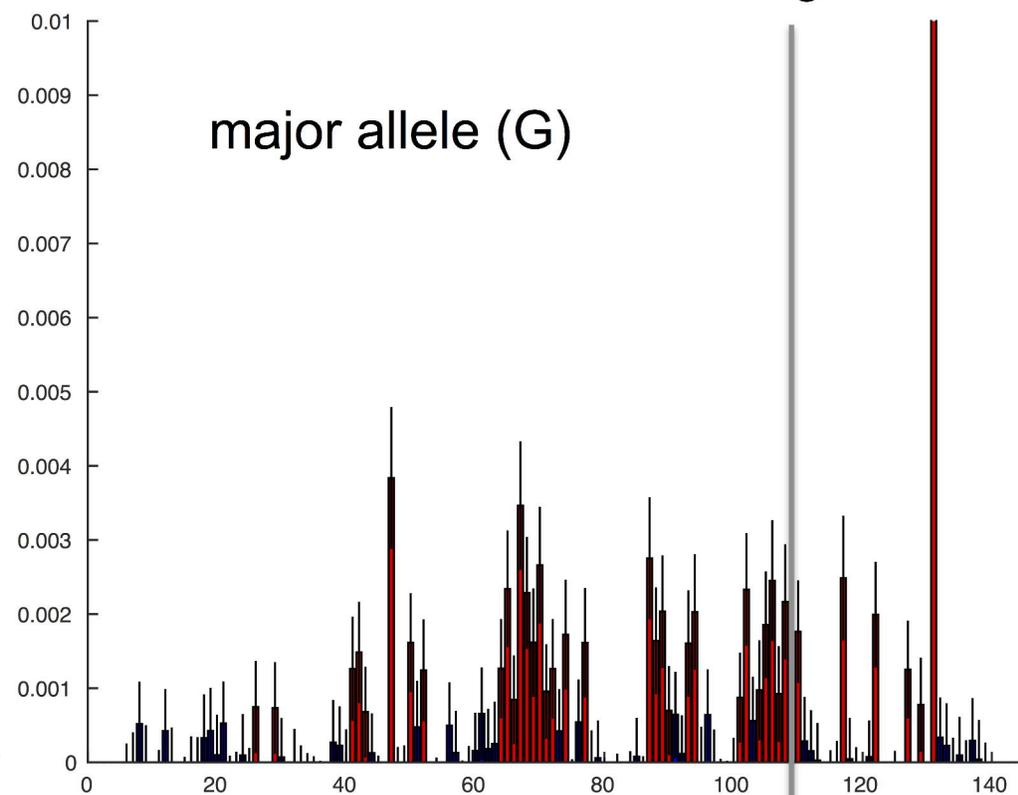
		Major allele signal	Minor allele signal	Log ratio	<i>P</i> -value
rs2277862	<i>Right</i>	0.0868	2.1326	-2.0458	2.21E-07
	<i>Center</i>	0.03925	1.5159	-1.5551	5.33E-07
	<i>Left</i>	0.29406	1.3633	-1.0693	7.22E-07
rs10889356	<i>Right</i>	1.3342	0.55495	0.77926	3.42E-04
	<i>Center</i>	2.825	1.4247	1.4004	1.28E-07
	<i>Left</i>	2.4236	0.50114	1.9225	5.01E-08

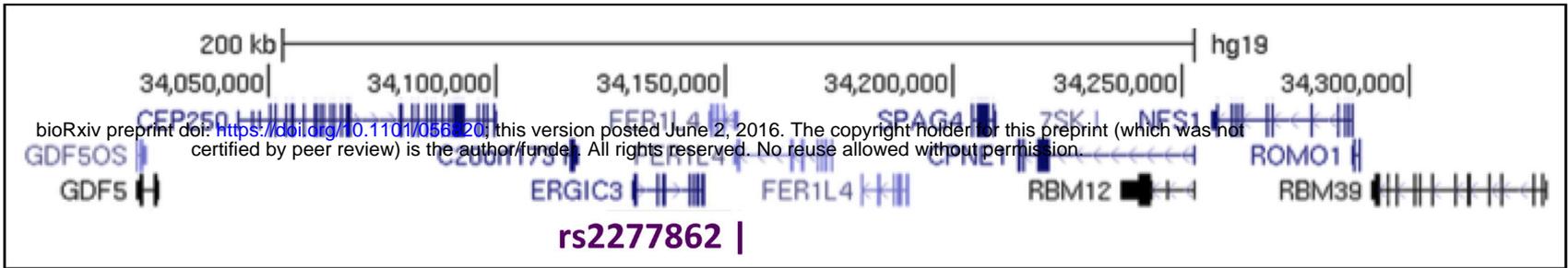
B

rs2277862 saturation mutagenesis

**C**

rs10889356 saturation mutagenesis

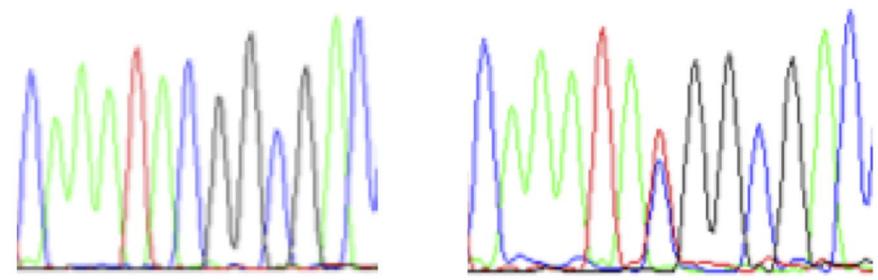


A**B**

Sense strand ACCCACGAGGTCATGATCAAATA**CGG**CGACCGTCAGCTCCGTCTCAGCTGGG
Antisense strand TGGGTGCTCCAGTACTAGTTTAT**G**CCGCTGGCAGTCGAGGCAGAGTCGACCC

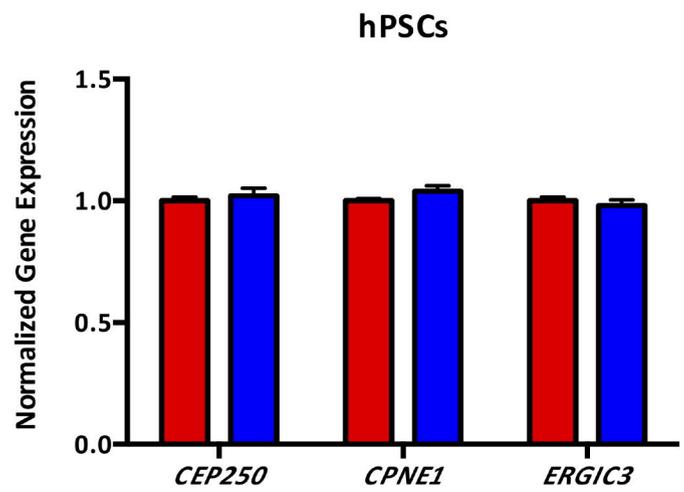
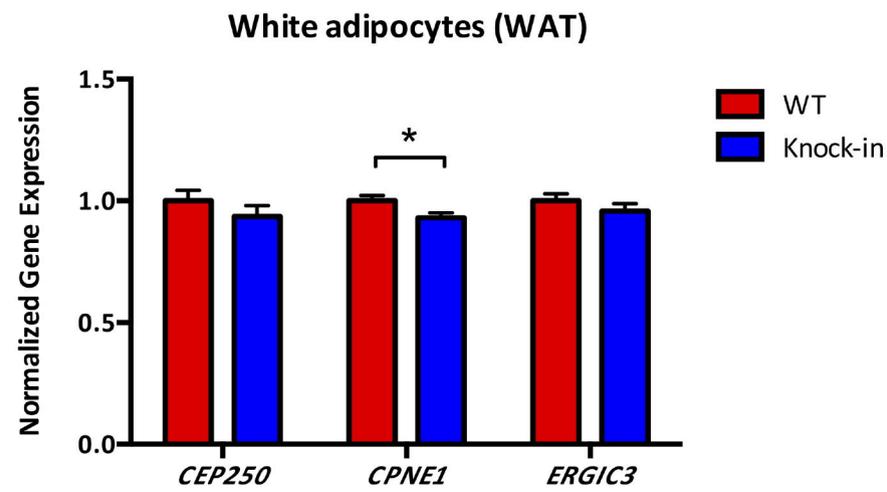
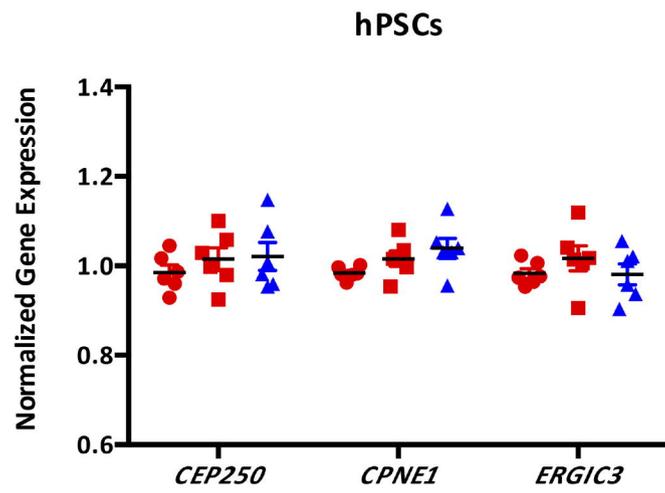
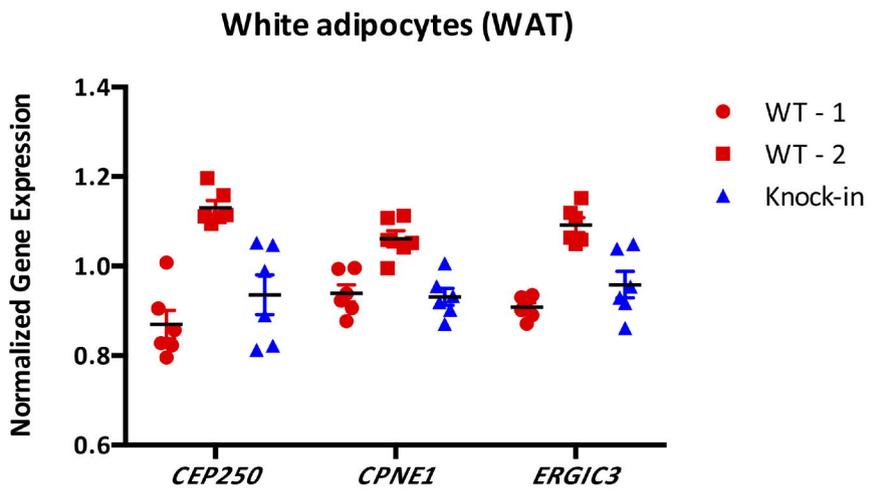
Sense strand (major allele only)
 ACCCACGAGGTCATGATC-----GACCGTCAGCTCCGTCTCAGCTGGG
 ACCCACGAGGTCA-----GCTCCGTCTCAGCTGGG

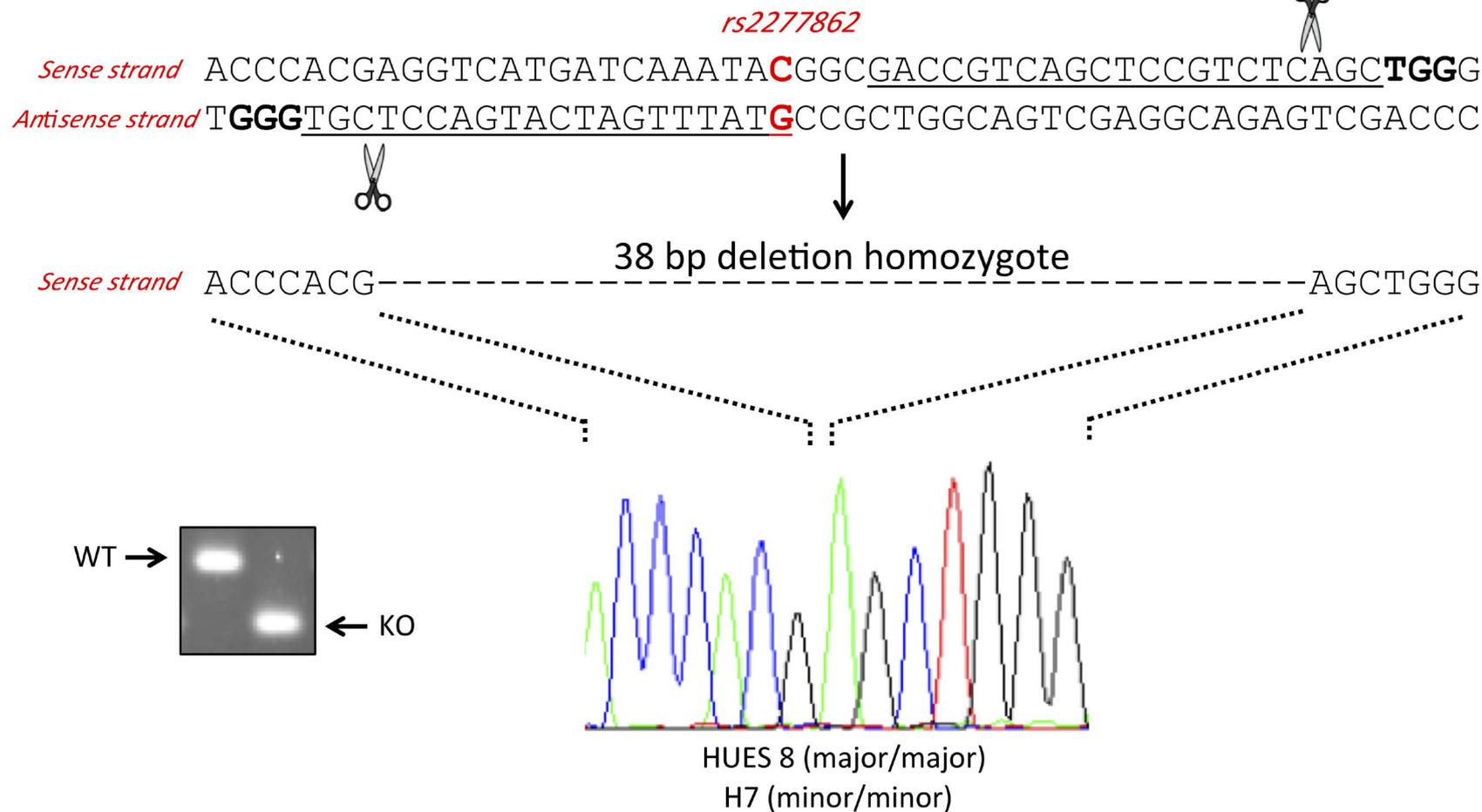
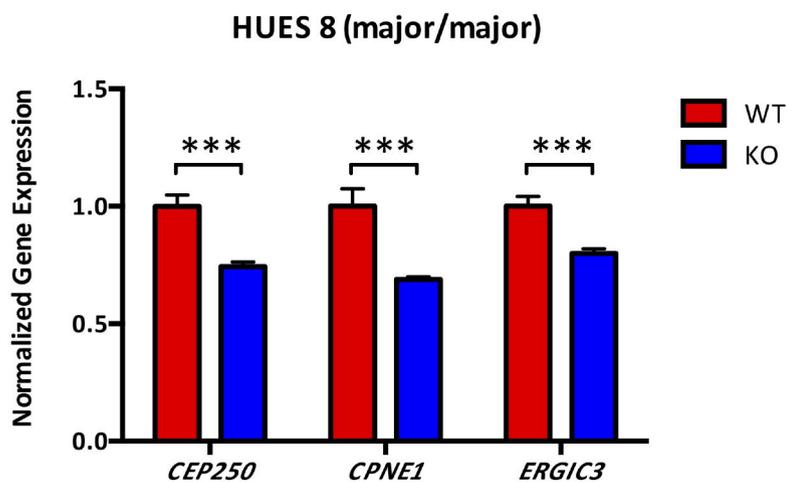
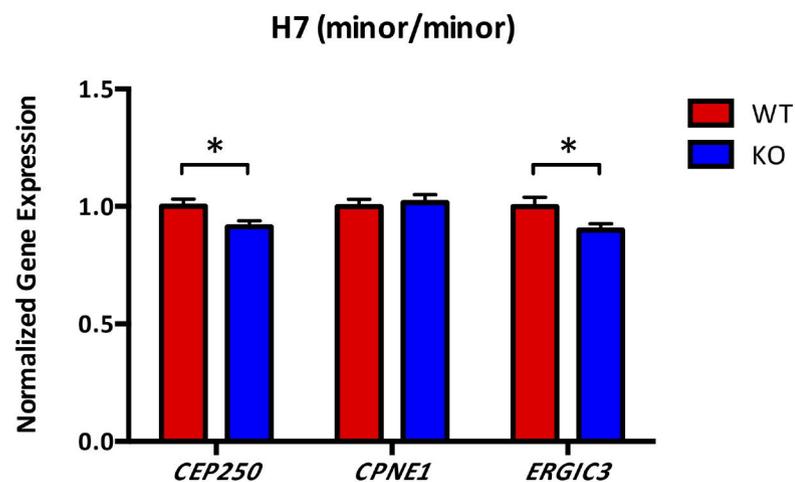
+ knock-in:

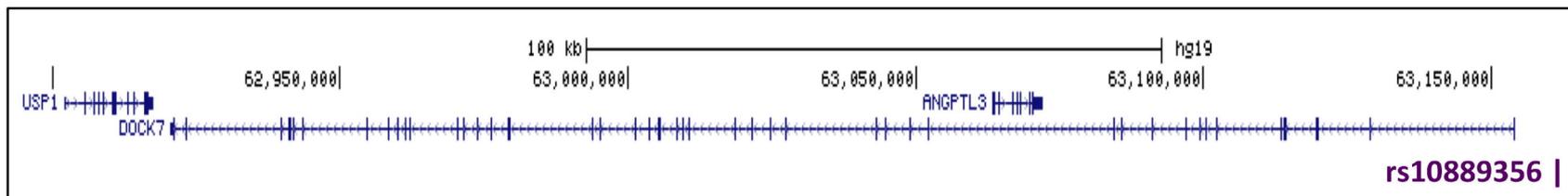
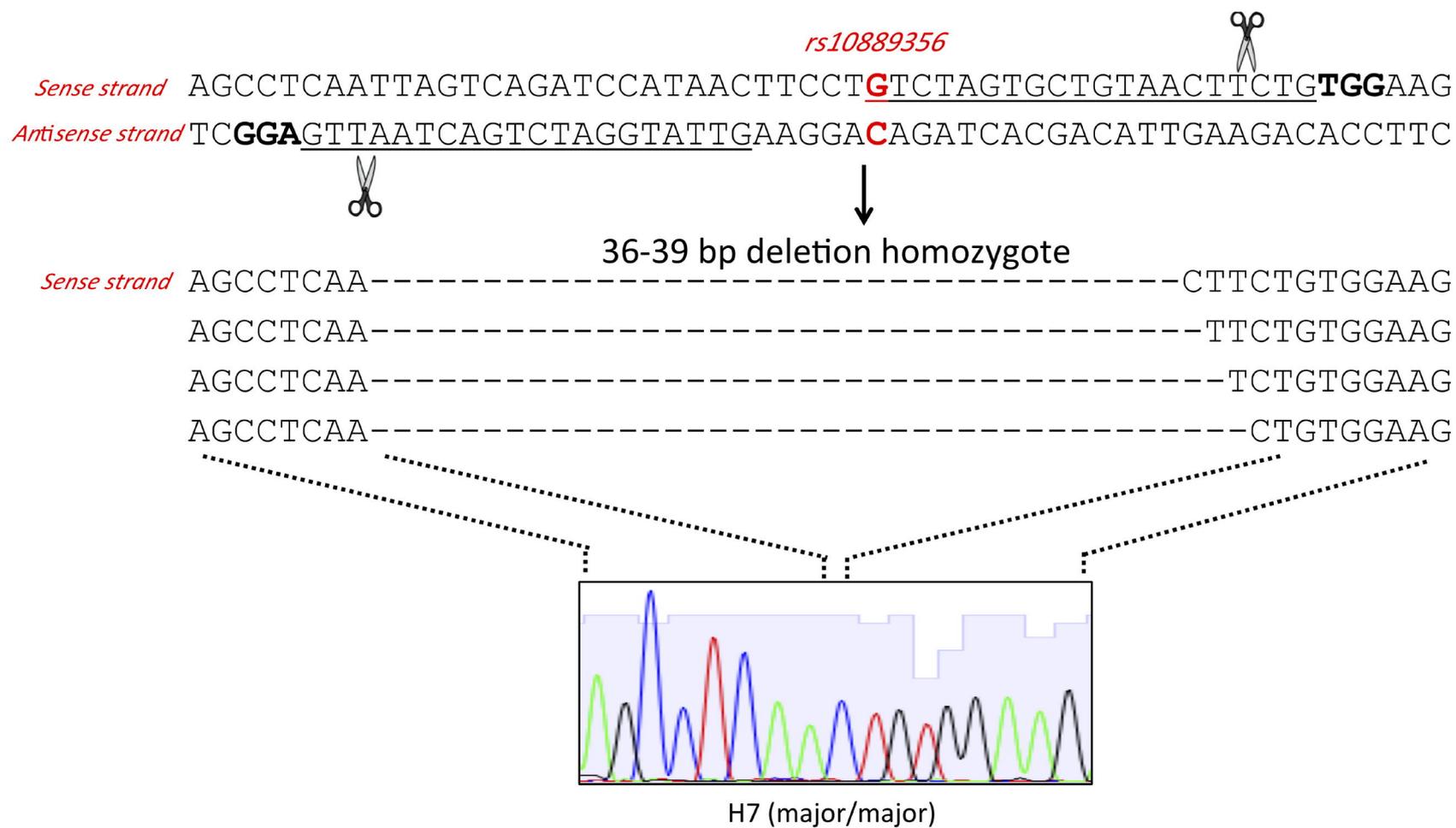
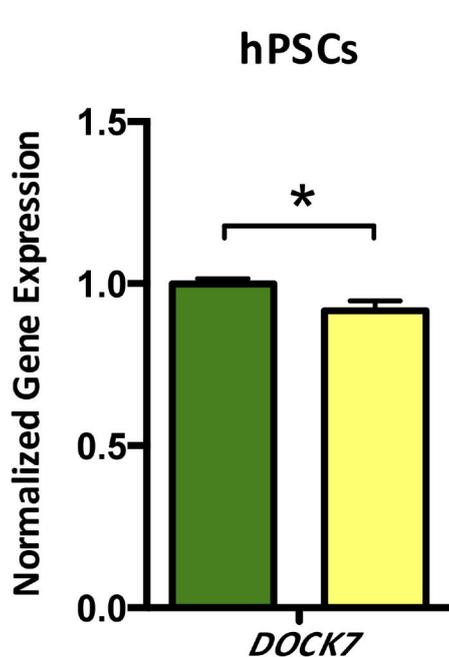
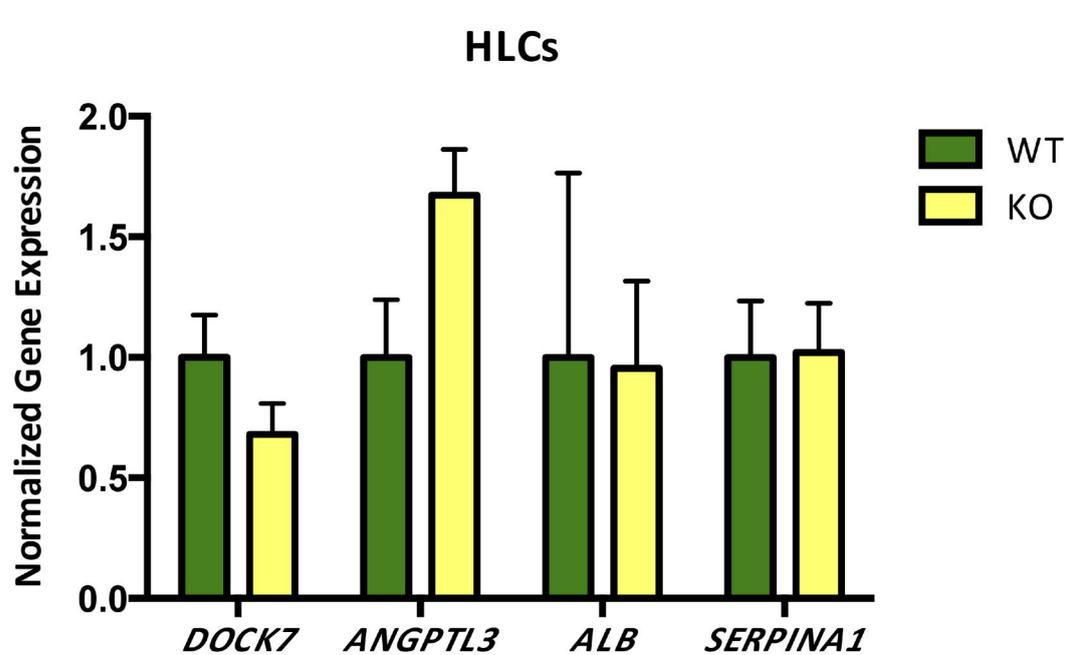


HUES 8 WT
(homozygous major)

HUES 8 recombinant
(heterozygous)

C**D****E****F**

A**B****C**

A**B****C****D**

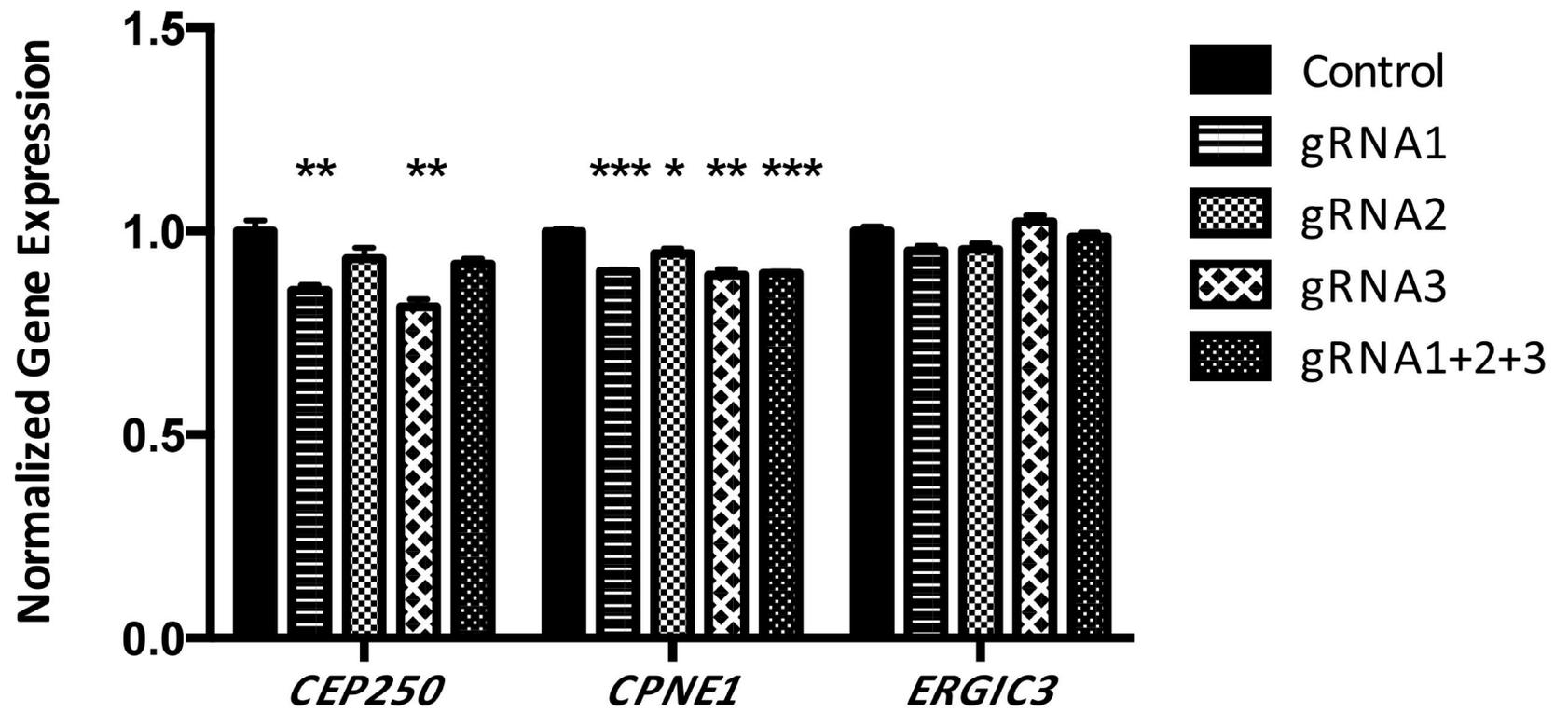
A

gRNA2
rs2277862
gRNA3

TCAGAACCCACGAGGTCATGATCAAATACGGCGACCGTCAGCTCCGTCTCAGCTGGG
 AGTCTTGGGTGCTCCAGTACTAGTTTATGCCGCTGGCAGTCGAGGCAGAGTCGACCC

gRNA1

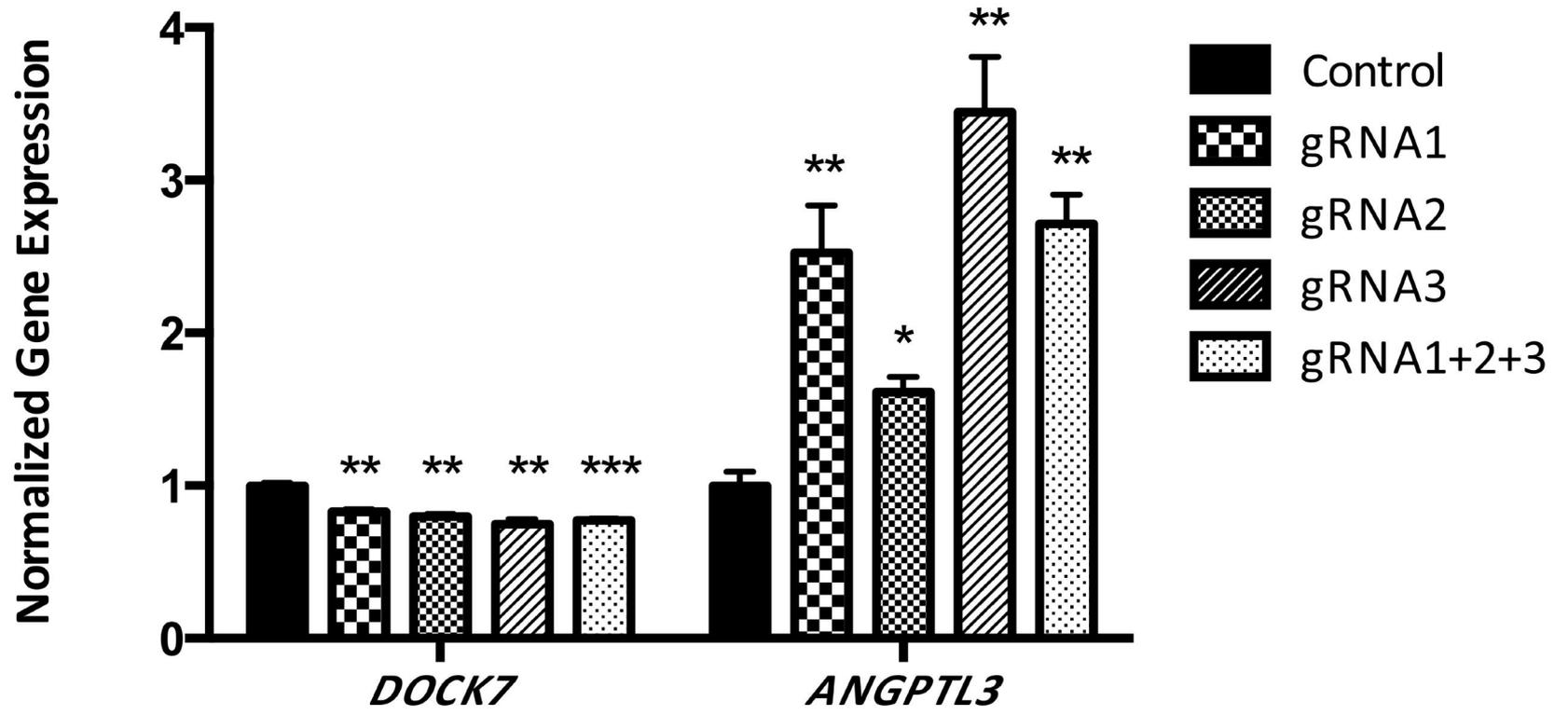
bioRxiv preprint doi: <https://doi.org/10.1101/056921>; this version posted June 2, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

B**C**

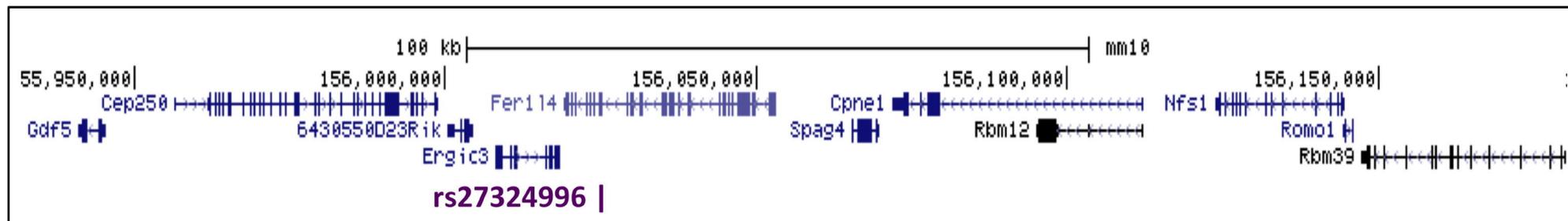
rs10889356
gRNA-3

AGCCTCAATTAGTCAGATCCATAACTTCCTGTCTAGTGCTGTAAGTTCTGTGGAAG
 TCGGAGTTAATCAGTCTAGGTATTGAAGGACAGATCACGACATTGAAGACACCTTC

gRNA-2
gRNA-1

D

A



B

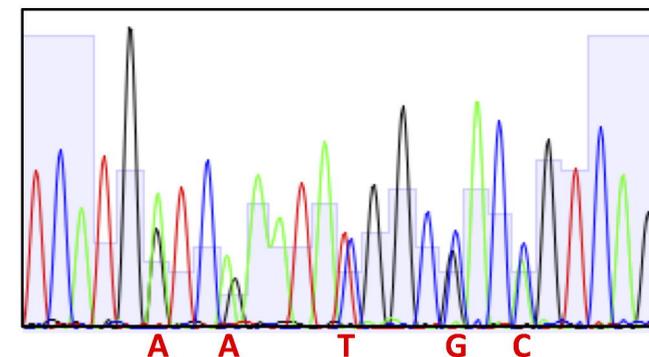
Sequence alignment:

rs27324996 *gRNA protospacer*

Mouse GAGGTCATGGTCTGAATACGGCCACAGTCAGCTCTGGCTGGGCT

Human GAGGTCATGATCAAATACTGGCCGACCGTCTCAGCT

rs2277862

Minor allele
knock-in

C

